

**TESTING REGIME CHANGE AS INNOVATION:
WASHBACK POTENTIAL OVER TIME**

by

Poonam Anand

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© Poonam Anand, 2018

Abstract

Research (Fox, 2004, 2009; Shohamy, 1993) highlights how policy makers use high-stakes tests as part of a testing regime to promote curricular innovation. In this dissertation, a *testing regime* is defined as a systematic approach to the use of assessment (often high-stakes tests) to impact curriculum and classroom practices. The influence of high-stakes tests on classroom teaching and learning is known as *washback*, and the beneficial influences of tests on teaching and learning is *positive washback* (Alderson & Wall, 1993). The present study explored washback as deeply contextual, arising from and influenced by a complex set of interrelated factors (e.g. stakeholders, power-relationships) within a testing regime. Specifically, it investigated whether a high-stakes test could be used to leverage positive washback in an English for Academic Purposes (EAP) program over time, and which factors were most influential.

This longitudinal case study examined washback from a newly introduced high-stakes test –the *innovation*– as evidenced in stakeholders’ accounts: two administrators, 15 teachers, and 201 students, over a period of 20 months. The study drew on Hughes’ (1993) principles of Washback, and Henrichsen’s (1989) Hybrid Model of Diffusion/Implementation of Innovation, and was conducted in three phases: Phase 1 during the former testing regime; Phase 2 during implementation of the new testing regime (immediate washback); Phase 3 after three semesters (delayed washback).

Results indicate that EAP teachers mediated washback to the satisfaction of their students in both the former and new testing regimes. Although the new high-stakes test had the potential for positive washback, based on “an evidential link” (Messick, 1996, p. 247) for washback between the new test and classroom practices, other meta-contextual

factors such as power relationships in the EAP program's testing regimes hindered this potential. It was problematic to use a stand-alone, high-stakes test to leverage positive washback over time.

Recommendations address stakeholder issues within the EAP research site, while calling for in-depth, longitudinal explorations in future washback research. Future studies can examine interactions between stakeholders' accounts and testing regime change contexts, in order to pinpoint specific factors that trigger positive or negative washback within such contexts.

Acknowledgements

There are many individuals to whom I am indebted for this dissertation. First and foremost my gratitude is towards my mentor and supervisor, Dr. Janna Fox, without whom, through her caring guidance and constant support, this work would not have been possible. She was my pillar of strength through the difficult periods of this study. I thank my Advisory Committee, Dr. Eva Kartchava and Dr. Beverly Baker, for their support and comments. My gratitude is also for Dr. Steven E. Noble, a dear friend, who provided valuable assistance throughout the process of writing this dissertation and was always ready to answer my questions. I would also like to thank Dr. Liying Cheng, Dr. John ApSimon, Dr. Natasha Artemeva and Dr. Kartchava for serving on my Examination Committee.

My sincere thanks are for the EAP program's manager, curriculum coordinator, teachers and students for their willingness to participate in this research. I am grateful for their constant support throughout the study. A special thank you is due to the ten PhD students (Adam, Britt, Christopher, Codie, Dimitri, Janna, Lisa, Parisa, Richard and Raouf) at Carleton University for their assistance in ensuring my coding was reliable.

Finally, this work would not have been possible without the blessings of my spiritual guru, Bhagawan Sri Sathya Sai Baba. This work is a humble offering at his feet. I am eternally grateful for the unconditional love and faith of my husband, Sanjay, and children Sathya, Sai Priya and Noémie. They have sacrificed a lot during this process, but never complained. Instead, they encouraged me to follow my dreams and provided all the physical and moral support that I needed. My acknowledgment is also due to my family and friends, for inquiring about this study and gently providing support whenever needed.

Table of Contents

Abstract.....	ii
Acknowledgements	iv
List of Tables	ix
List of Figures	x
List of Abbreviations	xi
Glossary of Key Terms or Concepts	xii
CHAPTER 1: Introduction	1
Researcher Lens	1
Washback.....	3
The Definition of Washback in This Study	6
Washback as Diffusion of Innovation.....	7
Overall Purpose of the Study	14
Specific Research Goals	16
Research Questions	17
Significance of the Study.....	17
Organization of the Dissertation	18
Chapter 2: Washback and Innovation.....	20
Introduction.....	20
Language Tests and Assessing Academic Language Proficiency	21
The Phenomenon of Washback and Its Complexity	32
Conceptual Dimensions of Washback.....	35
Washback and Innovation	40
Insights from Innovation Studies.....	49
The process of innovation.....	51
Models of Innovation	59
Henrichsen’s (1981) Hybrid Model of Diffusion.....	61
Research Questions	70
Chapter Summary.....	73

Chapter 3: Methodology	75
Introduction.....	75
Research Method	75
Researcher Positionality.....	82
Criteria for Judging Reliability and Validity of Case Studies.....	83
Eliciting Case Data by Interviews and Focus Groups.....	86
Study Context	91
Participants	94
Overall Research Design	100
Phase 1: The Antecedents phase – Former Testing Regime and Intended Washback of ISTs	100
Phase 2: The Process phase – Implementation Dynamics and Immediate Washback of the New Testing Regime.....	102
Phase 3: The Consequences phase - Washback of the New Testing Regime Over Time..	103
Procedures	103
Quantitative Data Analysis	114
Qualitative Data Analysis	115
Document Analysis: Evaluation of Test Tasks Characteristics	121
Chapter Summary.....	128
 Chapter 4: Results and Discussion of Phase 1: the Antecedents phase (<i>Former Testing Regime and Intended Washback of ISTs</i>)	 129
Introduction.....	129
Experiences of Previous Reformers.....	130
Traditional Pedagogical Practices	133
Characteristics of the User and the User System.....	135
Research Question 1: What Evidence is There of Washback in the Former Testing Regime and What is the Intended Washback of the New Testing Regime?	135
Washback from the ExitTest.....	138
Intended Washback of the ISTs	163
Chapter Summary.....	172

Chapter 5: Results and Discussion of Phase 2: the Process phase	
<i>(Implementation Dynamics and Immediate Washback of the New Testing Regime)</i>	175
Introduction.....	175
Research Question 2: What Evidence is There of Washback Factors Facilitating and/or Impeding the Implementation of the New Testing Regime?	176
The Innovation Itself: New Integrated Skills Tests.....	177
The Resource System.....	193
The User System.....	203
Inter-Elemental Factors.....	215
Consequences.....	220
Chapter Summary.....	221
Chapter 6: Results and Discussion of Phase 3: the Consequences phase	
<i>(Washback of the New Testing Regime Over Time)</i>	223
Introduction.....	223
Research Question 3: What Evidence is There of Washback in the New Testing Regime Over Time?	224
Evidence of Washback on Teachers.....	224
Evidence of Washback on Students	242
Chapter Summary.....	256
Chapter 7: Conclusion, Implications, and Future Research Directions.....	258
Introduction.....	258
Stakeholder Accounts and Washback.....	258
Summary of Results.....	261
Limitations of the Study.....	268
Implications and Contributions.....	270
Recommendations for Future Directions	276
References	279
Appendices	293
Appendix A: Research focus and methods of prominent empirical studies on washback	293

Appendix B: Phase 1 - Teacher interview questions.....	296
Appendix C: Phase 1 - Student questionnaire.....	296
Appendix D: Phase 1 – Student focus group questions.....	297
Appendix E: Phase 1- Administrator interview questions	298
Appendix F: Phase 2 - Teacher interview questions set 1	299
Appendix G: Phase 2- Teacher interview questions set 2.....	300
Appendix H: Phase 2 – Student questionnaire	301
Appendix I: Phase 2 – Student focus group questions	302
Appendix J: Phase 2 - Administrator interview questions	302
Appendix K: Phase 3 -Teacher interview questions	303
Appendix L: Phase 3 - Student focus group questions	304
Appendix M: Phase 3 - Administrator interview questions.....	304
Appendix N: Ethics Clearance and Teacher Consent Forms	306
Appendix O: Student Consent Form.....	309
Appendix P: Examples of first and second cycle coding sheets with analytical memos ..	311
Appendix Q: Sample division of categories.....	313
Appendix R: First and second cycle coding with emerging categories and themes	314

List of Tables

Table 1 <i>Tabular summary of Markee's (1997) framework of diffusion of innovation</i>	58
Table 2 <i>Case study tactics for testing of case studies (adapted from Yin, 2014)</i>	84
Table 3 <i>Comparison of former and new test structures</i>	93
Table 4 <i>Listing of interview and focus group participants</i>	98
Table 5 <i>Study instruments</i>	108
Table 6 <i>Sample interview and focus group questions</i>	112
Table 7 <i>Meta-analysis of teachers' interview data</i>	120
Table 8 <i>Educational philosophies and assessment practices of the participating teachers</i>	134
Table 9 <i>Teachers' views on a need to change former testing regime</i>	155
Table 10 <i>Students' views about assessment practices in the former testing regime</i>	159
Table 11 <i>Sample scale for score interpretation on rubric for the reading to write test</i> .	185
Table 12 <i>Student questionnaire results regarding the new ISTs</i>	190
Table 13 <i>Spearman's rho correlation of positive washback of test format</i>	191
Table 14 <i>Summary of the characteristics of the ISTs</i>	193
Table 15 <i>Summary of the characteristics of the resource system in the new testing regime</i>	202
Table 16 <i>Teachers' accounts of changes in classroom practices in the new testing regime</i>	207
Table 17 <i>Student questionnaire results of English learning experiences in their home country</i>	211
Table 18 <i>Student questionnaire results about understanding of learning outcomes</i>	212
Table 19 <i>Student questionnaire results about understanding of referencing and citation</i>	213
Table 20 <i>Summary of the characteristics of the users - teachers and students</i>	215
Table 21 <i>Teachers' accounts of teaching from Phase 1 to Phase 3</i>	227
Table 22 <i>Teachers' attitude towards the new testing regime</i>	230
Table 23 <i>Teachers' accounts of the characteristics of the resource system in Phase 3</i> .	235
Table 24 <i>Comparison of assessments at the EAP program and university assessments</i>	243

List of Figures

<i>Figure 1</i> The Hybrid Model of the Diffusion/Implementation Process (Henrichsen, 1989, p. 80)	65
<i>Figure 2</i> Basic types of designs for case studies (Yin, 2014, p. 50).....	77
<i>Figure 3</i> Overall Research Design.....	100
<i>Figure 4</i> Phase 1- Antecedents and Process Elements from Henrichsen (1989, p. 80). 101	101
<i>Figure 5</i> Phase 2 - Implementation dynamics and immediate washback (Process Phase)	102
<i>Figure 6</i> Quantitative and qualitative data collection and analysis for student questionnaires and focus groups	105
<i>Figure 7</i> Consequences of the importance afforded to "writing skills" in the EAP Program.....	140

List of Abbreviations

Abbreviation	Definition
CAEL	Canadian Academic English Language
CLOs	Course Learning Outcomes
EAP	English for Academic Purposes program
ESL	English as a Second Language
ESP	English for Specific Purposes
ET	Exit Test
GL	Graduating Level of the EAP program
IELTS	International English Language Testing System
ISTs	Integrated Skills Tests
MCQs	Multiple-choice questions
PEEL	Post Entry English Language Support
PLOs	Program Learning Outcomes
TLU	Target Language Use
TOEFL iBT	Test of English as a Foreign Language

Glossary of Key Terms or Concepts

Antecedents: In this dissertation, Phase 1, the conditions in the *Antecedents* phase are those viewed as the status quo within the EAP program under study before the introduction of a new testing regime. This notion aligns with Henrichsen's (1989) hybrid model and refers to the conditions in place in the educational context before an innovation is proposed.

Consequences: In this study, Phase 3, or the *Consequences* phase, involves the effects of change brought about through innovation, whether short term or longer-termed/delayed following a shift in the testing regime. This notion follows Henrichsen's (1989) definition of consequences, i.e., the decisions and outcomes of a change process.

Curricular Innovation: This includes all material and process changes to a curriculum to bring about an intended change within the educational enterprise of a learning program. In any program, key stakeholders view these changes as new.

High-Stakes Tests: High stakes tests are viewed by stakeholders as holding significant power to change the course of student's progress within a program, either negatively or positively. Examples of high-stakes tests for students are proficiency tests for universities in English speaking countries in which failure can mean being barred from entering a desired course of study to achieve a degree.

Graduating Level: In the context of the EAP program under study, Graduating Level is the last level of the program that is geared towards L2 learners who want to pursue degree programs in North American universities.

Innovation: Innovation in education and particularly in the context of curricular change is the bringing about and/or adopting something new into an existing system in order to improve that system in some intentional way. Positive change is the hallmark of innovation – and is typically viewed that way by all or most stakeholders (Henrichsen, 1989; Fullan, 2015; Markee, 1997; Rogers, 2003).

Integrated Skills Tests: Generally speaking, an integrated skills test is the incorporation of several skills within one test to determine whether a student can take on the complexity of a real-world task that requires a multiple skill set. Within the EAP program, integrated skills tests are those in which students are required to “produce written compositions that display appropriate and meaningful uses of and orientations to source evidence, both conceptually (in terms of apprehending, synthesizing, and presenting source ideas) and textually (in terms of stylistic conventions for presenting, citing, and acknowledging sources)”(Cumming, Kantor, Baba, Erdosy, Eouanzoui, and James, 2005, p. 34).

Intended User System: Intended-user system can be described as the educational context where an innovation is introduced. With any system, the various stakeholders, norms, communication styles, protocols, and rules have to be explored to better understand how inhabitants within the system are imagined, constructed, and shaped by these mechanisms. Henrichsen (1989) describes it as the nature of the school system and its structure, especially, “the power hierarchy among the various elements within these systems that affect the course of the diffusion/implementation process” (p. 79).

Negative Washback: Negative washback is the unintended outcomes of an innovation being adopted within a particular system. In a general sense, these are the “side effects”

of an intended adoption of change that have not been contemplated and therefore have not been managed by the stakeholders in any intentional way. Bachman and Palmer (2010) describe negative washback as “‘teaching to the test’ [which] implies doing something . . . that may not be compatible with teachers’ own values and goals, or with the values and goals of the instructional program” (p. 108).

Post-Entry English Language (PEEL) Courses: In this study, PEEL courses are the English language support courses that students can take after completing the Graduating Level of the EAP program. These courses can be taken in conjunction with students’ university courses.

Positive Washback: When understanding a testing regime and planning for the implementation of new tests, one has to be aware of the repercussions, or washback, that flow from testing practice. Tests can have immediate or long-term forms of washback, whether positive or negative. Positive forms of washback are the imagined, intended, planned, or actual actions undertaken by teachers and students, which are linked to the influences of a test or testing practices, and which improve the quality of their teaching and learning. In the context of curricular reform, where the introduction of a new test is viewed as an innovation, because these actions are anticipated by curricular planners, these can be more or less adjusted as needed. If such adjustments improve the quality of teaching and learning, they may be viewed as evidence of positive washback. Cheng (2005) describes positive washback as the feasibility and desirability to bring about “beneficial change in language teaching by changing examinations... [where] teachers and students have a positive attitude towards the test and work willingly towards its objectives” (pp. 29-30).

Process: In this study, Phase 2, the *Process* phase is viewed as the time period when the introduction of the new testing regime and the immediate washback of a given change are first experienced (Henrichsen, 1989; Fullan, 2015). This is usually during the first semester of a program, in this case an EAP program.

Target Language Use (TLU) Domain: This is the language that has been routinized in a way that is meaningful and viewed as “normal” and “expected” within a particular context. Many occupations and learning disciplines have language and language use that is particular to specific contexts. Meanings and definitions are narrowly constructed and viewed to fit a particular routinized way of inhabiting the world. Cheng and Fox (2017) define TLU domain as “certain language use tasks, which inform the design of test tasks and ultimately allow us to generalize from performance on the language test to performance in the TLU domain” (p. 229).

Testing Regime: In this dissertation, a testing regime is defined as a systematic or programmatic approach to the use of assessment practices and procedures that has control over the curriculum and is directed at realizing higher levels of student achievement as measured by high-stakes tests and their results.

Users (or Intended Users) Of Innovation: A system, where any innovation is introduced, is made up of individuals (or stakeholders) who share a number of common characteristics. In the context of the present study, teachers and students are considered as the main intended users. Further, these stakeholders come with a constellation of experiences, backgrounds, goals, and beliefs about learning. The attitudes, values, norms, and abilities of these individuals can influence the course of the implementation process.

Washback: Washback is the sum total of all the intended and unintended effects that flow from a set of testing protocols and affect stakeholders directly through these testing and assessment practices and, in the case of this dissertation, the intended and unintended effects brought about through an integrated assessment innovation. Washback can promote or inhibit learning, affect classroom practices directly or indirectly, and affect the actions of stakeholders because of the influence of high stakes testing (Messick, 1996; Tsagari & Cheng, 2017; Wall, 2012).

Washback Intensity: If a test is high-stakes, then its washback intensity will be strong because the positive or negative repercussions can be quite significant. Examples of high stakes tests include university admission or immigration whereby not doing well can shut down significant life chances at a better life. If a test is low-stakes, its intensity will be weak and the fallout is insignificant with regard to broader life effects. Examples of low stakes testing can include low value classroom tests and quizzes.

CHAPTER 1: Introduction

“...washback is not simply good or bad teaching or learning practice that might occur with or without the test, but rather good or bad practice that is evidentially linked to the introduction and use of the test” (Messick, 1996, p.254).

Researcher Lens

My background as a student in high-stakes testing environments in India and transforming to a new testing environment within the Canadian context piqued my curiosity about the effects of high-stakes testing. Growing up in a testing-dominant education system, I was fearful of assessments and felt that real knowledge could never be expressed in high-stakes tests. When I started teaching English for academic purposes, I was always conscious of not overburdening my students with quizzes and tests, as I wanted learning to be meaningful for them and also to avoid the anxiety and fear that my own testing experience had produced in me.

Two important life events led me to reflect critically on the role of high-stakes English proficiency tests and their influences on teaching, learning and life altering rites of passage such as access to higher education, certification or immigration. Firstly, I became a speaking and writing examiner for the Cambridge Proficiency Test – the International English Language Testing System (IELTS). I came across candidates with high language proficiency and low articulation, fluency, and coherence of thought and, conversely, candidates with low language proficiency, but profound wisdom and articulation of thought. This made me question the role of one-shot, high-stakes tests in measuring language proficiency.

Enrolling for an MA in Applied Linguistics, and later, doctoral studies at Carleton University was another impetus for me to evaluate my assessment beliefs. Here, instead of tests, my course work was assessed differently through assignments and projects. I learnt the differences between task-based learning (such as using projects and assignments as measures of competence) and test taking. Thus, in teaching my own students, I started incorporating more projects and assignments as meaningful measures of student learning.

Teaching in the EAP program was rewarding in the sense that I used all my newly acquired knowledge in assessment to measure my students' language proficiencies. However, when I taught at the graduating, or bridging (the last), level of this EAP program, which is the focus of this dissertation, I had to prepare my students for an external high-stakes English proficiency test – a policy of the testing regime in the program. In my dissertation, I define *testing regime* as a systematic or programmatic approach to the use of assessment practices and procedures that has control over the curriculum and is directed at realizing higher levels of student achievement as measured by high-stakes tests and their results. I saw many international students (graduate and undergraduate) who were ready, in my view as their teacher, for academic studies, but were unable to continue because they failed the high-stakes exit test. This was the impulse that first caused me to think deeply about the power, relevancy, use, and consequences of tests. The testing regime affected my classroom teaching, and my students' expectations for learning in preparation for the external high-stakes test. Other teachers and students in this program, too, voiced their opinions about the deleterious

effects of the dominance of the testing regime. The administration of the EAP program, too, was aware of the problem and decided to implement changes in the testing regime.

Reviewing my own life course and where I find myself today, I am left curious about the effects of testing regime change on teaching and learning. My belief was that testing regime change, if not done with forethought, could have a negative impact on teachers and students, especially if the process is imposed on them. Conversely, if changes were well planned, they could potentially have a positive impact on teaching and learning particularly if teachers and learners were considered important stakeholders and did not have the process imposed on them. These were my initial beliefs and I undertook this dissertation in order to investigate how testing regime change affected my teaching, my learners, my colleagues, and the program within which the study is situated. This dissertation is a longitudinal case study of the implementation of a testing regime change process within an EAP program, which adds to the literature on test washback and the diffusion of innovation. In the sections below, I provide an overview of washback and diffusion of innovation as a prelude to my study.

Washback

A prevailing and widely accepted notion is that tests have important consequences for students, teachers, institutions, and society at large. In the language testing literature, the influences or outcomes of testing on teaching and learning are frequently described as *consequences*, *impact*, and *washback*. Although some researchers distinguish these terms (e.g. Wall, 1997), all broadly refer to the different aspects of the same phenomenon - influences of testing on teaching and learning (Andrews, 2004). To distinguish between consequences, impact, and washback, researchers like McNamara (2000) and Wall

(1997) link the terms ‘consequences’ and ‘impact’ to effects of tests on societal-levels of education, and the term ‘washback’ to the effects of tests on the classroom-level of language teaching and learning. This dissertation uses the term “washback” to refer to the effects of tests on teaching, learning, and other stakeholders in the educational process. Where the term *impact* occurs in the dissertation, it is used in a non-technical sense, as a synonym of “influence.”

The concept of the washback phenomenon is rooted in the notion that “tests or examination can and should drive teaching, and hence learning” (Cheng & Curtis, 2004, p. 4). Cohen (1994) defines washback as “how assessment instruments affect educational practices and beliefs” (p. 41). For Alderson and Wall (1993), washback occurs when students and teachers “do things *they would not necessarily otherwise do* because of the test” (p. 117, emphasis added). For example, certain portions of a syllabus or typical test question formats might be practised more in class by teachers and students because of how often these items will appear on tests causing negative washback.

In the last three decades, interest in researching washback and the impacts of different high-stakes tests has stemmed from different political, practical and theoretical concerns. For example, for political and practical reasons, policy makers may use testing as a means of controlling curricula and to impose new teaching methods or materials in classrooms (Cheng, 1997; Heyneman & Rensom, 1990; Madaus, 1988; Nolan, Haladyna, & Hass, 1992; Shohamy, 1993; Shohamy, Donitsa-Schmidt, & Ferman, 1996). An influence of tests on classroom practices makes high-stakes tests a potential source of influence on curriculum and curricular innovation (Andrews, 2004). For example, the content and structure of tests may define curriculum that can negatively influence teacher

pedagogies and student learning. Shohamy (2001) critiques this negative influence, claiming, “there is evidence that tests are often introduced by those in authority as disciplinary tools, often in covert ways for the purpose of manipulating educational systems and for imposing the agendas of those in authority” (p. 374). Language testing researchers who investigate washback are interested in understanding the implications and impositions of this control on different stakeholders, such as teachers, administrators, and students.

From a theoretical point of view, researchers also consider washback an important step in test validation. For example, Messick (1996) argues that washback is one of the manifestations of the consequential aspect of the validity of inferences that we draw from test scores or performance. He notes, however, that “it is problematic to claim evidence of test washback if a logical or evidential link cannot be forged between the teaching or learning outcomes and the test properties thought to influence them” (p. 247). Of critical importance here, as Messick notes, is the evidence to support the link between tests and their washback.

In the field of language education, educators generally have divided opinions about test washback. For example, Pearson (1988) considers high-stakes tests as ‘levers for change’ to describe the positive influences of exams on curriculum. Similarly, when teaching the curriculum and teaching to the test are same, there is a potential for positive washback (Green, 2007). However, opponents of high-stakes testing (e.g. Madaus, 1988) consider that high-stakes tests negatively influence curriculum and claim that tests narrow the curriculum by encouraging teachers to teach to the test, thus exerting negative washback on teaching and learning. This is of particular importance, according to

Madaus (1988), if a test is high-stakes; i.e., “whose results are seen –rightly or wrongly – by students, teachers, administrators parents, or the general public, as being used to make important decisions that immediately and directly affect them” (p. 87). He further argues:

Measurement-driven instruction invariably leads to cramming; narrows the curriculum; concentrates attention on those skills most amenable to testing... constrains the creativity and spontaneity of teachers and students; and finally demeans the professional judgment of teachers. (p. 85).

Researchers have continued to focus on washback to find out more about the debate regarding the influences of a test to be advantageous or detrimental. Why have researchers focused on washback and engaged in research about washback? How do researchers identify positive, or negative, or neutral washback, particularly when a new language test is being introduced in an educational system? These questions mark the central interests of this dissertation.

The Definition of Washback in This Study

The phenomenon of washback is understood as the influence of “intended direction and function of curriculum change on aspects of teaching and learning by means of change in a public examination” (Cheng, 2005, p. 28). Pearson (1988) calls this intended direction as encouraging the use of beneficial teaching-learning processes, also known as *positive washback*. The ultimate product of positive washback is “the improved learning of the construct being measured (language proficiency in our case)” (Bailey, 1999, p. 11). Alderson & Wall (1993) state that washback is a neutral term, which assumes positive or negative direction depending on the context where it is used. They have also pointed out that washback becomes highly complex when intertwined with

curriculum change and development. Thus, along with the intended washback, there is a possibility of unintended (or side effects) washback, also called *negative washback*, when “teachers tend to ignore subjects and activities which do not contribute directly to passing the exam....excessive coaching for exams” (Alderson & Wall, p. 115). Similarly, Bachman and Palmer (2010) consider negative washback as “teaching to the test” [which] implies doing something . . . that may not be compatible with teachers’ own values and goals, or with the values and goals of the instructional program” (p. 108). I retain the original meaning of washback from these authors when referring to positive and negative washback in my study, namely, that intended washback is viewed as positive and unintended washback as negative and that the context changes washback negatively and positively depending on the interaction between the specific situation and motives of the stakeholders such as teachers.

Washback as Diffusion of Innovation

Cheng and Curtis (2004) suggest that there are two key types of washback studies. The first type concentrates on tests that are already in existence, i.e., traditionally multiple choice, large-scale, high-stakes tests. These tests are frequently considered to generally exert negative washback on teaching and learning (see e.g., Alderson & Wall, 1993; Green, 2007; Nolan, Haladyna, & Hass, 1992; Shepard, 1990). Madaus (1988) claims that such tests “transfer control over the curriculum to the agency which sets or controls the exam” (p. 97). The second key type of washback study has concentrated on a new test or several new tests that have been revised, modified, or improved in order for them to exert a positive influence (see e.g., Andrews et al., 2002; Hughes, 1989; Pearson, 1988; Popham, 1987). However, this second type of study (Alderson and Wall, 1993) of new

tests has also shown not only positive washback, but also negative and/or neutral washback on teaching and learning.

Early commentaries on washback (e.g., Heaton, 1990) considered tests central to the concept of washback. According to Heaton (1990), “If it is a good examination, it will have a useful effect on teaching; if bad, then it will have a damaging effect on teaching” (p. 16). However, following Hughes (1993), researchers (e.g. Wall and Alderson, 1993; Cheng, 1997; Shohamy, Donitsa-Schmidt, & Ferman, 1996) took a more nuanced approach, by researching ‘participants’ (e.g., teachers and students), ‘process’ (e.g., selection of content for teaching/testing), and ‘products’ (e.g., students learning). These studies found that washback is not generated and affected by only one factor (the test), but is in fact more complicated. It is mediated by numerous factors, such as the context, stakes, and personal factors like motivation (Wall & Alderson, 1993; Alderson & Hamp-Lyons, 1996; Messick, 1996).

Though this idea of washback as a multi-faceted phenomenon has become widely accepted, the challenge that has remained for language testers is to categorize these mediating factors into a coherent and user-friendly model (Tsagari & Cheng, 2017; Wall, 2000). In the early 1990s, White (1993) observed a particular gap in the literature that helped push the field forward noting that “Most applied linguistics and testing literature had skimmed over the issue of innovation” (p. 45). Citing Miles (1964, p. 14), Henrichsen (1989) defines innovation as

a species of the genus ‘change’ [It is] a deliberate, novel, specific change, which is thought to be more efficacious in accomplishing the goals of a

system”...In addition, an innovation is usually “willed and planned for, rather than ... occurring haphazardly” (p. 65).

With the exception of a few notable researchers in the language teaching/curriculum literature (e.g. White, 1988; Henrichsen, 1989), there has been little familiarity with “the voluminous literature that already existed in a number of disciplines on how and why innovations diffuse” (Markee, 1993, p. 229). In particular, Wall (1997) pointed out that little attention had been paid to the question of innovative implementation of revised or new tests. Wall (1997) considered introducing a test into an educational system as a form of curricular innovation. Her research bridged the gap between educational and curricular innovation and language testing by using innovation theory to examine the washback of a new or revised test. She recommended using innovation theory to study washback because a study of innovation practice would increase the understanding of *how* and *why* washback emerges. Furthermore, theories of educational innovation can be useful not only to study test washback, but also for language test developers and implementers to evaluate *how* or *why* their innovations were successful or not. Use of innovation theory to study washback can help identify factors in an educational context, which facilitate or hinder the successful implementation of curricular changes that have the intentions of producing positive or intended washback.

Today, washback research is ample: studies have found evidence of both positive and negative washback from many high-stakes tests. Still, the washback research agenda is far from exhausted. Some researchers (Andrews, 2004; Cheng, 1997; Wall, 2012) have suggested new lines of inquiry, including investigating washback as a more complex phenomenon. For example, the effects of tests as innovation can be positive or negative.

These influences can also be “immediate or delayed, direct or indirect, or apparent or not visible –e.g. changes in attitude that do not manifest themselves on overt behavior” (Henrichsen, 1989, p. 80). Alderson (2004) is concerned that curriculum reform and innovation are very complex and washback studies of revised, modified, or improved tests should not only take into account the context into which the change or innovation is being introduced, but also “all the myriad forces that can both enhance and hinder the implementation of the intended change” (p. xi). Wall (2012), meanwhile, suggests that there are methodological challenges to be addressed when investigating washback because changes in classroom practices can arise from other educational contextual factors than a test, especially when a new or modified test is introduced alongside or as part of curriculum reform.

Shifts in the literature on washback have left two significant gaps relating to the process of washback: First, it is clear that it is insufficient to know that washback exists; *how* it takes place is of equal importance. Second, most previous research on washback has been cross-sectional, relating to one test at one time. However, longitudinal washback research is needed to understand what happens when a new testing regime replaces an old testing regime and how washback develops at different points in time in a particular context.

The current dissertation proposes to address these two research gaps relating to the processes of washback. First, the use of innovation theory can help in understanding the interactions between a test, the curriculum on which it is based, and factors such as the characteristics of an educational context before and after the introduction of a new test intended to produce positive washback.

Second, most previous research on washback has been cross-sectional, relating to one test at one time. For a fine-grained portrayal of washback-as-process, this dissertation proposes that it is essential to understand how washback can vary over time. Thus, longitudinal washback research is needed to understand what happens when a new testing regime replaces an old testing regime, how washback develops over time in a given context and what factors are most influential in its development. Without understanding these features it is hard to say how washback has occurred and how an “evidential link” (Messick, 1996, p.247) can be established between teaching, learning, and testing.

A proposed step in addressing these gaps is an examination of a context where a new test is introduced. As Weir and Roberts (1994) point out: “We need to establish what the conditions are before a ‘treatment,’ which will help us to monitor any effects that occur during or after ‘treatment’” (p. 46). Wall and Horak (2006), likewise, suggest establishing the situation *before* an innovation is introduced, so that the intended effects can later be matched with the actual effects. In the case of this dissertation, these ideas are applied to study the former testing regime in an EAP program before the introduction of a new testing regime and to examine how the change in testing regime (i.e. innovation) plays out over time. I use the term “testing regime” with the view that such a regime is an evaluation system of power and control. I appropriate the term from Broadfoot (2005) and Fulcher (2009) who see testing regimes as centrally controlled, standards-based education systems with their core purpose being hyper-accountability. Although they use this term for the broader context of evaluating the impact of the national examinations in England, the context in my study was narrower. Their notion of “testing regime” evokes a profoundly powerful washback notion of testing. The final test in the EAP program

considered here was a powerful, high-stakes test involved in a system that exerted power and control over teachers and learners for the purposes of accountability and standardization.

The EAP program, the research site of this study, caters to university-bound adult L2 learners who want to pursue degree programs in a North American university. This program recently introduced Integrated Skills Tests (ISTs), replacing the traditional multiple-choice reading and listening tests as the exit criteria for the graduating students of the program. These tests were high-stakes in nature because the results determined whether students could enter their degree programs, or not, at this university. Until recently, to pass the EAP program, students at the graduating level (GL) had to take an external-to-the-program high-stakes ExitTest (ET), provided by the university's testing office. The ET was comprised of multiple-choice reading and listening tests and writing a decontextualized five-paragraph essay. Speaking was not formally assessed. Results of this test, in conjunction with classroom marks, decided three possible outcomes for the students: 1) If students achieved a final grade above 75%, they successfully commenced their full-time academic studies in undergraduate or graduate courses; 2) If their final grade was between 70-75%, depending on their final exam writing samples, students had to take one or two Post-Entry English Language (PEEL) courses in conjunction with their university courses; and 3) If students received a grade of less than 70%, they had to repeat the GL, which meant there would be a further delay in starting their university studies.

During the Fall 2016 term, many curricular changes were made to the EAP program. One major change was in relation to the testing regime in the program. Under

the new regime, instead of separate multiple-choice reading and listening tests, the final tests took the form of ISTs where students read and listened to a text, and then wrote an essay based on the knowledge gained from the texts. In addition, teachers in the program were to begin formally testing speaking skills. These ISTs were developed and delivered in-house, within the EAP program, instead of the external Testing Office of the university. However, there was no change in the exit criteria for the GL students. Their final exam writing performances were still reviewed by the Testing Office of the university and, similar to the former regime, if students received an aggregate grade of 75%, they could start their full time academic studies without any PEEL courses; if their grades were between 70-75%, students were required to take one or two PEEL courses, depending on the suggestions by the Testing Office; and upon failure to score at least 70%, students were required to repeat the GL. The case study reported here examined whether the change in the high-stakes test within this EAP program and its testing regime, resulted in positive washback (which was the intention of the change itself).

For studying a change process, Fullan (2007) provides support by suggesting two aspects to any educational change: *what* change to implement (theories of education) and *how* to implement the change (theories of change). Put differently, Fullan states, “we have to understand *both* the change and the change process” (p. 40), as these two aspects of change process interact, shape, and define each other. Therefore, to examine the phenomenon of washback of e.g., a newly introduced test (as in the case of the present research), the distinction is made between the ‘what’ and the ‘how’ of change in reform efforts, and both must be examined. To do so may involve paying attention to such things as the school context where innovation resides (Fullan, 2015).

Fullan (2015) also argues that the meaning of change can be elusive if a ‘shared meaning’ is not established. In other words, it is important to understand the meanings that different stakeholders attribute to changes to both a testing regime and to the change process, i.e., how the new testing regime was implemented. After all, “success is not just about being right; it is about engaging diverse individuals and groups who are likely to have different versions about what is right and wrong” (Fullan, 2007, p. 40). Therefore, to address the research gaps, just mentioned, this dissertation project undertakes a longitudinal, in-depth description of test washback over time from the viewpoints of various stakeholders involved in the process of change. More specifically, it provides a comprehensive description of the phenomenon of washback in an EAP program using a longitudinal case study design (Yin, 2014) and a conceptual framework grounded in innovation theory.

Overall Purpose of the Study

As discussed earlier, numerous factors (e.g., the context, the stakes, and personal factors) mediate washback; therefore, the main purpose of the present longitudinal case study is to highlight some of these factors when a new testing regime is introduced in an EAP program. Particularly, it considers how these factors facilitate or hinder the diffusion of innovation and promote positive (or negative/neutral) washback on the teaching and learning in the program.

The incorporation of ISTs was considered a positive innovation in the EAP program, and their implementation was seen as a source of potentially beneficial washback because the new tests were considered to be more pedagogically advantageous than multiple-choice tests and five-paragraph essays. Therefore, this particular EAP

program provided a rich context to investigate washback from a new test on teaching and learning, and the overall purpose of this study is to examine whether or not a high-stakes test can be used to leverage positive washback on teaching and learning as part of curriculum innovation in the EAP program under study.

Research findings regarding educational innovation across different contexts suggest that for innovation to take hold it takes time, even years, because adopters need time to understand and manage change (Cheng, 2005; Fullan, 2007, 2015; Henrichsen, 1989; Markee, 1997; Wall, 2000, 2005). Examining the influence of ISTs was, therefore, meaningful not only in the early stages of their implementation, but also several semesters after they were introduced, beyond the EAP program when participating students in this study were in their regular university courses and teachers were accustomed to the new testing regime. This longitudinal approach helped to explain whether any of the features present in the ‘after’ period of test introduction can be “evidentially linked” (Messick, 1996, p. 246) to the introduction of the new test.

Most English as a Second/Foreign Language (ESL/EFL)-related washback studies about innovation have taken place outside the North American context, especially in China and the Asian subcontinent (see Alderson & Wall, 1993; Andrews et al., 2002; Cheng, 1997, 2005). There are very few studies about innovation in EAP programs in the North American context. A notable exception is Stoller’s (1994) study about the analysis of innovation in a select higher education, intensive English program. Stoller included administrators’ perceptions, but not other stakeholders’, such as teachers and students. Cheng and Fox’s (2013) “Review of the doctoral research in language assessment in Canada (2006-2011)” does not cite a single doctoral study about innovation in English for

Academic Purposes (EAP) programs in Canada despite the popularity of the academic preparatory courses, commonly known as Pathways¹, in most Canadian universities. This dissertation represents a type of washback research that is lacking in North America generally, and in research done by doctoral students specifically.

Finally, washback studies related to innovation, such as Wall (1997, 2005) and Wall and Horak (2006, 2008, 2011), are related to standardized, high-stakes language proficiency tests that are external to EAP programs, such as the Test of English as a Foreign Language (TOEFL). The uniqueness of the current study also resides in the exploration of the washback related to standardized high-stakes proficiency tests (e.g. ExitTest used in former testing regime) as well as classroom assessment practices in the EAP program.

Specific Research Goals

The first goal of this study is to explore the washback of an innovation in a testing regime using Henrichsen's (1989) Linkage Model of Diffusion/Implementation of Innovation. This model has previously been used in impact studies of test reforms in specific educational settings (Wall & Horak, 2006, 2011; Wall, 1997, 2005).

Another goal of this study is to investigate how intended and unintended washback occurs, i.e. unforeseen or unplanned consequences of tests and testing. The potential advantage of examining the processes of washback over time is to learn how dynamics of implementation processes can be made more effective in future, leading to more meaningful outcomes.

¹ According to Illuminate's 2015 report, there are approximately 80 pathways programs (academic ESL, EAP, Foundation, and Award (2nd year Entry)) for international students run by various universities, colleges and private providers across Canada.

The final goal of the study is to examine the factors that can help or hinder the diffusion or implementation of an innovation (Henrichsen, 1989) by examining the accounts of key stakeholders (e.g., teachers, students, and administrators). In order to address the overall purpose of this study, i.e. can a testing regime change be used as innovation to engender positive washback, this study will address the following research questions:

Research Questions

1. What evidence is there of washback in the former testing regime and what is the intended washback of the new testing regime?
2. What evidence is there of washback factors facilitating and/or impeding the implementation of the new testing regime?
3. What evidence is there of washback in the new testing regime over time?

Each of these broad research questions guides one or more phases of this study. Each question is addressed as outlined in Chapter 3 and discussed in Chapters 4, 5, and 6.

Significance of the Study

This study contributes to the literature about washback potential as a result of the implementation of a new test and explores the value of innovation theory in accounting for changes within the context of an EAP program. Using diffusion of innovation theories (Fullan, 2015; Henrichsen, 1989; Markee, 1997), this study describes positive and negative contextual features previously unaddressed in the literature on washback from the point of view of diffusion of innovation.

Methodologically, the study is one of only a few (Cheng, 2005; Wall, 2005) to employ a longitudinal case study methodology that incorporates both qualitative and quantitative data to yield evidence of washback. Ultimately, this dissertation has practical implications for EAP programs, EAP test developers, EAP policy-makers, EAP teachers, language testers and researchers who are concerned about washback from high-stakes tests in understanding the role of coordination, collaboration and communication in creating positive or negative washback during the process of diffusion of innovation of using tests as a means for making curricular changes.

Organization of the Dissertation

This first chapter has introduced the phenomenon of *washback* and has outlined the research questions, research context, purpose, and significance of the study. This chapter has also made a case for the study of the diffusion and implementation of innovation when a test is used to leverage positive change in teaching and learning.

Chapter 2 reviews relevant literature on the phenomenon of washback and its complexity, as well as on issues of measuring academic English skills. The present study examines the phenomenon of washback through the lens of innovation theories, such as Fullan's (2007, 2015) views of change and House's (1981) critical treatment of educational innovation. These and several other theories will be discussed in Chapter 2 including: Hughes's (1993) Model of Washback on Participants, Process, and Participants (PPP); Markee's (1997) Curricular Innovation Model; and Henrichsen's (1989) Hybrid Model of Diffusion/Implementation of Innovation.

Case study research methodology is introduced in Chapter 3 along with a description of the study's participants, data collection instruments, and data analysis.

Chapters 4, 5, and 6 present and discuss the results in relation to the three phases of the research. The final chapter, Chapter 7, summarizes the previous three chapters and discusses the usefulness of a diffusion of innovation perspective in studying washback, the limitations and implications of the study, and directions for future research.

Chapter 2: Washback and Innovation

Introduction

Chapter 1 provided a brief overview of the dissertation. It also briefly described the EAP program under study and the introduction of a new testing regime (i.e., the ISTs) in the program. The ISTs were considered high-stakes as results of these tests had serious implications for students in the program. Therefore, it is a reasonable expectation that the ISTs would affect teachers and students. *How* and *why* these influences occur can be informed by theories of innovation and research on test washback. Therefore, the focus of this chapter is a review of the literature on the phenomenon of washback and its complexity, and the curricular literature on the diffusion of innovation.

Chapter 2 begins with a review of tests and test purposes and theoretical foundations and academic constructs in second-language testing. Then, the concept of washback is discussed with an emphasis on the mechanisms (i.e. how washback works). A discussion of the theories and models of diffusion of innovation follows. More specifically, the conceptual and theoretical perspectives informing the study include:

1. Models of language performance within the discipline of language testing (Bachman, 1990, revised in Bachman & Palmer, 1996 and 2010).
2. Washback: its mechanisms, dimensions and complexity (Alderson & Wall, 1993; Bailey, 1996; Cheng, 1997; Hughes, 1993, 2003).
3. Theoretical models for understanding the change process (House, 1981; Fullan, 2007, 2015; Markee, 1993, 1997).
4. Model of diffusion of innovation in English language teaching (Henrichsen, 1989).

I begin by considering language tests in general and issues with assessing academic language proficiency specifically.

Language Tests and Assessing Academic Language Proficiency

In the education literature, researchers generally differentiate among different terms such as ‘tests’, ‘assessment’, ‘measurement’, ‘examination’ and ‘evaluation’. However, Bachman and Palmer (2010) suggest all these terms usually refer to the same activity: “the process of collecting information about something that we are interested in, according to procedures that are systematic and substantively grounded” (p. 20).

In the language testing literature, one of the confounding issues is the use of language as both the means and the end of measuring proficiency. In other words, language is both the object and instrument of measurement (Bachman, 1990). Therefore, it is sometimes problematic to understand the relationships and roles of ‘language abilities’, ‘language tasks’, ‘learning contexts’ and their ‘interactions’ especially for the purpose of constructing language tests (Bachman, 2007). Various language-testing researchers have tried to capture and interpret these language-related concepts from different perspectives (Canale & Swain, 1980; Chalhoub-Deville, 1997, 2003). As a result, over time, different models of testing language proficiency (from testing knowledge about language to testing abilities in language user-in context) have evolved to discuss these concepts in different terms with slightly different meanings (Bachman, 1990; Bachman & Palmer, 1996, 2010; Canale & Swain, 1980; Chalhoub-Deville, 1997, 2003; Kramsch, 1986; McNamara, 1996). However, Pawlikowska-Smith (2002) is of the view that most models of testing communicative proficiency of second language are either a reconceptualization of Bachman (1990), and Bachman and Palmer (1996, 2010)

or “updated versions of the classic Canale and Swain (1980) and Canale (1983) models of communicative competence, and all owe much to Hymes (1971) and his concept of communicative competence” (p. 9). Most models of language assessment perceive language ability as a product of interactions between different components, e.g. linguistic, grammatical, sociocultural, and pragmatic. Canale and Swain (1980) call the interaction of the above components strategic competence. Bachman and Palmer (2010) also use the term *strategic competence*, but they use it to refer to a set of metacognitive strategies that are involved in “planning, monitoring and evaluating individual’s problem solving” (p. 49).

For Bachman and Palmer (2010), viewing language use in terms of separate skills is reductive and not the best basis for thinking about language-use abilities. They suggest that an ability to use language in different ways and settings varies within the same individual across tasks and with other language users. As a test taker uses language, all these attributes and the test taker’s strategic competence (see above) interact to affect the test taker’s performance on the language task.

For the purpose of this research, the most useful model of language testing is Bachman and Palmer’s (1996, 2010) model of assessing communicative language ability because the model describes the test tasks; individual characteristics of test-takers; and the interaction of test tasks with test-takers. In this study, Bachman and Palmer’s model will be useful to: a) interpret the old and new tests from the perspective of teachers, test-takers, and administrators; and b) examine how individual characteristics of test-takers influence their interaction with the old and new tests.

Test purposes. Second language tests and assessments can generally be divided into two types depending on their roles, purposes, uses and the contexts: internally mandated and externally mandated (Davidson & Lynch, 2002). Internally mandated tests are used for placement, achievement and diagnosing difficulties in individual learners. They are related to “the needs of the teachers and learners working within a particular context and ... are generally ecologically sensitive” (Fulcher, 2010, pp. 1-2). These internally mandated purposes are also sometimes referred to as *formative* (i.e., assessments *for* learning or classroom-based assessments), where such tests are an essential component of classroom work used to inform teaching and learning and ultimately to raise standards of achievements of students (Black & Wiliam, 1998). These tests, usually held at the end of a unit of study or a course, are generally used as tools to help teachers and student (or sometimes administrators) to better understand their learning.

On the other hand, externally mandated tests come from outside the local context. For externally mandated tests, decisions to test are made by a group of people who often “do not know a great deal about the local learning ecology [context], and probably don’t even know the teachers and learners who will have to cope with the required testing regime” (Fulcher, 2010, p. 2). Such tests (e.g., General Certificate of Secondary Education (GCSE) examinations in England) are also called *summative* (or assessment *of* learning). Based on the results of this type of test, policymakers or other stakeholders make judgments about proficiency/achievement at the end of a study period and learners are expected to have reached a particular standard. Fulcher (2010) explains:

The motivations for external mandates may also appear extremely vague and complex; indeed, policy makers often do not clearly articulate the purpose of the required testing, but it serves a very different function from internally mandated tests. External tests are primarily designed to measure the proficiency of learners without reference to the context in which they are learning (Fulcher, 2010, p. 3).

Such tests can typically be *high-stakes* because they can either determine the immediate future of a test taker or decide the longer-term prospect of each test taker (Madaus, 1988). These tests are generally used to make decisions as to “whether learners can communicate with people outside their immediate environment, in unfamiliar places, engaging in tasks that have not been directly modeled in the test itself” (Fulcher, 2010, p. 3). Although it is useful to classify tests into different categories, tests are essentially used to collect information for making decisions and these decisions have serious consequences for all stakeholders: individuals, programs and society at large (see Bachman & Palmer, 2010; Bailey, 1996; Cheng, 1997, 2005; Hughes, 2003; Messick, 1996; Shohamy, 2001; Wall, 2005).

Tests of English for academic purposes. Measuring language proficiency of university-bound international students is now a well-established practice (Chalhoub-Deville & Turner, 2000; Fox, 2001; Xi, Bridgeman & Wendler, 2014), and university applicants are generally given two types of tests: ‘placement and admission’ (Xi et al, 2014). While placement tests determine if students need special resources (such as English or subject-specific support), admission tests determine whether the applicants have a “requisite level of certain knowledge, skills and abilities (such as English

proficiency) deemed necessary for success” (Xi et al, 2014, p. 1). These tests can also be useful as diagnostic tools; if it is determined that a student needs special help in the English language, they are generally placed at the appropriate level in an English language support program at the host university (Fox, 2009).

However, the literature on EAP assessment has also suggested that using external standardized proficiency tests is problematic in deciding the placement as well as appropriate exit criteria and/or concomitant academic readiness of students from pre-sessional EAP courses (Banerjee & Wall, 2003; Elder, 2017; Fox, 2009; Green, 2007; Moor & Morton, 2005). Banerjee and Wall (2003), reporting on their survey of UK institutions state that while some institutions accept results of external tests as an exit criteria, where students are required to re-take the same standardized tests such as TOEFL and IELTS (which were used for placement) at the end of their pre-sessional EAP courses, other institutions design their own tests “which may or may not be modeled on tests created elsewhere for other purposes” (Banerjee and Wall, 2003, p. 51). Yet, other ESL/EAP institutions may have an additional goal of preparing students for high-stakes tests such as TOEFL and IELTS. For example, Moor and Morton’s (2005) study of some Australian universities found that instructors designed their EAP classes to help learners prepare for high-stakes proficiency tests. Furthermore, other programs may prefer to judge students “on their in-course performance, combining internal tests scores, performance on written assignments, formal presentations and classroom participation” (Banerjee and Wall, 2003, p. 51). Finally, there are some institutions where EAP teachers use ‘benchmarks’ or ‘can-do’ scales to indicate if students are able to achieve certain performance outcomes or not.

Two reasons can be attributed to this variation in the exit criteria: firstly, the complexity of the relationship between English proficiency and academic success; and secondly, the description of academic language abilities in target language use (TLU) and the construct of academic assessment (Cheng, Myles & Curtis, 2004; Cumming, 2013; Fox, 2001, 2004; Fox, Cheng & Zumbo, 2014; Hamp-Lyons & Kroll, 1996; Stoyhoff, 2009; Weigle, 2002). For instance, Xi et al (2014) suggest:

the academic domain description is based on the warrant (or generally held belief) that the observations on the language test represent relevant knowledge, skills, and abilities for use in the target domain of academic discourse in an English-medium university. Support for this warrant should show the link between the critical language tasks and skills in the target use domain and the observations of performance (tasks) on the test (p. 8).

However, to understand the measurement of language use in academic domain, first it is important to know what academic language proficiency is. Davies (2007) suggests academic language proficiency is:

the language of argument, of analysis, and of explanation and reporting,..... it is skilled literacy and ability to move easily across skills it is the literacy of the educated, based on the construct of there being a general language factor relevant to all those entering higher education, whatever specialist subject(s) they will study (p. 85).

Thus, Davies defines academic language in terms of a particular discourse requiring inferences, and not just vocabulary or separate skills. For assessing academic skills, Weigle and Malone (2016), in their article on assessment of English for academic

purposes suggest that there are three main differences between EAP tests and general proficiency tests of English: a) content or topic, b) the nature of language used in assessment, and c) the nature of test tasks which reflect academic settings.

As mentioned earlier, tests - whether in the field of general education or language testing - have played different roles for assessment and accountability purposes (Fulcher, 2010; Linn, 2000; Resnick & Schantz, 2017). Historically, parallel advancements have taken place in the fields of educational measurement and language testing. For example, during the 1950s, tests were mainly used for tracking and selection because of the limited chances of getting into higher education; in the 1960s, test purposes changed to program accountability because of the great expansion of public sectors to control the quality of the growing educational programs (Linn, 2000). These tests were mainly influenced by measurement theory, and the emphasis was on the reliability of scoring and minimizing measurement error (Behizadeh & Engelhard, 2011, as cited in Weigle & Malone, 2016). Weigle and Malone (2016) further state that influenced by these measurement theories in the 1950s and 1960s, scholars (e.g. Carroll, 1961; Lado, 1961) used tests using discrete items (mainly in the form of MCQs) to test language knowledge. These tests were also consistent with the dominant pedagogical tradition of grammar-translation or audio-lingual methods prevalent at that time.

From the 1970s onwards, while in the field of education the idea of minimum-competency tests that focused along the lower ends of the achievement-spectrum started emerging (Linn, 2000), the trends in language teaching started shifting towards more communicative language teaching. In communicative teaching, the goal of instruction is not the mastery of discrete skills, but developing the overall communicative ability of the

learners (Bachman, 1990; Canale & Swain, 1980). Language ability was considered to include both linguistic and pragmatic knowledge as modeled by Bachman and Canale and Swain. These communicative models, suggest Weigle and Malone (2016), influenced “a new emphasis on analyzing real-world language use situations that an examinee would be likely to encounter in school and other language use contexts” (p. 662). There was also a shift in testing from discrete items to other parallel assessment methods called ‘integrative tests’ (e.g., Oller, 1979, as cited in Plakans, 2013). Plakans (2013) suggests that the ‘integrative’ tests aimed to assess a grammatically driven unitary language-proficiency construct and one popular format was the cloze test. In cloze tests, a certain portion of text is removed, either selectively (e.g., articles, or verbs) or mechanically (every nth word) and the test taker is asked to choose the removed word from the given multiple-choice options. This unitary construct of language ability “competed with the idea that it is valuable to separate language ability into component skills for assessment practices” (Plakans, 2013, p. 2).

With the increase of international students at universities, many standardized tests such as TOEFL and IELTS were developed aiming at testing academic language. However, since the 1980s, these tests have also undergone major revisions in the content, format and delivery modes because of test-user demands and advancements in the theories and practices of language learning, teaching, and testing (Xi et al, 2014).

Also, until the 1980s, language testers divided language competency into four separate skills: reading, writing, listening and speaking. Increasing acknowledgement of the limitations of this four-skill view of language assessment led to new test formats such as ‘assessment of integrated skills’, in which the examinees are required to read/listen to a

text and then speak/write about it are also encouraged (Chapelle & Plakans, 2013). Closer to home, in the early 1990s, the Canadian Academic English Language (CAEL) assessment was developed to represent the construct of EAP (Fox, 2001, 2009). According to Plakans (2013), the conceptualization of language as four skills, plus grammar and vocabulary, has been a mainstay in language testing for some time, and:

The current trend of integrated skills assessment may on the surface appear to return to the earlier notion of integrative assessment but in fact the construct assessed and the format of the test tasks for integrated tests today are very different from those of integrative testing of the past (p. 1).

Assessment of integrated skills refers to the use of test tasks that combine two or more language skills to simulate authentic language-use situations (Plakans, 2013, p. 1). Cumming et al. (2005) describe integrated tasks as those in which test-takers are required to “produce written compositions that display appropriate and meaningful uses of and orientations to source evidence, both conceptually (in terms of apprehending, synthesizing, and presenting source ideas) and textually (in terms of stylistic conventions for presenting, citing, and acknowledging sources)” (p. 34). According to Cumming (2014) there are five advantages of integrated skills tests, namely:

...provide realistic, challenging literacy activities; engage examinees in writing that is responsible to specific ideas and content; counter test method or practice effects associated with conventional item types on writing tests; evaluate language abilities in accordance with construction integration models of literacy; and offer diagnostic value for instruction or self-assessment (p. 5).

Furthermore, Chapelle and Plakans (2013) suggest these integrated tests are important avenues in tests of academic language ability “whose scores are intended to reflect how well examinees’ language will allow them to perform on academic tasks, which typically involve a combination of skills” (p. 2).

Even with the use of integrated skills, one of the most confounding issues, however, in assessing academic language is the definition and description of the language to be tested, also called ‘the construct’ (Bachman, 2007; Douglas, 2013; Fox, 2001). According to Douglas (2013), the academic language construct is defined in terms of academic situations rather than linguistic features of a language. There are constraints related to these academic language constructs. Weigle and Malone (2016) describe some of these constraints:

1. The demands for academic language differ for different academic settings; these differences include age of the test-takers (high-school students vs. graduate students), purpose (writing technical report vs. writing general essays) and specific demands of a context (English as Second Language vs. English for Specific Purposes), so academic language must be defined within a specific setting. Also, there are constraints based on sociolinguistic demands across these settings either in written or spoken interactions, for example, writing professional emails or interacting with colleagues.
2. Academic constructs are not only written, but spoken as well. Students need to have a command of both in order to excel.
3. Academic demands at the university level include both specific and general academic contexts (Weigle & Malone, 2016). While examples of specific contexts

include reading and writing assignments, as well as materials and instructions related to a particular course, general contexts include reading and writing involved in a variety of texts beyond a specific course.

4. Finally, the academic construct, itself, poses several challenges. Any test for academic purposes can only measure a portion of all the academic skills mentioned above. “A standardized test of any kind cannot accurately evaluate all the idiosyncratic linguistic requirements of any particular academic setting, nor can it reflect the interaction inherent in spoken and written academic environments” (Weigle & Malone, 2016, p. 663). Furthermore, because these tests do not generally measure the subject area content, they are not necessarily true reflections of a student’s ability to be successful in his/her university studies.

Thus, it can be concluded that the construct of academic language is complicated (Elder, 2017; Fox, 2001; Jamieson, 2014; Weigle & Malone, 2016) and Fox (2001) rightly suggests, “Constructs are neither stable nor permanent. They are perpetually evolving in response to research and theory generation” (p. 3). Elder (2017) further reiterates:

There are still uncertainties about how best to define and capture the academic language proficiency construct for testing purposes, in ways which can serve highly diverse student populations in contexts which are increasingly internationalized (p. 271).

Although tests have changed and will continue to change (as just stated), the level of influence of test content, format and design on teaching and learning is still unresolved in the washback literature. The washback literature has suggested that it is not necessary

that newly introduced tests will definitely achieve beneficial washback. Wall and Horak (2006) in their impact study about introducing speaking and integrated tasks on the TOEFL iBT in many European countries pointed out:

...given what is known from a decade of research into impact and washback, it is no longer logical to predict that because there will be speaking and integrated tasks on the new TOEFL test there will automatically be adequate and beneficial practice of these skills in the classroom. (Wall & Horak, 2006, p. 121).

After the discussion of tests and their constructs, how tests influence different stakeholders in different settings will be discussed in the next section, which focuses on the phenomenon of washback and its complexity.

The Phenomenon of Washback and Its Complexity

As mentioned earlier, the phenomenon of *washback* has been defined and described differently by various researchers. In simple terms, it can be defined as “the influence of language testing on teaching and learning” (Cheng, Watanabe & Curtis, 2004, p. xiii). Cohen (1994) defines it more broadly as “how assessment instruments affect educational practices and beliefs” (p. 41). Messick (1996) explains washback as “the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning” (p. 241); and finally Bachman and Palmer (2010) regard washback as “the effects of assessment use on teachers’ instructional practice...[and] the broad effects of an assessment on learning and instruction in an educational system” (p.109).

Most definitions of washback, just described, have originated from the seminal work of Alderson and Wall (1993): ‘Does washback exit?’ Alderson and Wall argue on the basis of their empirical research, that both test developers and test users cannot simply assume that tests will have an effect on teaching, they must investigate the specific areas, direction, and extent of these presumed effects. By area, Alderson and Wall mean the content of teaching, teaching methodology, teaching materials, methods of assessment etc., and by direction they mean washback can be neutral, have a positive influence, or a negative influence on teaching and learning.

Although the phenomenon of washback is described differently from different perspectives, most researchers agree on one point: studies related to washback should focus on the influence of the test on teachers, learners and their classroom practices rather than the test itself (See Appendix A for method and research focus of prominent empirical studies on various language tests’ washback in the last three decades). The concept of washback is generally related to the results of high-stakes tests. This study investigated two tests – the previous ExitTest and the new ISTs – which were both viewed as high-stakes by their test-takers.

Mechanisms of washback. Although there is a great deal of interest in investigating the phenomenon of washback, Tzagari and Cheng (2017) suggest that “researchers in the field of language education continue to wrestle with the nature of washback” (p. 366). Different researchers have put forth various models and hypotheses to illustrate the workings or the mechanisms of washback. Arguably, the most influential among these are Alderson and Wall’s (1993) hypotheses and Hughes’ (1993) model, which clarify the mechanism of washback.

Alderson and Wall (1993) have suggested 15 possible hypotheses that describe the relationship between tests, teaching and learning as a result of their extensive research. With these hypotheses, they hoped that there would be “the eventual refinement of the washback construct in empirical investigations” (Bailey, 1999, p.6). These hypotheses are:

1. A test will influence teaching.
2. A test will influence learning.
3. A test will influence **what** teachers teach; and
4. A test will influence **how** teachers teach; and by extension from (2) above,
5. A test will influence **what** learners learn; and
6. A test will influence **how** learners learn.
7. A test will influence the **rate** and **sequence** of teaching; and
8. A test will influence the **rate** and **sequence** of learning.
9. A test will influence the **degree** and **depth** of teaching; and
10. A test will influence the **degree** and **depth** of learning.
11. A test will influence attitudes to the content, method, etc. of teaching and learning.
12. Tests that have important consequences will have washback; and conversely,
13. Tests that do not have important consequences will have no washback.
14. Tests will have washback on **all** learners and teachers.
15. Tests will have washback effects for **some** learners and **some** teachers, but not for others (pp. 120-121).

The work of Alderson and Wall (1993) helped in developing the constructs of future washback studies. Their work assumed a dichotomous relationship between teachers and students through a dichotomous process of teaching and learning in examining washback. Advancing on their work, Hughes (1993) suggested a tripartite relationship among participants, process and products in teaching and learning, stressing that “all three may be affected by the nature of a test” (p. 2). Besides teachers and

students, Hughes added other participants, such as material writers, curriculum designers and researchers, improving and adding to Alderson and Wall's hypothesis. He adds:

The trichotomy into participants, process and product allows us to construct a basic model of backwash [or washback]. The nature of a test may first affect the perceptions and attitudes of the participants towards their teaching and learning tasks. These perceptions and attitudes in turn may affect what the participants do in carrying out their work (process), including practicing the kind of items that are to be found in the test, which will affect the learning outcomes, the product of that work (Hughes, 1993, p. 2).

As mentioned in Chapter 1, the ultimate product of beneficial washback is “the improved learning of the construct being measured” (Hughes, 1993 as cited in Bailey, 1999, p. 11). Adding to Hughes' model, Bailey (1996) suggests not all of the participants' processes lead directly to learning. Other researchers (e.g. Green, 2007; Shih, 2007; Watanabe, 1997, 2004) too have developed various models to investigate washback's complex dimensions (see next section). These dimensions can mediate the process of washback and other relationships among testing outcomes and various processes of teaching and learning.

Conceptual Dimensions of Washback

Positive and negative washback directions. Most washback studies are about either the advantageous or detrimental effects of a test in a particular context. These attributes are generally termed as *positive* and *negative* washback directions, respectively. These directions are also described in different terminologies by different researchers.

For example, Bachman and Palmer (1996, 2010) call ‘intentional’ and ‘planned’ effects as ‘positive’ and unintended effects as either positive or ‘negative’. Similarly, Hughes (1989, 2003) and Buck (1988) call positive effects as ‘beneficial’, and negative effects as ‘harmful.’

Researchers have posited *positive* or *negative* washback as either residing in the test design or with the test users (Bailey, 1996; Green, 2007; Messick, 1996; Watanabe, 2004). For example, Bailey (1996) is of the view that any test, whether good or bad, can have either positive or negative washback depending on whether “it impedes or promotes the accomplishment of educational goals held by learners and/or program personnel” (p. 268). On the other hand, Green (2007) suggests:

The key relationship determining the direction of washback is not that between test and curriculum, but that between both test and curriculum and the construct to which they are directed, the better a test represents target skills (whether these are based on a specified curriculum, or a target domain), through content, complexity, format, scoring procedures and score interpretation, the more beneficial washback effect is predicted to be (Messick, 1996). Arguments over washback direction are, at root, variations on arguments over construct definition (p. 13).

Bailey (1996) recommends that to promote positive washback and to reduce tension among pedagogical and ethical decisions, it is useful to incorporate language learning goals, authenticity, learner autonomy and self-assessment, and detailed score reporting in the assessment process. Similarly, to promote beneficial washback, Hughes (2003) has suggested to:

...test the abilities whose development is desirable; use direct testing; make test criteria-referenced; base achievement tests on objectives; ensure that the test is known and understood by students and teachers; and where necessary provide assistance to teachers (pp. 53-56).

Messick (1996) is of the view that promoting positive washback and reducing negative washback can be achieved by minimizing two threats to construct validity: construct under-representation and construct irrelevance. Construct under-representation, as defined by Messick occurs when “the assessment is deficient: the test is too narrow and fails to include important dimensions or facets of focal construct” (p. 244). He defines construct irrelevance variance, as “the assessment is too broad, containing excess reliable variance that is irrelevant to the interpreted construct” (p. 244).

Watanabe (2004) suggests, a “distinction between positive and negative [washback] could usefully be made only by referring to the audience” (p. 21), and it is important to know “who the evaluation is for” (p. 21). Similarly, Cheng, Sun and Ma (2015), in their review of washback studies, suggest “in a sense, positive or negative washback is likely defined by test stakeholders, possibly differentially, as they see how a test serves its purpose from their points of view” (p. 437).

Specificity, intensity, length, intentionality. In addition to positive and negative directions of test washback, many studies have also looked into other dimensions of washback, such as specificity, intensity, length, and intentionality to represent its various characteristics (Bachman & Palmer, 1996, 2010; Cheng, 1997, 2005; Watanabe, 1997, 2004). Watanabe (2004) describes specificity as general or specific effects of any test. Watanabe considers general washback as test-takers’ motivation to prepare for a test, and

specific washback as related to “only one specific test or one specific test type” (p. 20). For example, if writing a 5-paragraph essay were included in a test, most students and teachers would emphasize a 5-paragraph essay in their learning and teaching. Similarly, Cheng (1997) describes washback intensity as “the degree of washback effect in an area or a number of areas of teaching and learning affected by an examination” (p. 43). This degree can be strong or weak. Other researchers (Wall, 2012) have suggested washback effects as immediate or delayed, direct or indirect, or apparent or not visible, i.e., changes are not manifested in explicit behavior.

Washback and validity. Researchers, such as Andrews and Fullilove (1994), Fredriksen and Collins (1989), Hughes (1989), and Morrow (1991) have claimed that washback should be a criterion for developing and evaluating language tests and have introduced terms such as ‘washback validity’ or ‘systemic validity’. For example, Hughes (1989) states, “potential backwash [or washback] effects should join validity and reliability in the balance against practicality” (p. 146). Hughes defines validity as the degree of correspondence between test results and what is being measured. Andrews and Fullilove (1994) argue, “as far as possible the test should embody the characteristics of a ‘good’ test. In particular, [the test developers] kept in mind the consideration of validity (especially face and content), reliability, and washback” (p. 64). One of Morrow’s (1991) five criteria for the evaluation of any communicative language test is the idea that “tests should reflect and encourage good classroom practice” (p. 111). In Canada, too, test development teams, such as at the Ontario Institute for Studies in Education (OISE), have utilized the washback concept in creating communicative language teaching (see Canale & Swain, 1980). One of OISE’s four tenets for the design of communicative tests has

been “work for washback’ – the notion that communicative tests should be explicitly designed to bring about positive washback” (Green 1985, pp 218-223, as cited in Bailey, 1999). However, Messick (1996) is of the view that “one should not rely on washback, with all its complexity and uncontrolled variables, to establish test validity... Rather, one can instead turn to the test properties likely to produce washback - namely, authenticity and directness - and ask what they might mean in validity terms” (Messick, 1996, p. 242).

Regarding washback, Messick (1996) stresses it is “not simply good or bad teaching or learning practice that might occur with or without the test, but rather good or bad practice that is *evidentially linked* to the introduction and use of the test” (p. 254; emphasis added). Messick points out two important points in testing to produce positive washback: firstly, test tasks should be criterion samples. This means they should be “authentic and direct samples of the communicative behaviours of listening, speaking, reading and writing of the language being learnt... and transition from learning exercises to test exercises should be seamless” (p. 241). According to this idea, there should not be any difference between the activities for learning and activities (or tasks) for testing. If there is a problem with any of these two, Messick suggests this difficulty could cause potential conceptual and methodological challenges in studying washback because “it is problematic to claim evidence of test washback if logical or evidential link cannot be forged between the teaching or learning outcomes and the test properties thought to influence them” (p. 247). My study is an exploration of ‘logical or evidential links’ that Messick believes must be in place between testing and washback on teaching and learning. I look specifically at what occurs when changes are made in a testing regime in

an EAP program and examine how much washback has occurred because of the new test content, structure and administration.

Washback and Innovation

One reason for a notable increase in the number of studies devoted to washback and impact of high-stakes standardized tests has been a major paradigm shift in assessment practices around the world, which, according to Cheng and Curtis (2004), is a reaction to “the direct instruction model under the influence of a behaviorist “tell-show-do- approach ...resulting in rote memory and isolated skills” (p. 16). The shift has resulted in an increased interest in washback of test innovations among researchers. Various studies regarding the washback of test innovations have been investigated in: Australia (Burrows, 2004), Canada (Cheng & Fox, 2007; Turner, 2009), China (Qi, 2005, 2007), England (Green, 2007), Hong Kong (Andrews, 1994, Cheng, 1997, 2005), Israel (Shohamy et al, 1996), Japan (Gorsuch, 2000; Watanabe, 1996, 2004), New Zealand (Hayes & Read, 2004), and Sri Lanka (Alderson & Wall, 1993; Wall and Alderson, 1993; Wall, 1999, 2005). However, as discussed previously, washback is a multi-faceted phenomenon. Cheng and Watanabe (2004) stress that “simply changing test contents or methods will not necessarily bring about direct or desirable changes in education as intended through a testing change” (p. xiii). They further suggest that specific educational contexts, testing cultures, where tests are being used, should be taken into consideration when examining washback. The following section will review research studies that examine the effects of change in tests on teachers and students.

Influences of testing on teaching. Stakeholders’ judgments about a test are an important tool for determining its washback (Chapelle, 1999; Cheng, 1997, 2005;

Shohamy, 2000; Wall, 2005). According to Winke (2011), “teachers and school administrators have a unique insight into the collateral effects of tests” (p. 633). Between these two stakeholders, teachers are the more noticeable and observable participants. Bailey (1999) calls them “the front-line conduits for the washback processes related to instruction” (p. 17). Alderson and Wall (1993) have also highlighted the importance of teachers in six of their fifteen washback hypotheses. The vast majority of washback empirical research has included teachers as the most prominent participants since teachers know their students well and have a unique vantage point from which to recognize differences between classroom practices and the effects of testing on teaching and learning.

As mentioned earlier, policy makers are aware of the importance of tests as a major tool in bringing about change in classroom teaching and learning. Alderson and Wall’s (1993) was the first empirical study in Sri Lanka’s educational context, which examined the top-down efforts to bring changes in teaching and learning by introducing changes in national English examination. The changes also included innovation in textbook materials and teacher training courses. Their study incorporated classroom observations to explore the potential positive, negative or neutral washback in teaching and learning. Their classroom observations yielded neutral washback in teaching methodology and they suggested this could be due to the lack of teachers’ understanding of the appropriate ways to prepare their students for the new examination. Alderson and Wall also reported negative washback where teachers skipped teaching listening skills because these were not being tested.

Wall (2005) reexamined Alderson and Wall's (1993) study about changes in Sri Lanka's educational context by using the analytical lens of Henrichsen's (1989) Hybrid Model of Diffusion/Implementation. Wall analyzed documents and teachers' interviews, and came up with a list of factors, which facilitated or hindered the implementation of the intended outcomes such as the examination itself and the new textbooks.

Other researchers have described the power of tests to promote positive washback. For example, Swain (1985) described 'Work for Washback' about a test that she and her colleagues developed for French immersion situations in Canada. They considered that positive washback could be achieved by involving teachers in all stages of the testing process and by developing a strong support system for teachers in terms of making and administering the tests and by providing "alternative teaching-learning strategies" (p. 44).

In terms of innovation and washback, multiple studies have tried to investigate the deliberate attempts to engineer change in teaching. For example, Lam (1993) studied how teachers' pedagogical practices changed in relation to preparing for the revised Hong Kong national examination. He surveyed 33 experienced (teaching under both, the old and new syllabus) teachers and 28 novice teachers who had taught only under the new syllabus, and found that the experienced teachers were "much more examination-oriented than their younger counterparts" (p. 91). Lam's conclusion was that it is insufficient to change exams; rather "the 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). Similar to Lam (1993), Andrews (1995) also administered questionnaires to test developers, experienced and novice teachers involved in teaching

oral English in the Hong Kong exam context. Andrews found that, the test developers highlighted teachers' readiness to devote time to improve students' speaking skills, but teachers felt the actual impact of the new syllabus was not as strong as the intended impact anticipated by the test developers.

Chapman and Snyder (2000) studied the educational quality in different countries where high-stakes national tests were introduced with the intention of improvement. Their conclusion was that changing high stakes national testing can improve instructional practices only in certain settings, but these effects are rather indirect. The intermediate conditions, such as teachers' understandings of the relationship between their pedagogical practices and the changes in the test should be well communicated by the implementers of change. For the desired impact on teachers' classroom practices, they suggest that the level of resources and their distribution among different schools are also important reasons for successful implementation of change.

Cheng (1997, 2005) studied the impact of the Hong Kong Certificate of Education Examination in English (HKCEE) from the time of the first official announcement up until the implementation of the exams over a few years. She used different methods of data collection, such as questionnaire, interviews and classroom observations. She suggests that the changes teachers made in their teaching, as a result of the new exams being implemented, were mainly about the "form" than of "substance" (Cheng 1997, p. 52) and teachers were more influenced by the publishers' understanding of the new exam, rather than by their own understanding of the exam. Cheng suggests that bringing changes in teachers' beliefs and behaviors is unlikely to happen too quickly because it is not easy to bring about changes just by changing the exam format. There is a lot going on

in any classroom, and “a change in examination syllabus itself will *not*, on its own, fulfill the intended goal. Teacher education and professional development must be involved in the process” (Cheng, 2005, p. 246, emphasis added).

Qi (2005, 2007) is of the view that in order for tests to be successful change agents, first it is necessary to understand the intentions of the test developers. These intentions, then, should be compared with teachers’ practices, so that the gaps between the intended and actual change can be identified. In Qi’s (2005) study of the writing tasks in the National Matriculation English Tests (NMET) in China, she found that the writing tasks used by teachers in preparation for the final communicative exam tended to be similar in word limit and format. Teachers mainly concentrated on content, organization, and accuracy, instead of stressing the importance of communicative features, such as appropriateness of language. Qi’s suggestion was that the NMET test failed to bring about intended changes in the classroom.

Similarly, Turner (2009) examined the effects of the early stages of an educational reform on English-as-a-second-language (ESL) teachers. Her notion of washback is similar to Messick’s (1996) explanation that the effects are only washback evident if they can be linked to the introduction and use of the test. She studied how teachers mediate between classroom assessment activities and preparing students for external tests. Contrary to the general notion in the washback literature about the negative effects of a change, she noticed a positive washback context where teachers integrated the exam characteristics (in this case speaking activities) in their everyday teaching. Turner says “it appears they [teachers] did this [integration] intentionally, and within the context of wanting to align teaching with assessment i.e., positive washback” (p. 120).

Wall and Horak (2006, 2008, 2011) conducted a longitudinal study of the impact of the newly introduced TOEFL exam on teachers' awareness of the change and their reactions to the change, and how they incorporated the new exam in their test preparation classes. Adopting Henrichsen's (1989) framework, Wall and Horak, first, collected the baseline data where they found that teachers did not have much awareness of the impending change at the beginning, and that teachers' awareness about change grew during the process of data collection. Teachers perceived, mainly, the task differences in the old and new tests and suggested that these tasks generally had a positive impact on their teaching. Wall and Horak, however, suggested that teachers' understanding of change was not complete, and this could be because of the lack of information provided about the new exam. They, therefore, suggested that communication channels disseminating information about change are extremely important.

The studies, just explored, have revealed that using tests, as engines of educational change are not simplistic tools and do not always work according to policy makers' or test developers' intentions. Teacher factors in washback may vary and depend on features such as teachers' understandings of the assessment innovation and its underlying principles, their professional knowledge, and their own beliefs (Alderson & Hamp-Lyons, 1996; Alderson & Wall, 1993; Cheng, 1997, 2005; Qi, 2005, 2007; Watanabe, 2004). For the possible beneficial effects, Hughes (2003) suggests that teachers must have an understanding of the test demands including teaching methods, syllabus design and material writing expertise.

Influences of testing on learning. Compared to washback research on teachers and teaching, research about the effect of washback on learners and learning is still

under-researched (Bailey, 1999; Cheng, 2014; Fox, 2004; Tsagari & Cheng, 2017). In any testing situation, it is student learning that is most strongly impacted. Other stakeholders' efforts, instead, are to promote language learning (Bailey, 1999; Fox, 2004). Fullilove (1992), in his study of Hong Kong examination, noted "students often feel that they are very small components of an enormous examination system which is highly impersonal on the one hand but personally highly important on the other" (p. 138). Bachman and Palmer (1996) suggest that test-takers are affected by a) "the experience of taking and, in some cases, of preparing for the test; b) the feedback they receive about their performance on the test; and c) the decisions that may be made about them on the basis of the test" (p. 31).

In relation to examination change, Cheng's (1998) investigation was probably the first attempt to examine students' perceptions and practices. She concluded that the HKCEE exam change did not have much impact on students' motivations and learning strategies, but had some effect on students' learning activities. Other studies have also underlined a variety of reasons as to why and how students make changes to their learning and studying practices including test preparation (Cheng, 1998, 1999, 2007; Green, 2007; Lummlley & Stoneman, 2000; Shi, 2007; Tsagari, 2009; Xie, 2011). These studies have revealed that learners determine for themselves how best to prepare for tests, and for the learners, washback does not flow in a straightforward manner either from the test or from the teacher (Green, 2007). For example, Ferman (2004) investigated the washback of an oral examination on learning among high-ability and low-ability students in Israel. Ferman concluded that cramming was more common among low-ability students and these students tended to study more intensively for the test than high-ability

students. Also, low-ability students tended to have more private help than high-ability students.

Similarly, the conclusion of Scott's (2007) study of a primary school context in the UK was that the degree of test impact on learners varied according to different grades and students in the higher grades felt more testing effects than students from the lower grades. Other studies have also explored the relationship among testing, learners, learner attitudes, and the learning process. For example, Lummley and Stoneman's (2000) study of students preparing for the local university graduation test in Hong Kong revealed that students relied on their past learning strategies and test-taking experiences that had major influences on their exam preparation methods, such as going through past tests and relying on test preparation books. Similarly, Damankesh and Babaii's (2015) study of Iranian high school students preparing for their high-stakes final exam revealed primary test preparation strategies included studying only what was tested, memorizing and reviewing past papers. Finally, Green (2007), in his study of IELTS preparation, concluded that students' understandings of the test demands are very important and these understandings are also related to their test preparation practices. His study also suggested that in high-stakes public tests contexts, if enough test-related information is unavailable to students and they do not fully understand the test construct, then it is difficult to predict that the exam will exert a positive washback.

Questioning the relationship between the exam preparation, test scores and actual learning, Cheng (2018) in her recent article on 'Geopolitics of assessment' concluded:

...teaching test-taking skills and drilling on multiple-choice worksheets is likely to boost the scores but unlikely to promote general understanding

(Montgomery & Lilly, 2012). As a result, students are forced to cram for examinations, rather than prepare for a broad curriculum. There are a number of studies that raise questions about whether improvements in test score performance actually signal learning. Others point to large-scale standardized tests' narrowness of content, their lack of match with curricula and instruction, their neglect of higher-order thinking skills, and the limited relevance and meaningfulness of their multiple-choice formats. According to these and other researchers, rather than exerting a positive influence on student learning, testing may trivialize the learning and instructional process, distort curricula, and usurp valuable instructional time (p. 5).

Reviewing the just-explored literature about washback on students, it can be predicted that tests that are used as mediators to promote desirable change in learning are not necessarily always efficient, and they may not have the desired consequences as predicted. More understanding of learners' beliefs, attitudes and expectations in particular testing situations are necessary to better understand the washback on learners (Cheng, 2014).

The literature review presented above reveals many factors from the point of view of teaching and learning and also describes the complex relationship between test washback and educational change. These factors are consistent with Messick's (1996) view that it cannot be assumed that a good test will automatically produce positive washback and a bad test will produce negative washback. To understand washback, it is also necessary to understand the educational context in which the test resides (Cheng,

2005; Messick, 1996; Wall, 2005), but what is unclear, so far, is how these contextual factors occur or interact with each other (Wall, 2012). Insights gained from the area of diffusion of innovation in educational contexts may help to understand the interactions of different contextual factors (Alderson, 2004; Alderson & Wall, 1993; Markee, 1997).

The aim of the following section is to discuss the change process and the role of contextual factors in it, while also identifying a framework that can be helpful to judge whether the newly developed innovations (the new tests) are likely to bring about intended positive washback in the EAP program being studied.

Insights from Innovation Studies

I begin this section with the definition of the term ‘innovation’ and then turn to Fullan (2007, 2015) to describe different phases of a change process and the meaning the different stakeholders make of the process of diffusion of innovation. This review is followed by a discussion of House’s (1981) classic and critical treatment of educational innovation through the analytical lenses of three perspectives. After that discussion, I summarize the composite question about innovation “who adopts what where, when, why and how?” (Cooper 1989, cited by Markee, 1993, 1997) and finally I end the discussion with the explanation of Henrichsen’s (1989) Hybrid Model of diffusion of innovation, which serves as the framework for my data gathering and analysis.

Definition of innovation. Rogers (2003) defines the term ‘innovation’ as “an idea, practice, or object that is perceived as new by an individual or other unit of adoption (p. 11). Thus, it is not the newness of the ‘idea,’ per se, but it is the perception of those who may be using it for the first time. Henrichsen(1989), citing Miles (1964, p. 14),

defines innovation as “a species of the genus ‘change’.... [It is] a deliberate, novel, specific change, which is thought to be more efficacious in accomplishing the goals of a system”...In addition, an innovation is usually “willed and planned for, rather than ... occurring haphazardly” (p. 65). Some researchers, such as White (1993), are explicit in distinguishing differences between the terms ‘change’ and ‘innovation’. White says “the difference has to do with intentionality: while *change* is any difference that occur[s] between time one and time two, an *innovation* requires human intervention” (p. 244). Other researchers use the terms synonymously. Although I understand that the distinction in using different terms, such as ‘change’ and ‘innovation,’ is valid, I use these terms interchangeably in my study.

There are many theories and models of diffusion-of-innovation available from different disciplines, such as education, sociology, psychology, business administration, and public health. (Fullan, 2007, 2015; Henrichsen, 1989; House, 1979, 1981; Kennedy, 1988; Lamie, 2004; Markee, 1997; Rogers, 2003; Waters, 2009; White, 1993). These theories and models have explained innovation from different angles and perspectives; however, Henrichsen (1989) suggests, “wherever new ideas can benefit practitioners, a diffusion of innovation approach can be valuable” (p.4). For example, Fullan’s (2007, 2015) view that educational change processes are slow and complicated is valuable when studying long term educational reform processes or House’s (1979, 1981) three analytical lenses of technological, political and cultural perspectives are useful in analyzing any educational innovation process. Similarly, Rogers’s (2003) distinction between innovations itself and ‘its diffusion’ is valuable, or are Kennedy’s (1988) subsystems of sociocultural context in which the innovation operates.

In language education, Markee (1993, 1997) made the most use of innovation theory and advised language teaching professionals to make use of a “diffusion of innovation perspective” in order to understand why certain attempts to innovate meet with success or failure. He suggested investigating not only the innovation (new test), and the context in which it operates, but also the role of different stakeholders and the process of implementation. The washback literature has also identified these factors as the most influential in producing intended washback. Finally, according to Henrichsen (1989), the experiences of previous educational reforms have demonstrated that “although change in a desired direction is possible it seldom happens by itself. Innovation is seldom sufficient on its own. Neither is merely communicating the news of an innovation to the appropriate audience enough to bring about change” (p. 4). Henrichsen’s (1989) hybrid model, most pertinent to my study, suggests investigating an innovation at three stages of implementing— antecedents, process, and consequences.

The process of innovation

Fullan’s view (2007, 2015). One of the most comprehensive books about innovation and change management in education is Fullan’s *The New Meaning of Change* (2007, 2015). For Fullan (2007), “*change is a process, not an event*” (p. 58, emphasis added). He suggests every change process needs on-going support to help individuals to cope up with change. Further, the reform process can be investigated in two ways: an ‘innovation-focused’ approach or a ‘capacity-building’ approach. While the innovation-focus approach is to “examine and trace specific innovations to see how they fare” (p. 65), a capacity- building approach is about how “to develop the innovative capacity of

organizations and systems to engage in continuous improvement” (p.65). Fullan suggests these approaches are not mutually exclusive, but build on each other.

According to Fullan (2007, 2015), the overall innovation management can be conceptualized in three broad phases:

- a) initiation – when a decision is taken to proceed with any innovation,
- b) implementation – when the attempts are made to put the innovation in practice
- c) institutionalisation – when the attention is directed towards innovation

sustainability. For Fullan, change is not a linear process but an iterative cycle of these three phases where outcomes are at the core. Ultimately, the end product of any change process enhances student learning and increases subsequent capacity to deal with future changes.

The *initiation* process, suggests Fullan (2007), is influenced by eight factors: “existence and quality of innovations, access to innovation; advocacy from central administration, teacher advocacy, external change agents, community pressure/support, new policy – funds and problem-solving and bureaucratic orientations” (p. 60).

The *implementation* stage can be affected by: a) characteristics of change – need, clarity, complexity and quality/practicality of innovation; b) local characteristics, such as district, community, teachers and students; and c) external factors such as government and other agencies. The *outcomes*, in Fullan’s framework, depending on the objectives of the change process, could refer to several different types of results. These could be “improved student learning and attitudes, new skills, attitudes, or satisfaction on the part of teaches and other school personnel, or improved problem-solving capacity of the

school as an organization” (p. 56). Fullan (2015) suggests a number of variables that can be operationalized in his model:

1. Numerous factors operate at each phase of the change process.
2. Change is not a linear process rather one in which “events at one phase can feedback to alter decisions made at previous stages, which then proceed to work their way through in a continuous interactive way” (Fullan, 2015, p. 57).
3. The scope of change and the question of who develops and initiates the change. The scope could be from “large scale externally developed innovations to locally produced ones.” (p. 57). Also, teachers’ role in the decisions and development of change process is important.
4. Time is a complication in the change model. There are no set time boundaries between different phases of initiation, implementation and institutionalization. For example, Fullan suggests it may take more than 2 to 3 years for a moderately complex change to become implemented. Although some outcomes can be assessed in a short time, the complete institutionalization (in case of large scale efforts) can take 5 to 10 years.
5. Finally, the most important variable in the change process is the ‘meaning of change’ that stakeholders make in the change process. Fullan states, “In the process of examining the individual and collective settings, it is necessary to contend with, both, the ‘what’ of change and the ‘how’ of change. Meaning must be accomplished in relation to both these aspects” (Fullan, 2015, p. 7). ... “we are not only dealing with a moving and changing target; we are also playing this out in social settings. Solutions must come through the development of *shared*

meaning. The interface between individual and collective meaning and action in everyday situations is where change stands or falls” (Fullan, 2015, p. 11, emphasis added).

It is complicated to fully comprehend the change process as the number or dynamics of interacting factors is too overwhelming and vast. That’s probably one reason that in the field of English language teaching, most research focus has been either one or more than one, rather than all of the above phases (Wall, 2005).

House’s (1981) three perspectives on innovation. House (1979) defines educational innovation as “the deliberate systematic attempt to change the schools through introducing new ideas and techniques” (p.1). He distinguishes innovations as relatively isolated, technical or programmatic alterations or a low-level change as compared to the larger structural educational reforms. To analyze any educational innovation process, House (1981) has suggested an interpretive framework of three analytical perspectives: *technological*, *political*, and *cultural*. He considers these perspectives as “screens” of “facts, values, and presuppositions” (p.1) to view innovation. Each perspective describes different issues and problems in the phenomenon of innovation.

The *technological perspective* focuses on the innovation, its characteristics, components, its production and introduction into the educational system. Innovation is considered a mechanistic process, which is based on rational analysis and empirical research and it is in the common interest of all stakeholders. Ethics of innovators are authoritative and the goals are predetermined, and there is one best way to accomplish those goals. The main principles and assumptions under this perspective are that the role

of consumer is passive and cooperation among different stakeholders is automatic. The image of innovation is that of production-orientation.

The *political perspective*, on the other hand, focuses on the innovation in a particular context, provider relationships, and innovation recipients in this context, as well as the rewards and costs of the dissemination. Unlike technological perspective, the ethics are contractual in political perspective, and the innovations are not necessarily in the best interest of all stakeholder groups. From this perspective, power and authority are the focal points and innovation is considered as a set of conflicts and resolutions among different groups where power struggle dominates. Under such circumstances, cooperation sometimes becomes problematic and there are conflicts over interests. The image of innovation is that of a conflict-orientation.

The *cultural perspective* is focused on the context, structure of work, as well as how the innovation is understood and interpreted in this context. Participants (as a group) are considered as cultures and subcultures and the innovation is seen as interactions of these. Effects of innovation are not easy to interpret and cooperation among different groups can be problematic as shared values among different groups can be different. Further, change may have different “meanings” for different groups. Therefore, innovations may have unanticipated consequences. Autonomy becomes an issue and different stakeholders may be in conflicts over values and interests. From a cultural perspective, a focus is on the meaning and values of the participants. House suggests that in order to analyze and explain the process of innovation, it is difficult to utilize only one of the perspectives because a single perspective is inadequate for a comprehensive explanation of any innovation process. Therefore, in my study, to analyze the

implementation process of the new testing regime, I will be combining these three perspectives for a more descriptive explanation of the change process at the EAP program.

The meaning of change. Another theme that has emerged in the innovation literature is the *objective reality* of educational innovation and *subjective meaning* for the individuals who make use of innovation. Fullan (2015) is of the view that the objective reality of an educational innovation process can be related to three different components: “the possible use of new or revised *materials*, the possible use of new *teaching approaches*, and the possible alteration of *beliefs*” (p.28). In educational contexts, changes in these three areas are necessary for the change to take place, but these areas do not necessarily require the same amount of amendment. Further, some adjustments are easier to affect than others. For example, it is easier to modify teaching materials than *subjective* teachers’ beliefs. Similarly, it is not easy to bring about changes to teaching approaches because these require learning a whole new set of skills. Kennedy (1988) believes behavioural changes are only “a surface phenomenon”... and deeper and complex changes come from the way people think about certain issues” (p. 329).

Markee’s (1997) framework. Markee’s (1997) framework is useful for language educators who are presented with the problem of how innovation may be designed, implemented, and maintained. Markee defines curricular innovation as “a managed process of development whose principal products are teaching (and/or testing) materials, methodological skills, and pedagogical values that are perceived as new by potential adopters” (1997, p. 46).

Like Fullan, Markee is also of the view that three major products of innovation are teaching (and/or testing) materials, methodological skills, and pedagogical values. These three are intertwined and affect each other. As mentioned in the meaning of change, materials of teaching or testing are the tangible products of innovation and that could be one of the reasons that most curricular innovators in education initiate the innovation from changing the existing materials (Markee, 1997). However, without considering teachers and engaging them in reflecting on their pedagogical values and developing more methodological skills, the process of innovation cannot be considered complete. That's why washback and innovation studies of new or revised tests should consider teaching materials, methodological skills and pedagogical values to generate a more comprehensive picture of washback. These are also relevant for my study because I am looking at how stakeholders differ in implementing/adopting the assessment change in the EAP program and what factors hinder or encourage the process of adoption of the new testing regime. Markee attempts to explain these principles by answering Cooper's (1983,1989) questions about innovation: *Who adopts, what, where, when, why and how*. Markee's explanation is summarized in Table 1.

Table 1 *Tabular summary of Markee's (1997) framework of diffusion of innovation*

Who	Adopts	What	Where	When	Why	How
<p>Social roles played by different participants as adapters/resisters, Implementers, etc.</p> <p>These roles are not mutually exclusive; some participants can play many complex roles at different stages of implementation.</p> <p>If all participants are not involved in the change process, power relations can cause misunderstandings.</p>	<p>Decision-making processes potential adopters go through;</p> <p>four types:</p> <ol style="list-style-type: none"> 1. gaining knowledge, 2. being persuaded of its value, 3. making preliminary decisions to adopt 4. confirming their decision to continue using the innovation. <p>These stages are not linear, but interwoven.</p>	<p>The curricular innovation. It is of two types:</p> <p><i>primary innovations</i>: teaching/testing materials, methodological skills, and pedagogical values;</p> <p><i>secondary innovations</i> rely on organizational development to enable primary innovations.</p> <p>The developmental change is related to the 'subjective' realities (of participants) and 'objective' realities (of the innovation)</p>	<p>Sociocultural context of innovation. Different factors affect implementation e.g., cultural, historical, political, administrative, and economic.</p> <p>Ranking of these factors is influenced by researchers looking at diffusion of innovation from different perspectives</p>	<p>Diffusion is seen as an interaction between time and the adoption of innovation.</p> <p>The speed of adoption varies for different users. Cooper's (1982) diffusion curve suggests change takes time to put in place and the implementers should be patient and allow more time for innovation to take hold.</p>	<p>Refers to psychological profiles of different adopters</p> <p>Relationship between an innovation and variation between adoption behavior of different participants</p> <p>Attributes of the innovation itself that lead to success or failure of the diffusion</p>	<p>Different approaches to effective change. Five models are proposed:</p> <ol style="list-style-type: none"> 1. The social interaction model, 2. Center-periphery model, 3. Research development, and diffusion model, 4. Problem-solving model, 5. Linkage model e.g., Henrichsen's (1989) model of diffusion

Based on his experience in different countries and borrowing from the field of general education, sociology, and development planning, Markee (1997) has put forward a set of general principles of curricular innovation as:

...curriculum innovation is a complex phenomenon; good communication is a key to successful curricular innovation; the successful implementation of educational innovations is based on a strategic approach to managing change; innovation is an inherently messy, unpredictable business; it always takes longer to effect change than originally anticipated; there is a high likelihood that change agents' proposals will be misunderstood. (pp. 171-180).

Models of Innovation

I have summarized a number of ideas from innovation theory, which are about the process of change; the meaning of change; the context and different participants in any change process; different perspectives in analyzing the change process; and strategies and approaches which affect change. Wall (2005), citing Snow (1973), draws out the differences between a theory and model as: “a theory is a symbolic construction designed to bring generalizable facts or laws into systematic connections and a model is a well-developed descriptive analogy used to help visualize, often in a simplified or miniature way, phenomena that cannot be easily or directly observed” (pp. 101-102). Models exist in different forms and these forms could be either “mathematical rigor of the closely articulated symbolical or suggestive metaphor and simile” (Snow as cited in Wall, 2005, pp.81- 82). The purpose of the next section is to present a model that is useful for the present study to investigate the phenomenon of washback in the context of testing regime change in the EAP program.

Markee (1997) in the ‘how’ section of the change process refers to five different approaches to effecting change. These five approaches highlight the role of the change agents and leadership style in bringing about desired changes. In the first *Social Interaction Model*, Markee (1997) suggests communication between different colleagues in the diffusion process is important. This model does not relate to any strategies of change and leadership style. By establishing communication networks, innovations can be spread easily. In the second, *Center-periphery Model*, which is more top-down; the decisions to make changes rest with a small number of people in authority. Teachers are on the periphery of the change process, and are merely the implementers of the decisions made by “power-coercive change strategies and mechanistic leadership styles” (Markee, 1997, p. 63). Similarly, the third type, *Research Development, and Diffusion (RD&D) Model*, is also adopted by top-down centralized educational systems and is based on the assumption that to develop good innovations, academics’ research efforts are needed than people in authority taking the decisions. However, Markee (1997) warns that “teachers-who are at the bottom of this expert-driven decision-making hierarchy – still do not own the products of this approach” (p. 65). The fourth model suggested by Markee (1997) is the *Problem-Solving Model*. This model is different than the previous two models and is a “bottom-up” approach to the change process. The potential users of an innovation decide whether a change is needed and they come up with possible solutions, trials and evaluations of innovations. Markee points out that this approach “assumes that people’s actions and beliefs are governed by their social values” (p. 67) and the strategy of change is normative-reductive, which means that a change in behavior involves change in beliefs.

The leadership style is more about discussion and persuasion than power. Teachers in such approaches are initiators and decision makers in the change process.

The fifth model is the *Linkage Model*. As the name suggests this model is a synthesis of other models of change (Markee, 1997). In this model, “a change agent’s decision to use a particular change strategy is contingent on the problem to be solved” (Markee, 1997, p. 68). Further, for different sociocultural situations, different approaches are required depending on the types of problems that need to be solved. In this model, change is seen as the product of “collaborative interaction” (Henrichsen, 1989, p. 68). The model recognizes “the complex nature of change processes, but it has the disadvantage that change agents who use this model can be confronted with a very steep learning curve” (Markee, 1997, p. 68). Henrichsen’s (1989) Hybrid Model of Diffusion of Innovation is an example of a linkage model. I use this model to investigate the phenomenon of washback in the context of testing regime change in the EAP program.

Henrichsen’s (1981) Hybrid Model of Diffusion

Within English language education, Waters (2014) suggests, the best-developed research framework for “identifying factors within an innovation project that are likely to influence the potential for innovation institutionalization” is the ‘Hybrid model’ (p. 99). Wall (1999, 2005) and Wall and Horak (2006, 2008, 2011) recommend using Henrichsen’s (1989) Hybrid Model to study test washback because it can account for a multitude of contextual variables as the model is “comprehensive without being overly detailed” and it gives a “clear picture of the factors which are most important at different stages in the innovation process” (Wall, 1999, p. 184). For example, while analyzing a Sri Lankan educational context, Wall (1999) cited many contextual factors from

Henrichsen's model. These factors were helpful in explaining why teaching did not change after a new exam was introduced. Henrichsen has recommended using the hybrid model before an innovation or reform is undertaken, while it is in progress, and after it has been implemented.

Before describing Henrichsen's model in detail, I would like to bring out the differences among different types of English language educational institutes and my rationale for excluding a few contextual factors mentioned in Henrichsen's (1989) Hybrid Model of Diffusion in my study.

Waters (2014), citing Holliday (1994, 2005), distinguishes two English language institution 'archetypes', namely BANA (British, Australasian and North American) and TESEP (tertiary, secondary and primary). The two types of institutions differ in some of their basic characteristics. First of these is the class size and composition. While TESEP have large class size (usually 35+) and all students share the same L1, the class size in BANA is relatively small and students have varied L1. Waters (2014) elaborates further on their differences as:

The overall purpose of learning in the two kinds of institutions also differs, being largely INSTITUTIONAL (education system needs based) in the case of TESEP entities, and mainly INSTRUMENTAL (learner communication needs-based) in BANA settings..... their respective dominant PROFESSIONAL-ACADEMIC CULTURES – the paradigm of educational values and related practices to which teachers in each of the two types of institution typically adhere to. TESEP teachers are seen as generally favouring a COLLECTIONIST educational orientation, which

results in a '[d]idactic, content-based pedagogy' (Holliday 1994a: 72).

BANA teachers, on the other hand, are regarded as more likely to adopt an INTEGRATIONIST educational perspective, leading to the use of a more '[s]kills-based, discovery-oriented, collaborative pedagogy' (pp. 93-94).

These differences are pertinent for my research for two purposes: first, the rationale for excluding a few factors from Henrichsen's diffusion of innovation framework in my data analysis and discussion; and second, when describing teaching and learning in the EAP and comparing the washback to teachers and washback to students, as most teachers in the program had the BANA ideologies and the majority of the student body in the program were from TESEP cultures.

Henrichsen (1989) applied the Hybrid Model to investigate an intended English Language Teaching (ELT) innovation in Japan. He used different data sets, such as interviews, document analysis and reviews of the literature in three distinct elements: antecedents, process, and consequence. Henrichsen related the antecedent elements to the eventual outcomes of the reform process of ELT innovation. He observed that the teaching after the introduction of the ELT innovation was not because of the innovation itself, but rather because of various contextual factors in the Japanese educational settings (e.g., traditional ways of teaching) before the introduction of the innovation and other factors in the context (e.g., insufficient teacher support, resourcing and issues with communication regarding the test). The three key elements of Henrichsen's model are described next (for a detailed description of these elements, see Figure 1).

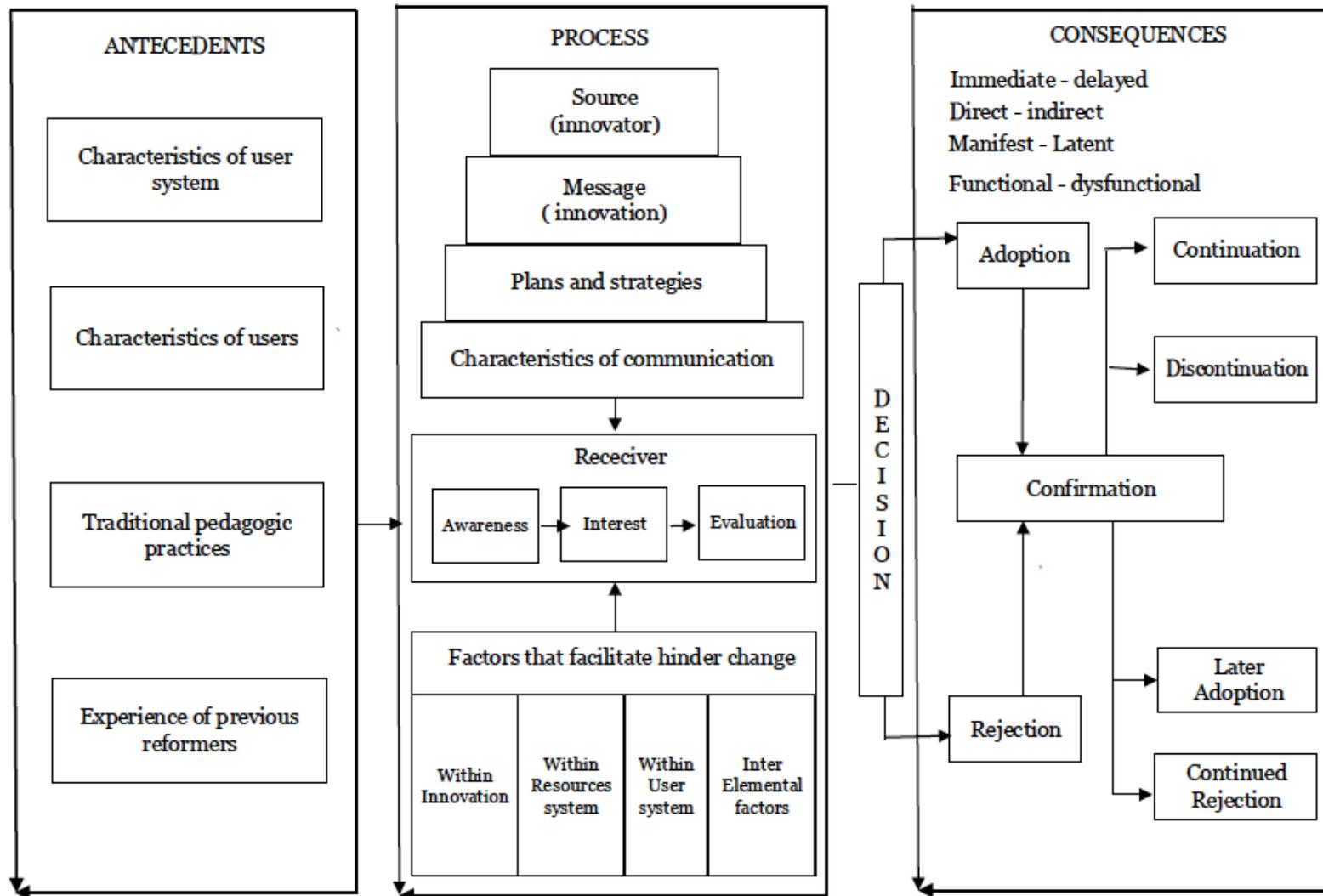
Antecedents. The *antecedent* elements in Henrichsen's model refer to the conditions in place in the educational context before an innovation is proposed. In the

innovation literature, this antecedents stage is also known as: Fullan's (2015) Initiation; Roger's (2003) 'Prior Conditions'; and Kennedy's (1988) 'Interrelating Systems'.

Henrichsen suggests before deciding on the type of innovation that would be suitable for their context, change agents should be aware of the following four factors:

1. Characteristics of the *Intended User System* – put simply as context, where the innovation is introduced, the intended user system is the structure and power relationships in institutes because these are the products of long historical development.
2. Characteristics of *Intended Users of the Innovation* – These are the attitudes, values, norms, and abilities of different users of innovation (Henrichsen, 1989). These characteristics can affect the diffusion/implementation effort. For example, if the users are dedicated to the status quo, there are few chances for the innovation to take hold.
3. *Traditional Pedagogical Practices* — culturally specific traditions of teaching and learning shape the form and content of much school learning (Henrichsen, 1989).
4. *Experiences of Previous Reformers* – To view how an innovation will be adopted in the same socio-cultural context, an understanding of previous reforms can provide valuable guidance.

Figure 1 The Hybrid Model of the Diffusion/Implementation Process (Henrichsen, 1989, p. 80²).



² This figure is included by permission under the *Fair Dealing Exception* of the Copyright Act.

The characteristics of both the intended user system and the intended users are not only the foundations for the new practices, but also are the most influential in the *process* phase of the model (Wall, 2005). I will be appropriating the just-explained, four factors (e.g. user system, users, pedagogical practices and experiences of previous reforms) as evaluative criteria for examining the antecedents of the former testing regime in the EAP program in Chapter 4.

Process. Miles (1964), as cited in Henrichsen, suggests, “innovations are almost never installed on their merits. Characteristics of the local system, of the innovating person or group, and of other relevant groups often outweigh the impact of what the innovation is” (p. 82). Henrichsen is of the view “in the development of an implementation strategy, a consideration of these ‘implementation factors’ and how they can work for or against the intended innovation is important. In fact, the implementation strategy should take precedence over other aspects of planning” (pp. 81-82). Literature on innovation has also stressed the process of implementation requires attention to myriad concerns, and these concerns should be identified in the planning or Initiation stage (Fullan, 2007, 2015; Stoller, 1994) because many facilitating or inhibiting factors can be present in the context.

Therefore, in addition to paying attention to the source, message, receivers of innovation, plans and strategies of change implementation, Henrichsen recommends investigating various factors that facilitate/inhibit the change process. These are: *within the innovation itself, within the resource system, within the user system, or inter-elemental factors.*

The discussion of these factors is important for this dissertation as these factors are appropriated as the evaluative criteria for *the Process* of implementation of the new

testing regime change in Chapter 5. The following section describes, in detail, the characteristics of these four factors (see Figure 1), which facilitate or inhibit the change process.

Factors within the innovation itself. Also called the attributes of innovation (Rogers, 2003), the characteristics of innovation as perceived by the intended adopters are important in themselves as well as in their interactions with other implementation factors (Henrichsen, 1989). Henrichsen listed eleven characteristics of an innovation, which could make it more or less acceptable to potential users: *originality, complexity, explicitness, relative advantage, trialability, status, practicality, flexibility, primacy and form.* Among these factors within the innovation itself, Henrichsen considers ‘originality’ and ‘complexity’ as the most important factors in any implementation process. He suggests *originality* of an innovation relates to the origin of the innovation and whether it is perceived as being appropriate for local circumstances. Further, the originality also depends on whether it is invented locally within the context or it is adapted or borrowed as is from some external source.

The *complexity* of an innovation is defined in many different ways in the innovation literature. For example, Dow, Whitehead and Wright (as cited in Henrichsen, 1989) suggest complexity comes first on the list of barriers to change and it is related to the amount of change and the number of people involved in the change process. Rogers and Shoemaker (as cited in Henrichsen, 1989) define complexity as “the degree to which an innovation is perceived as difficult to understand and use” (p.82). Pelz (as cited in Henrichsen, 1989) suggests that complexity can also be defined in terms of technical complexity, organizational complexity, sophistication or intellectual complexity and radicalness.

The previous discussion, of the eleven characteristics of innovation, is important for my study because I appropriated these characteristics as the evaluative criteria for the new integrated skills tests (ISTs) during the implementation of the new testing regime.

The results of the evaluation are presented and discussed in Chapter 5.

Factors within the Resource System. Henrichsen (1989) suggests, “the resource system that promotes an innovation also has characteristics that affect the course and success of implantation efforts” (p. 86). The resource system is further sub-divided into four characteristics: *capacity, structure, openness and harmony*. Wall (2005), citing Henrichsen (1989, p. 87-94), gives the following definitions:

Capacity refers to the system’s ability to ‘retrieve and marshal diverse resources’ and to convey, store and retrieve large amounts of information. It also refers to the system’s ability to influence potential adopters.

Structure refers to the division of labour and coordination of efforts within the system, the coherence of its view of the client system, and its ability to plan and get messages across in a structured way.

Openness refers to whether the system has ‘a willingness to help and a willingness to listen and be influenced by users’ needs and aspirations’.

Harmony refers to the social relationship between all individuals within the resource system. (p. 86).

These characteristics are important to understand the availability of resources and power relations in the context of study where innovation is implemented.

Factors within the Intended-User system. The characteristics of the context itself can be the powerful determinants of the success or failure of the diffusion process.

Henrichsen divides these factors into ten different characteristics: *geographical location, centralization of power and administration, size of the adopting unit, communication structure, group orientation and tolerance of deviancy, openness, teacher factors,*

educational philosophies and examinations. Most of these characteristics are fairly easy to understand from their labels.

Inter-elemental factors. Henrichsen (1989) suggests, “a number of factors exist “between” rather than “within” the elements involved in the diffusion and implementation of innovations” (p. 92). He suggests five critical inter-elemental factors as: *compatibility, linkage, reward, proximity and synergism.* Wall (2005), citing Henrichsen (1989, p. 87-94), gives the following definitions of the inter-elemental factors:

Compatibility refers to the degree of ‘fit’ between the innovation and the intended users, as well as to the fit between the resource system and the intended-user system.

Linkage refers to ‘the number, variety, and mutuality of contacts between the resource system and the user system’.

Reward refers to ‘the frequency, immediacy, amount, mutuality of, planning and structure of positive reinforcements’ – such as economic gain, status, satisfaction etc.

Proximity refers to ‘the nearness in time, place, and context’ of the resource system.

Synergism refers to ‘the number, variety, frequency, and persistence of forces that can be mobilized to produce a knowledge utilization effect (p. 86).

These factors are important when describing the interactions between the innovation, the resource system, and the intended-user system, i.e., the context of the innovation. As mentioned above, the discussion of the above mentioned facilitative and inhibitive factors is important for this dissertation because these are used as the evaluative criteria for *the process* of implementation of the new testing regime in the EAP program.

Consequences: The decisions and outcomes of a change process are as complicated as the process of implementation. Henrichsen’s Hybrid Model conveys this through the complexity of the consequence phase. This phase is similar to Fullan’s (2007,

2015) outcomes and Roger's (2003) decision and confirmation stage. However, sometimes the adopters can later change their decisions to adopt or reject changes in the consequence phase. There are many other outcomes possible in the change process. For example the outcomes can be *immediate or delayed*. If there are many outcomes, it is possible that one outcome is influenced by another outcome; then the results can be 'direct or indirect'. If the users recognize the outcomes, the results are 'manifest,' but if the change is neither recognized nor intended, then the results are 'latent' (Henrichsen, 1989, pp. 90-95).

Research Questions

The purpose of this longitudinal case study is to investigate the phenomenon of washback by adopting a diffusion of innovation (Henrichsen, 1989) perspective in the context of a systematic change of a testing regime in an EAP program. The new integrated skills tests (ISTs) were introduced as a key component of curriculum change, in keeping with the program's new learning outcomes and the role of high-stakes testing in the program. This study employed the models of Hughes' (1993) Principles of Participants, Process and Products of the working of washback, Henrichsen's (1989) Hybrid Model of Diffusion/Implementation Process, and Markee's (1997) Curricular Innovation Model to investigate whether or not a high-stakes test can be used to leverage positive washback as part of curricular innovation in the EAP program.

The principles of the phenomenon of washback as change (e.g. Cheng, 1997) and the theories of curricular innovation (e.g. Fullan, 2015) provide an assumption for my study that can be stated as: if a new test is used in an educational context to leverage positive washback, then this positive washback is a type of diffusion of innovation. I

argue that the process of diffusion/implementation affects not only the intended users of the innovation, but also the final outcomes/washback of such initiatives. Any test washback generally does not occur as it is intended, but is dependent on the specific context (characteristics of the user system), different stakeholders (characteristics of the users), and the test itself (innovation). If studied together, as I have attempted in my study, these factors may provide empirical evidence of test washback. As Messick (1996) has noted, “it is problematic to claim evidence of test washback if a logical or evidential link cannot be forged between the teaching or learning outcomes and the test properties thought to influence them” (p. 247). My study endeavors to find a) a logical or evidential link between teaching and learning outcomes in the EAP program and a new test’s properties that may have influenced them; and b) how other contextual factors in the EAP program influenced washback of the new ISTs.

Research Question 1: What evidence is there of washback in the former testing regime and what is the intended washback of the new testing regime?

This research question investigates the antecedents i.e. the situation on the ground before the change in the testing regime was proposed. The initiation phase of any innovation implementation process is similar to the antecedents (Fullan, 2015). Since previous research on washback of new or revised tests has highlighted the importance of studying conditions before “treatment” (Henrichsen, 1989; Wall, 2005; Weir & Roberts, 1994), the first question explores the washback from the ExitTest, used in the former testing regime, on teaching and learning in the EAP program from the vantage point of three different stakeholders – teachers, students and administrators. This question also enquires about the types of intended effects that administrators proposed for the new testing regime. The purpose of this question is to match the intended effects with the actual effects in later

phases of the study. Previous researchers (e.g. Bailey, 1996; Cheng, 2005) have identified a lack of multiple perspectives in previous washback studies. Therefore, data were collected from different stakeholders (e.g. teachers, students and administrators) in the EAP program.

Research Question 2: What evidence is there of washback factors facilitating and/or impeding the implementation of the new testing regime?

In this research question, the contextual factors that could facilitate and/or hinder the implementation of the new testing regime and the washback of the Integrated Skill Tests (ISTs) are examined. This period is the “during or implementation” phase of the ISTs when the attempts were made to put innovation in practice (Fullan, 2015). To understand the immediate washback of the ISTs on teaching and learning in the EAP program, data were collected from teachers, administrators and students in the first semester of the implementation of the new testing regime.

Research Question 3: What evidence is there of washback of the new testing regime over time?

In this research question, the delayed washback from the new testing regime is explored. Fullan (2015) suggests that in this “institutionalization phase” of a change process, attempts are made for innovation sustainability. Research on both washback of new or revised tests and the diffusion of innovations has suggested that it takes time for any innovation to take hold (Bailey, 1996; Markee, 1997; Wall, 2005). Therefore, the purpose of this research question was to study the actual washback of the new testing regime three semesters after its implementation.

Chapter Summary

This chapter has reviewed the literature that gave direction to this study. I, first, discussed the theoretical foundations and issues with measuring academic constructs in second language testing. I, then, reviewed the phenomenon of washback, explored its complexity and its mechanisms, as well as the reasons to use innovation theories in identifying and explaining the factors in educational contexts that may contribute to the potential washback of a new test.

Most research on the phenomenon of washback has suggested that it is far more complex than it initially seemed (Alderson & Wall, 1993; Cheng, 1997, 2014; Spratt, 2005; Wall, 2012). The studies that have been conducted about new or revised tests have emphasized that it is important to look at factors beyond a particular test when attempting to predict the form of washback in the new surroundings. These studies have also suggested that there are different participants involved at every stage of a process and these participants all have their own needs and constraints (Wall, 2005). Further, studying different stakeholders' (e.g., students and teachers) characteristics has been reported as useful for explaining the phenomenon of washback and the direction that test washback might take. As Green (2007) points out, "Differences among participants in the perception of test importance and difficulty, and in their ability to accommodate to test demands, will moderate the strength of any effect, and perhaps, the evaluation of its direction" (p. 25). Therefore, it was important to take accounts of teachers and students in consideration when investigating whether a high-stakes test can be used to leverage positive washback on teaching and learning.

Furthermore, successful educational innovation requires change on at least three levels: content, methodology, and attitudes (Fullan, 2015; Markee, 1997; Wall, 2005). For

example, as noted by Cheng (1997), it is easier for teachers to change the content or their behaviours in their teaching than their beliefs, attitudes, and values. Innovation theories are useful for identifying and explaining such factors in the educational context. Finally, the innovation process is long and complicated, which consists of many phases and stages. Innovators need to ask different questions during each phase to make sure that the innovation is going the way it was intended to unfold (Fullan, 2015).

What is significant in all the studies and work cited (e.g. Cheng, 1997; Henrichsen, 1989; Wall, 2005), about both washback and innovation, is the importance of context. Rea-Dickins and Scott (2007) have rightly suggested “washback can be viewed as a context-specific, shifting process, unstable, involving changing behaviors in ways, which are difficult to predict” (p. 5). The next chapter will describe the present study’s methodology and the overall research design of the study. It will also describe the participants, procedures used for data collection and data analysis.

Chapter 3: Methodology

Introduction

This chapter explains the longitudinal case study method employed in the present qualitative study. I start by positioning the study within a case study framework using a single-case design with four embedded units of analysis. Following this, the positionality and procedures I used to ensure the validity and reliability of the case are explained. I, then, describe the use of interviews and focus groups to elicit case data. After that, the research context and selection of the participants and other data collection instruments are explained. Following that, the study's research design is elaborated. The design includes three distinct phases adapted from Henrichsen's (1989) Hybrid Model of Diffusion. These three phases align with my three research questions. Dividing the model allows me to address the research questions separately before determining if a high-stakes test can be used to leverage positive washback on teaching and learning in the EAP program. I, then, include a discussion of the quantitative and qualitative data collection instruments, to end with a detailed description of the procedures that were followed and data analysis.

Research Method

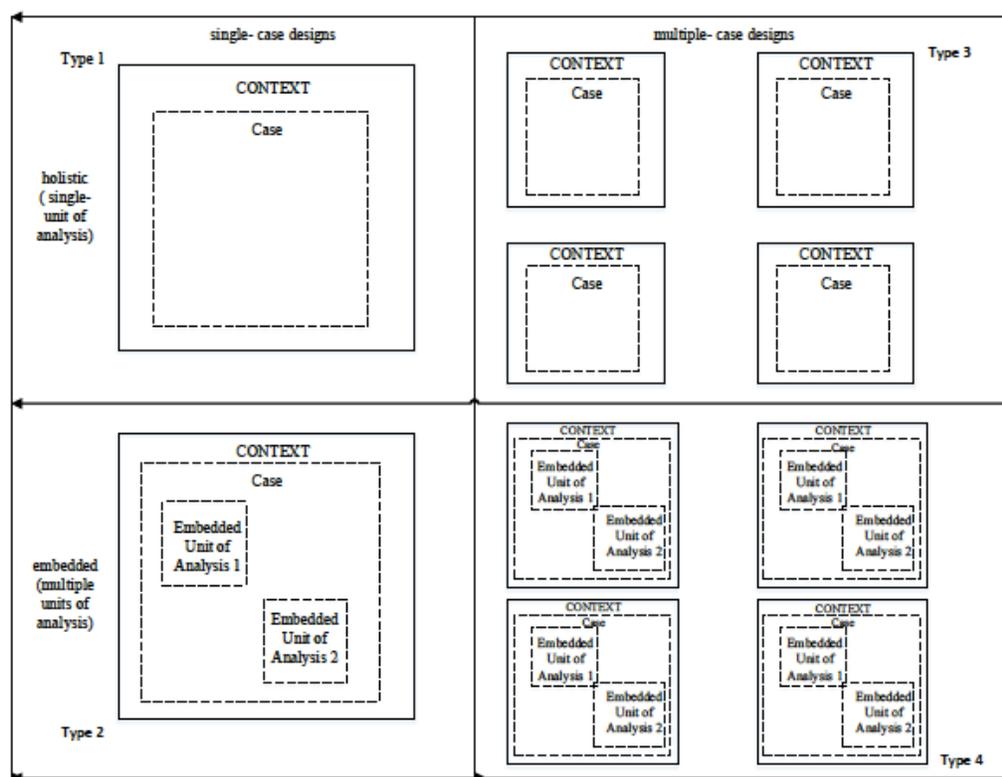
As discussed in the opening chapters of this dissertation, this study adopts a diffusion of innovation perspective. It presents a study that was undertaken to investigate the phenomenon of washback in an EAP program where a new testing regime (innovation) was introduced as the new exit criteria for the Graduating Level (GL) students. As demonstrated in the previous chapter, a number of washback studies have approached the washback phenomenon by investigating the influence of a single test, at

one point in time, and mostly by eliciting accounts from single stakeholder groups (e.g., teachers or students). The uniqueness of this study lies in examining two sets of testing regimes (before and after change) at different time intervals in a twenty-month data collection period. Most importantly, this research was undertaken from the vantage points of three groups of stakeholders: teachers, students, and administrators. Thus, a longitudinal case study approach with four embedded units of analysis (Yin, 2014) was an appropriate choice of methodology for this study.

Yin (2014) suggests two main types of case studies: single-case studies and multiple-case studies (Figure 2). Single-case studies are conducted in one context and can have different embedded units of analysis depending on the case. Each embedded unit of analysis is “a unit lesser than the main unit of analysis, from which case study data also are collected” (p. 238). When the case itself is the unit of analysis it is called a holistic single-case study, and if there are two or more embedded units of analysis, it is called a single case study with embedded units of analysis. Multiple-case studies, on the other hand, include two or more cases in different contexts.

The overall design of this qualitative study was a single case study with four embedded units of analysis (see, Yin’s Type 2, Figure 2). The case was the phenomenon of washback at one level of an EAP program at a Canadian university. The four embedded units of analysis were the accounts of: program administrators, teachers and students at the graduating level of the program and document analysis during the period of study.

Figure 2 Basic types of designs for case studies (Yin, 2014, p. 50³)



The three key stakeholder groups were purposefully selected to gain a multi-faceted perspective of the phenomenon of washback as innovation and its implementation process. The first unit of analysis was the administrators' accounts of the former and new testing regimes. These accounts allowed me to understand their rationale behind the change implementation. The other units of analysis were the teachers' and students' accounts of assessment practices before, during and after the change implementation. It can be argued that examining the phenomenon of washback from different stakeholders' perspectives could be used to arrive at a more comprehensive understanding of it. Both qualitative interviews and focus groups, and quantification provided by data from surveys were used to inform my interpretation of the stakeholders' accounts of washback in the

³ This figure is included by permission under the *Fair Dealing Exception* of the Copyright Act.

former and the new testing regimes. It is important to note that qualitative data collection techniques such as interviews are known to be useful in drawing out in-depth insights into attitudes and reported practices of participants. This enables a rich, multi-layered account of differing perceptions of washback (Burrows, 2004; Scott, 2007; Wall, 1996, 2005; Wall & Horak, 2008, 2011; Watanabe, 2004). Furthermore, according to Miles, Huberman, and Saldaña (2014):

Using the same instruments as in prior studies is the only way we can converse across studies. Otherwise, the work will be noncomparable, except in a very global way. We need common instruments to build theory, to improve explanations or predictions, and to make recommendations about practice (p. 39).

In sum, the quantitative data from surveys in the study was mostly descriptive, intended to gain triangulating information such as students' accounts of assessment practices in the EAP program.

Why a Case Study? In education, case studies are most often qualitative. In the field of applied linguistics, case studies “involve rich contextualization and a deep, inductive analysis of data from a small set of participants, sites, or events in order to understand aspects of language learning or use” (Duff, 2012, p. 1). Case study is preferred when *how*, *what* and/or *why* questions are asked and a number of data collection measures are incorporated, such as documentation, archival records, interviews, direct observation, participant observation, and physical artifacts (Yin, 2014). Yin suggests that explanatory “how” and “why” research questions are better answered with case studies because “such questions deal with operational links needed to be traced over time, rather than mere frequencies or incidents” (p. 10).

Rooted in a constructivist perspective, case studies consider truth as relative and dependent on one's perspective (Baxter & Jack, 2008; Stake, 1995; Yin, 2014).

According to Crabtree and Miller (1999), a constructivist perspective “recognizes the importance of the subjective human creation of meaning, but doesn't reject outright some notion of objectivity. Pluralism, not relativism, is stressed with a focus on the circular dynamic tension of subject and object” (p. 10). They suggest that the advantage of the constructivist approach in case studies is the collaboration between the researcher and participants. This allows participants to tell their stories by describing their views of reality and enables the researcher to better understand their actions. In this way, case studies can help gain an in-depth understanding of the situation and meaning for those involved.

Additionally, in case studies, the interest of the researcher is “in the process and not the outcomes, in context rather than a specific variable, and in discovery rather than confirmation” (Merriam, 2009, p. 19). Dornyei (2007) suggests that case studies help in obtaining a thick description of a complex social issue embedded within a cultural context. This helps a researcher examine and understand how “an intricate set of circumstances come together and interact in shaping the social world around us” (p. 155).

As discussed in the literature review, many washback studies have been quantitative in nature, relying on data from surveys and questionnaires (Lam, 1993; Qi, 2005). Bailey (1996) stresses the importance of studying washback in naturally occurring settings rather than in laboratory conditions. While some outcomes of assessment, for example test results, can be easily measured, other “outcomes and processes can only be analyzed and described” (Wall, 2012, p. 88). Washback effects, whether intended or unintended, are controlled by factors within the classroom, program, and/or other

sociocultural factors. Understanding the interaction of these factors is important for the promotion and implementation of diffusion of innovation (Markee, 1997). In this regard, case study is a highly appropriate strategy for the investigation of the phenomenon of washback, as it “offers a means of investigating complex social units consisting of multiple variables of potential importance in understanding the phenomenon” (Merriam, 1998, p. 32).

By including different stakeholders’ perspectives, a case study approach can reveal interactions and processes of the situation that may become invisible within large-scale studies or if relying on a singular vantage point (Yin, 2009). Moreover, in case studies, the researcher “can be seen as one of the participants” (Casanave, 2015, p. 124) and totally immersed in the case. This may allow the researcher to attend to every detail or piece of evidence during the process. In this way, rich, in-depth, and integral information can be gathered in a single case study. Because I was a researcher and a teacher in the program under study, I was able to get both an outsider (as a researcher) and an insider (as a teacher in the program) view of different sources of evidence.

One of the strengths of a case study is its fieldwork that can allow a researcher to be immersed in a particular setting and capture situations of an innovation or program. Specifically, in examining educational innovations and evaluating educational programs, case study methods have been found to be very useful (Merriam, 2009; Patton, 2005; Yin, 2014). Yin (2014) proposes that survey and experimental strategies are too simplistic to capture the complexity of program implementation and their effects. As case studies are situated in real-life settings, they can provide a rich and holistic account of a phenomenon under study as the end product (Merriam, 2009). Thus adopting a case study approach

helped in obtaining a thick description of how the introduction of ISTs influenced classroom teaching and learning in the EAP.

Binding the case. Miles and Huberman (1994) define a case as “a phenomenon of some sort occurring in a bounded context” (p. 25). Yin (2009) defines case study as an empirical inquiry that “investigates a contemporary phenomenon in depth and within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident” (p. 18). My study encompasses these qualities because the boundaries surrounding the assessment innovation and the EAP program are indistinguishable spatially and temporally. Additionally, the innovation (introduction of a new testing regime); the phenomenon of washback, and the key stakeholders (i.e., administrators, teachers and students) are all essential parts of the same EAP program. My study was bounded by place, time, activity, and perspective. The study examined the assessment innovation in five terms of study within one EAP program (January 2016 to August 2017).

To avoid common pitfalls associated with case studies, such as broad questions or a topic/study with too many objectives, researchers (e.g. Creswell, 2012) have suggested *binding* a case. Binding a case means that once a researcher has determined what the case will be, they must also consider what the case will NOT be. In the case study literature, the following are some of the suggestions to bind a case:

- By definition and context (Miles & Huberman, 1994);
- By time and place (Creswell, 2012);
- By time and activity (Stake, 1995).

These boundaries help in deciding what will and what will not be researched in the study. Baxter and Jack (2008) compare these boundaries to “the development [of] . . . inclusion criteria for sample selection in a quantitative study” (p. 547). However, for case studies these boundaries not only include the sample, but also the breadth and depth of the study.

The main activity studied in this research was the phenomenon of washback within one level of an EAP program. Although meaningful, what were NOT studied in my research were other curricular changes in the program, test development or validation of the former and new tests, and students’ final grades/outcomes during the period of this study. Of importance was focusing on and limiting perspectives to only those with the closest proximity to the phenomenon of washback: administrators, teachers, and students.

Researcher Positionality

As a researcher and a teacher in the EAP program, I wanted to investigate the washback of the new testing regime on teaching and learning in the program. I had worked in the program two years before the proposed changes were to be implemented. I continued teaching there during the data collection period. The EAP program presented an excellent research setting for my study. I knew the teachers, administrators and the students- especially those in the graduating level of the program. I was also part of the curriculum renewal committee and one of my roles was to provide feedback on developing new learning outcomes. My involvement with the new learning outcomes represented additional incentives for me to research the effects of the new testing regime.

On the one hand, my role as both a teacher and researcher allowed me to explore different aspects of the new testing regime from two different perspectives. This dual-

positioning helped me develop a better sense of whether the testing regime as innovation was effective in producing positive washback. On the other hand, my role as an insider may have limited to a certain degree my openness or perceptions of the change process. I have attempted to mitigate this with rich and thick description over time. Although I taught at the graduating level during all phases of this study, I did not include any of my students in the data collection process. (See Appendices N and O, for copies of the ethics certificate and consent forms)

Criteria for Judging Reliability and Validity of Case Studies

One common criticism of qualitative research is its lack of representativeness, reliability, and validity. Yin (2014) states critics often ask “How can you generalize from a single case ... case studies, like scientific experiments, are generalizable to theoretical propositions and not to population or universe” (pp. 20-21). Therefore, the goal of case studies is “to expand and generalize theories (analytical generalizations) and not to extrapolate probabilities (statistical generalizations)” (p. 21). It is another type of generalizability, not a lack of it and the preference is for the notion of transferability or resonance.

Assuring reliability and validity was paramount in the present study. Yin (2014) suggests several tactics to employ throughout the course of a case study to maintain reliability and validity. Those that are relevant to my study are outlined in Table 2.

Table 2 *Case study tactics for testing of case studies (adapted from Yin, 2014)*

TESTS	CASE STUDY TACTICS	PHASE OF MY STUDY
Construct Validity	Use multiple source of evidence Establish chain of evidence	Data collected in phases 1, 2 and 3 of the study
	Have key informants review draft case study report	Data analysis and writing stages of the dissertation
External Validity	Use theory in single-case studies	Use of the proposed models (e.g Henrichsen, 1989; Hughes, 1993, and Markee, 1997) which incorporated different theories of innovation and the concept of washback
Reliability	Use case study protocol Develop case study database	Data collection in phases 1, 2 and 3 of the study

To meet the test of *construct validity*, a researcher must cover two steps: defining and relating specific concepts to the objectives of the study; and identifying correct operational measures for these concepts (Yin, 2014). To achieve these goals, Yin has advised three tactics (Table 2):

- 1) using multiple sources of evidence to triangulate different sources of data;
- 2) establishing a chain of evidence when collecting data; and
- 3) presenting the draft case study report to participants for member-checks.

This will allow participants to review the results and make sure the conclusions drawn are appropriate and unbiased. These tactics can also help case study researchers achieve construct validity (see also Creswell, 2012).

In my study, I used multiple sources of evidence in the form of administrator and teacher interviews, and student focus groups to collect perspectives from key stakeholders. I compared my findings from teacher interviews with that of student surveys and focus groups. The quantitative data (on-line student surveys) complemented

the qualitative data (student focus groups and teacher interviews). Additionally, because I was in the field for a prolonged period, my field notes provided a rich description of the participants and the setting. Finally, the draft case study report was presented to and reviewed by two participating teachers as part of member checks to ensure the construct validity.

To meet the test of *external validity*, i.e., to know if a study's findings are generalizable beyond the study, Yin (2014) suggests seeking analytical generalizations. Posing 'how' and 'why' questions can help arriving at analytical generalizations. The purpose of this study was to explain if a high-stakes could be used to leverage positive washback in a changing testing regime. In my study, this question was not only pertinent to the EAP program, but to the phenomenon of washback and diffusion of innovation, which, in turn, will have much wider applicability than my study.

Reliability of a study means that if a future researcher were to conduct similar research following the same procedures, similar findings and conclusions would result (Yin, 2014). The emphasis, however, "is not on *replicating* the results of one case by doing another case study.... the goal is to minimize the errors and biases in a study" (Yin, 2014, pp. 48-49). The methods and results sections of my study are as descriptive as possible. Readers can compare my case at the EAP program to their future research context to decide on the transferability of this research themselves. To enhance reliability, a section of this study's qualitative data was sent to ten first-year PhD students for peer review and independent coding. This was done in order to ascertain inter-rater reliability. Within qualitative research inter-rater reliability is carried out by having two data passages, independently coded in order to see how close two people in how they view the same data. The closer the coding the inter-rater reliability is said to be high; when far

apart the reliability is said to be low. Inter-rater reliability was examined through the use of Cohen's Kappa and was determined to be satisfactory, with the Kappa Measure of Agreement at .82 and a significance of $p < .05$. According to Peat (2001), agreement above .7 is considered to be evidence of good agreement, above .8 is considered to be very good.

In summary, if case study researchers pose 'how' and 'why' questions, triangulate different data collection sources, establish chain of evidence, have their participants review the results of the draft case study report, and use case study protocol, the results of case studies should be relatively more analytically generalizable, transferable to other contexts, trustworthy, reliable, and valid.

Eliciting Case Data by Interviews and Focus Groups

Interview and focus group data are one of the most important sources of evidence in case studies because case studies are mainly used to obtain the descriptions and interpretations of others (Duff, 2012; Stake, 1995; Yin, 2014). Interviews and focus groups provide opportunities to explore what goes on in respondents' heads: participants can talk about memories, experiences, and inner thoughts. According to Seidman (2013), "the root of in-depth interviewing is an interest in understanding the lived experience of other people and the meaning they make of that experience" (p. 9). In applied linguistics, interviews and focus groups are incorporated in case studies, mixed-methods research, and ethnographic explorations to examine language use and experiences of language learners (Duff, 2012; Roulston, 2013).

Interviews. DeMarrais (2004) describes an interview as "a process in which a researcher and participant engage in conversation focused on questions related to a

research study” (p. 87). Interviewing is necessary in qualitative studies, where the researcher is focused on interpreting people’s behaviour (past or present) and feelings, but is unable to observe them directly. Multiple views of a case can be discovered and portrayed through interview participants. Thus, interviews can “add an inner perspective to outward behavior” (Patton, 2002, p. 109). Within social constructivist ideologies, interrogations provide a better chance for personalization and probing, making it a “construction or joint production by interviewer and interviewee” (Duff, 2012, p. 133). Duff also cautions that interview data produces a truth elicited in a particular time and space for a specific purpose; therefore, data gathered during this social interaction cannot be “taken as decontextualized, independent facts or observations” (p. 134), thus making non-generalizability a key feature of qualitative data.

Importantly, interviews by nature are *constructions* of data rather than *collections* of data in the strictest sense: an interviewer asks follow up questions based on the information provided by the interviewee, while the interviewee shapes his or her responses based on how the interviewer is perceived. For example, interviewees may present their responses differently based on whether the interviewer is male or female, aggressive or friendly, older or younger, etc. Because of this, the interview is likened to a “dance” of understanding rather than plucking bits of data from a tree.

Types of interviews. There are mainly three main types of interviews (Dornyei, 2007; Seidman, 2013; Stake, 1994):

Structured interviews are when a researcher follows a prescribed and elaborate interview schedule with set questions.

Unstructured interviews are when the interview is like an open-ended conversation and takes an unprepared direction with only minimal direction from the research agenda.

Semi-structured interviews offer compromise between the two extremes. Although there is a set of prescribed guiding questions and prompts, the format is open-ended and “the interviewee is encouraged to elaborate on the issues raised in an exploratory manner” (Dornyei, 2007, p. 136). In semi-structured interviews, the interviewer provides guidance and direction (the “-structured”), but is also keen to follow up on interesting turning points in the respondent’s narrative and to let the participant elaborate on certain issues (the “semi-”).

According to Dornyei (2007), the semi-structured interview is most suitable for researchers who have a strong overview of the phenomenon or domain of inquiry. A researcher in a semi-structured interview develops broad questions about the topic in advance, but does not use “ready-made response categories that would limit the depth and breadth of the respondent’s story” (p. 137). Therefore, the interviewer will ask “the same issue questions to all of the participants, although not necessarily in the same order or wording” (p. 137). The interviewer may also supplement the main questions with various probes. Patton (2002) suggests that in semi-structured interview questions, a probe or a contrast probe should ask about how a particular experience/feeling/action compares to other similar concepts. Using probes increases the richness and depth of responses. Probes can also include detail-oriented and clarification questions. Most importantly, probes should also be used to corroborate what a respondent is relaying by asking for negative instances of something, e.g., when something related to the phenomenon did not go as planned.

The semi-structured approach of interviewing is best for case studies because this approach uses predetermined, but flexibly worded questions, the answers to which provide “tentative answers to the researcher’s questions” (Algozzine & Hancock, 2006, p.

40). Also, follow-up questions can be asked to probe in more depth the issues in which the researcher is interested. Semi-structured interviewers invite participants to freely express themselves from their own perspective.

Semi-structured interviewing was an appropriate approach to data collection for the present longitudinal case study research. My research explores the phenomenon of washback as testing regime change in the chosen EAP program. I already had a strong overview of the washback phenomenon in language testing and wanted to further explore how an introduction of a new high-stakes testing regime affected teaching and learning in the program. Since my study was highly contextual, i.e., related to one program, being a teacher in the program gave me a good vantage point from which to study the phenomenon of washback from different angles. Semi-structured interviewing also became a reliable source of data triangulation (Yin, 2014) or concept resonance because of my direct and easy access to all key stakeholders of the program.

Focus groups. Instead of one-to-one interviews, case studies also often use focus group interviews to receive several people's perspectives, sometimes called 'multivocality' (Duff, 2012), on an issue in a fairly short time frame. Group interactions are generally more informal than one-to-one interviews and may prompt others to comment on themes that they may not have thought about earlier. They are highly versatile, flexible, and information-rich. According to Dornyei (2007),

The focus group format is based on the collective experience of group brainstorming, that is, participants thinking together, inspiring and challenging each other, and reacting to the emerging issues and points. This within-group interaction can yield high-quality data as it can create a synergistic environment that results in a deep and insightful discussion (p. 144).

Focus groups also follow an interviewer's guide or protocol, whether a structured, unstructured, and/or semi-structured approach. Like semi-structured interviews, focus groups include both open- and closed-ended questions. However, the topics discussed are not sensitive, personal, or culturally inappropriate. The focus groups work best for "topics people could talk about to each other in their everyday lives – but don't" (Macnaghten & Myers, 2006, p. 65). According to Patton (2002), gathering data from different focus groups yields a variety of perspectives, which increases the researcher's confidence in interpreting the emergent findings. Using student focus groups in my study helped me to find a diversity of views regarding testing regime change among the same group of students and across groups of different students.

Organizing focus groups. To organize a focus group it is important to consider three things: the selection of participants, the drawing up of an interview guide, and the role of the moderator. First, the selected participants should be "information-rich" (Krueger & Casey, 2000), i.e., participants should be able to provide a great deal of information about the issues that are central to the purpose of the research (Patton, 2005). If there are too few participants, the data generation may not be sufficient; if there are too many participants, everybody may not get enough time to contribute meaningfully to the conversation. Dornyei (2007) suggests in any research project it is standard practice to run several focus groups. This helps in mitigating idiosyncratic results that can occur because of either internal or external factors affecting group dynamics. For this study, I tried to run focus groups for students at least once during each semester this study took place, totaling four focus groups over the course of the entire study. To maintain the diversity of perspectives in the focus groups, I selected participants from different cultures and linguistic backgrounds, but always from the same level in the EAP program.

A second consideration, when organizing focus groups, is to design the interview guide with a thorough understanding of the research agenda and the purpose of conducting focus groups. Ho (2013) suggests that questions should be sequenced in such a manner that general, unstructured questions come first and more specific ones at the end. I followed Ho's (2012) suggestion and began the focus group interviews with general unstructured questions before moving on to specific questions.

Study Context

This present study was conducted in a Canadian university's EAP program. I chose the EAP program purposefully as I have previously worked as a part-time English language instructor in this program and had access to teachers, administrators and students. This program offers English language academic support for students planning to enter this university's academic programs at the undergraduate or graduate levels. In addition to providing language support, this EAP program also familiarizes students with Canadian culture through socio-cultural activities. The program is an accredited member of "Languages Canada"⁴. The program offers a full-time curriculum of three sessions a year with four language levels ranging from beginner ESL to advanced academic English. Students with conditional admission offers from the university's regular academic programs are placed in one of the EAP program's four different levels based on the results of standardized tests such as International English Language Testing System (IELTS), Test of English as a Foreign Language (TOEFL) iBT, CanTest, etc. For new

⁴ Canada's national language education association representing more than 201 language education programs across Canada.

students without a conditional offer from the university, placement is determined by the university's in-house online test.

Most students in the EAP program are from China and Arabic-speaking countries such as Saudi Arabia and Libya. Classes run for 21 hours of instruction per week. The teaching load of 21 hours is divided among two teachers and a teaching assistant (TA). The core teacher teaches nine hours per week and generally attends more to writing skills than to other language skills such as reading and listening. The other teacher teaches six hours per week and focuses more on listening and reading skills. TAs, usually Masters' and Ph.D. students at the university, are commonly responsible for reinforcing the two teachers' lessons, while also providing socio-cultural support through the arrangement of outings in the local region.

To pass the EAP program, in both testing regimes students at the graduating level (GL) had to take a high-stakes test. Results of this test determined if students could commence their full-time academic studies in this university.

During the Fall 2016 term, the testing regime in the program was changed. Table 3 describes the differences between the former and new tests. According to an official memo sent by the administration in August 2016, the aim of the assessment updates was to provide students with more authentic academic experiences and to improve the validity of the program's assessments.

Table 3 *Comparison of former and new test structures*

The ExitTest	The ISTs (Sept 2016)	The ISTs (August 2017)
Reading MCQs: Skimming and Scanning test Reading comprehension Cloze test	Reading-to-Write test: One/two reading passages + Two writing tasks: 1. Summary writing for one reading passage 2. Response essay	Reading/Listening-to-Write test: One reading passages + One listening text + Two writing tasks: 1. Summary writing for any one (reading/listening) 2. Responding to any one reading/listening
Writing: Timed- impromptu five paragraph essays on genres e.g., comparison and contrast, argumentative, opinion, cause and effect etc.		
Listening MCQs: Listening to different genres e.g., dialogues, university lectures, and instructions. (listening tasks were played twice)	Listening-to-Speak test: Listening to one/two audio/video texts Speaking tasks: 1. Summarize one listening 2. Responding to one of the audios	Speaking: Individual interviews with teachers
Speaking was not assessed as part of this exam		

Under the new regime, instead of separate multiple-choice reading and listening tests, the final tests took the form of integrated skills tests (ISTs) where students read and listened to a text, and then wrote an essay based on the knowledge gained from the texts. Instead of decontextualized content in the former tests, the content used in the new tests was taken from the course textbooks. Speaking skills were not assessed in the former testing regime, but in the new testing regime teachers were to formally test speaking skills. However, the test format for assessing speaking changed considerably from first introduction in September 2016 to August 2017 (see Table 3).

Further, the new ISTs were developed and delivered in-house, within the EAP program, instead of by the external testing office of the university. However, there was no change in the exit criteria for the GL students. Their final exam writing performances were still reviewed externally by the testing office of the university.

Participants

Aligning with case study method, this research included multiple data sources (interviews, surveys, and focus groups) from multiple participant groups (administrators, teachers, and students) in order to gather rich and thick descriptions of the context and different perspectives relating to the phenomenon of washback-as- change. The key informants of this study were two administrators, 15 teachers, and 201 students in the EAP program. I conducted individual interviews with administrators and teachers, and focus groups with students before, during and after the testing regime change. Interviews focused on participants' views of how the former and new testing regimes affected classroom teaching and learning. The rationale for selecting these stakeholders is explained next.

First, to study washback, Watanabe (2004) suggests it is normal to select various groups of participants rather than selecting one single population because there is the potential that some aspects of washback “exist for learners but not for teachers, whereas other aspects exist for teachers but not for learners” (p. 29).

Second, looking at the same situation from various angles, e.g., using more than one data set would also be methodologically sound for triangulation purposes (Bailey, 1996; Stake, 1995; Wall, 2012; Yin, 2014). Using three different groups of stakeholders was expected to yield different perspectives on the effects of testing regime on the program. By gathering data from different stakeholders on the same topic, evidence that confirms or denies participants' stated beliefs about assessment practices could be detected (Creswell, 2012; Denzin & Lincoln, 2011).

Third, since qualitative research allows the actual voices of the participants to be heard (Scott, 2007; Wall, 2005), researchers (Scott, 2007, Watanabe, 2004) have

advocated for in-depth insights into attitudes of different stakeholders. Such insights can obtain a rich and multi-layered account of potential differences in perceptions of washback.

Finally, although many washback studies such as Alderson and Hamp-Lyons (1996) and Cheng (1997; 2005) have used observation of actual classroom teaching to explore and confirm participant's perceptions with their behavior, this study did not include classroom observation as a data collection method. The goal of this study was to explore washback in relation to change implementation, and not the confirmation of participants' perceptions with their behavior. The balance between a focus on content (that is, academic English) and/or classroom interactions, and implementation of language support a relation to the changes in testing regime.

Teachers

Participating teachers in this study were either Masters or PhD holders and had significant experience teaching in the EAP program (see Table 4). Most teachers had either TESL Ontario/TESL Canada or some equivalency of these accreditations. Further, as most teachers had their Masters in teaching English as a Second Language, they had some knowledge in second language acquisition and assessment. Ten teachers were interviewed in Phase 1 of the study. I interviewed these teachers towards the end of Winter 2016 and Summer 2016 terms. Not all teachers were teaching at the graduating level at that time, but they had previously taught at the GL. The use of peripheral sampling i.e. to include participants who are not central to the phenomenon under study, but neighbor to it (Miles & Huberman, 1994) was "to obtain contrasting and comparative information that may help understand the phenomenon" (p. 34). In Phase 2 of the study, when the changes in testing regime were introduced, seven teachers teaching at the

graduating level were interviewed. I interviewed these teachers twice in the semester. First, just after the implementation of the new ISTs in October 2016, and second at the end of that semester in December 2016 to better understand the implementation dynamics and the immediate washback of the newly introduced testing regime. In Phase 3 four teachers who taught at the graduating level at that time were interviewed in August 2017. The number of teacher participants varied each semester because of various factors such as the number of students enrolled in a semester, number of classes being offered, and teachers' seniority in the program. Only two out of fifteen interviewed teachers taught in all the phases of this study. 60% (9/15) of participating teachers were women and 40% (6/15) were men. Further, 60% (9/15) of teachers had at least eight or more years of teaching experience at the EAP program.

Administrators

The program manager and the curriculum coordinator of the EAP program were the two administrators who participated in this study (see Table 4). The program manager was interviewed once in Phase 1 in May/June 2016 and again after the change implementation at the end of Phase 2 in December 2016. The curriculum coordinator was interviewed three times, in total, once in each phase (Phase 1, 2 and 3) of the study, as he was more involved in the implementation of the new ISTs than the other program administrator. I conducted the last interview with him in Phase 3 in August 2017. The curriculum coordinator was also a teacher in the program. Both administrators had significant experience in teaching and administration in the EAP program. They both were with the EAP program for almost ten years.

Students

201 students at the graduating level of the EAP program participated in the study. The online questionnaire was administered four times during the entire period of study: twice in Phase 1, once in Phase 2, and once in Phase 3. Similarly, I conducted two focus groups in Phase 1, one in Phase 2 and one in Phase 3. In Phase 1, 137 students completed an online survey about the assessment practices in the former testing regime and six of these students participated in focus groups. In Phase 2, 64 students completed an online survey about the ISTs and three students participated in a focus group. In Phase 3, an online questionnaire was administered, but these data were not considered in this study because these students were new to the program (two or three semesters) and were not in a position to hold informed views on the former testing regime. The focus group in Phase 3 comprised of six students who had earlier participated in a focus group in either Phase 1 or Phase 2. This focus group data analysis is used in the presentation of results in Chapter 6. Unfortunately, the ethics cover did not allow me to have access to the test marks of the students in my study. However, I was able to recruit four students from Phase 1 (former testing regime) and two students from (new testing regime) in Phase 3 (the consequences phase) who volunteered to talk about their academic language development three to four semesters after completing the EAP program.

Table 4 lists the interview and focus group participants in this study. For the confidentiality purposes, all participants in this study were given a pseudonym.

Table 4 *Listing of interview and focus group participants*

Interview and Focus Group Participants			
Teachers			
Teachers	Education	Teaching experience in the EAP program	Data Collection by Phase (P) (i.e. P1, P2, P3) & Date
Alan	MA Communication, TESL	9 years	<i>P2- Nov'16 & Dec'16</i>
Ana	PhD (teaching & learning)	17 years	<i>P1 – April'16 and P3- August'17</i>
Ashley	PhD (ABD)	2 years	<i>P1 – July'16,</i>
Derek	PhD (ABD)	12 years	<i>P1 – July'16 and P3- August'17</i>
Gerald	MA	13 years	<i>P1 – April'16</i>
Jill	MA TESOL	8 years	<i>P1 – April'16, P2- Oct'16 & Dec'16, and P3- August'17</i>
Joshua	MA (App. Lin), TESLA	1.5 years	<i>P1 – April'16</i>
Kathy	MA (App. Ling)	10 years	<i>P2- Oct'16 and Dec'16</i>
Lisa	PhD (ABD)	2 years	<i>P1 – April'16, and P2- Nov'16 & Dec'16</i>
Mary	PhD (ABD)	3 years	<i>P2- Oct'16 & Dec'16</i>
Maria	MA (App. Ling)	2 years	<i>P1 – April'16, and P2- Oct'16 & Dec'16</i>
Raymond	MA	10 years	<i>P1 – April'16</i>
Ruth	MA (TESOL)	10 years	<i>P1 – April'16</i>
Stacey	MA (App. Ling), TESL, Sp. Ed	17 years	<i>P1 – April'16, P2- Oct'16 & Dec'16, and P3- August'17</i>
Travis	MA	2 years	<i>P2- Oct'16 and Dec'16</i>
Administrators			
Anderson	MA	10 years	<i>P1 – May'16, and P2- Dec'16</i>
Christopher	PhD	9 years	<i>P1 – August'16, P2- Dec'16, and P3- August'17</i>

Focus Group Students			
Name	Country	Going to:	
Phase 1			
			<i>P1 – April 2016</i>
Sheikha	Saudi Arabia	Graduate school	
Mohammed	Libya	Undergraduate	
Seiko	Japan	Graduate	
			<i>P1 – August 2016</i>
Tony	China	Undergraduate	
Sheng	China	Undergraduate	
Jessie	Indonesia	Undergraduate	
Phase 2			
			<i>P2 - December 2016</i>
Charlie	China	Undergraduate	
Valdo	Uzbekistan	Graduate	
Ali	Afghanistan	Undergraduate	
Phase 3			
			<i>P3 - August 2017</i>
Sheikha	Saudi Arabia	Graduate	
Mohammed	Libya	Undergraduate	
Charlie	China	Undergraduate	
Valdo	Uzbekistan	Graduate	
Johnny	Brazil	Undergraduate	
Sheng	China	Undergraduate	

Overall Research Design

This study made use of different stakeholders' (Hughes' *participants*) accounts to better understand the phenomenon of washback as change (Henrichsen's *diffusion of innovation*), which took place in an EAP program. The data were collected in three distinct phases (see Figure 3 below):

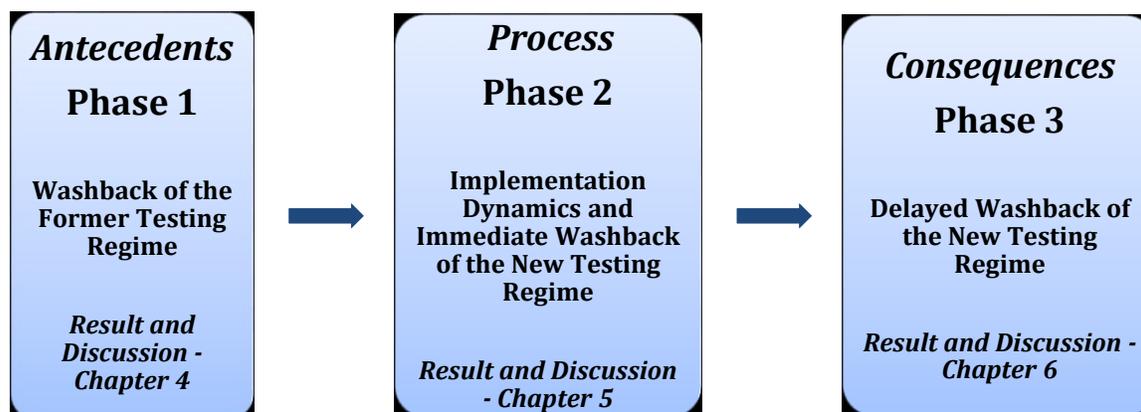
Phase 1: the Antecedents phase - former testing regime

Phase 2: the Process phase - Implementation Dynamics and Immediate Washback

Phase 3: the Consequences phase - Delayed Washback

A brief description of each phase with the relevant guiding questions follows.

Figure 3 Overall Research Design



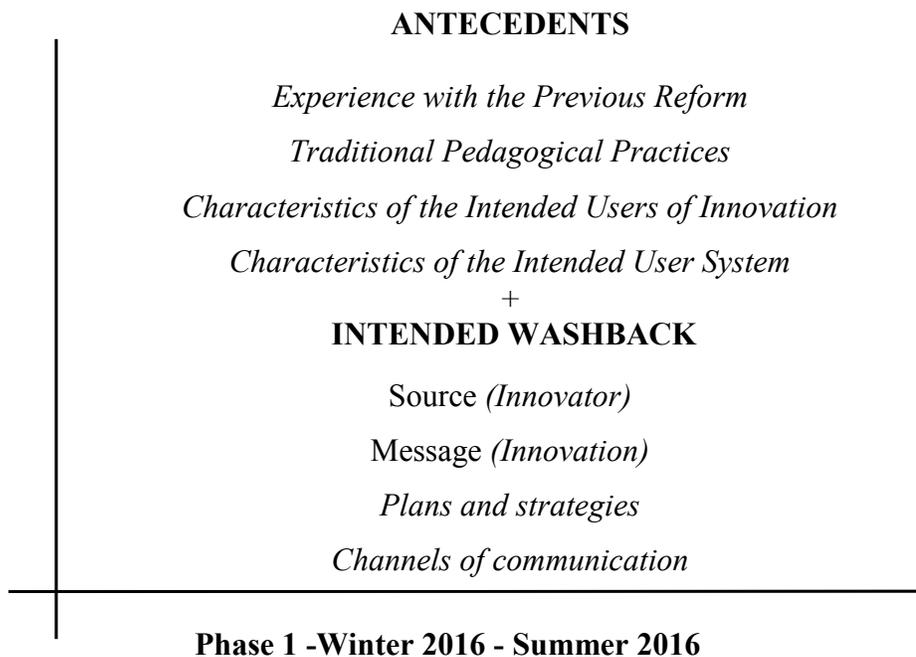
Phase 1: The Antecedents phase – Former Testing Regime and Intended Washback of ISTs

Phase 1 address Research Question 1:

What evidence is there of washback in the former testing regime and what is the intended washback of the new testing regime?

Antecedents of the situation, also called a baseline study, were required to understand the characteristics of how teaching, learning, and assessments were carried out before the introduction of the new ISTs (innovation). In keeping with Henrichsen's (1989) model and seeing high-stakes tests as a potential source of innovation (i.e., which leverage change), the research question was related to Henrichsen's antecedents and Markee's what and why (i.e., intended change and rationale for change). Phase 1 included (see Figure 4): Users (teachers, students, and program administrators); traditional pedagogical practices; plans and strategies of implementation; and the User system (the educational context of the EAP before the innovation was introduced). During this time, program administrators were planning to introduce testing regime change only at one level, i.e., the Graduating Level, of the program. Data collection for Phase 1 was over two semesters: Winter 2016 and Summer 2016.

Figure 4 Phase 1- Antecedents and Process Elements from Henrichsen (1989, p. 80)



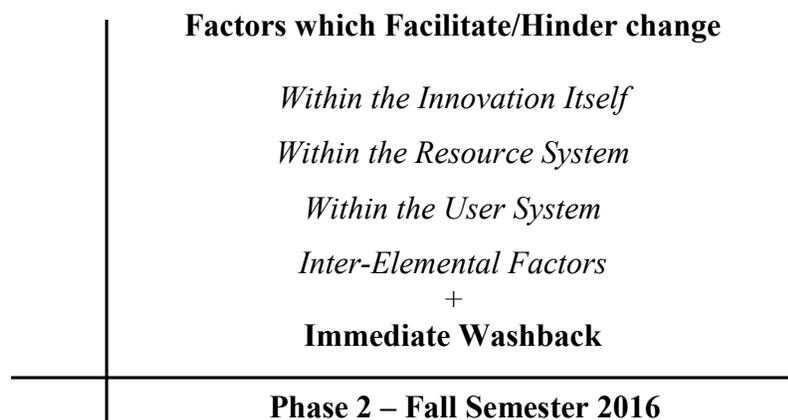
Phase 2: The Process phase – Implementation Dynamics and Immediate Washback of the New Testing Regime

Phase 2 was used to address Research Question 2:

What evidence is there of washback factors facilitating and/or impeding the implementation of the new testing regime?

The process of implementation of the new testing regime and immediate washback were the focus of Phase 2. This phase occurred in the Fall 2016 semester (Figure 5), when the new ISTs were first used. This phase included the implementation dynamics when the changes to testing regime were being introduced (September 2016 to November 2016), and the immediate washback when the first cycle of the implementation of the new testing regime was complete (December 2016). The evaluation criteria for the implementation dynamics and immediate washback were the following factors: Innovation (new ISTs), Receivers (teachers and students), and Factors that facilitate/inhibit the implementation dynamics (see Figure 5). Each of these factors was examined independently to determine the type of washback existed during the implementation period.

Figure 5 Phase 2 - Implementation dynamics and immediate washback (Process Phase)



Phase 3: The Consequences phase - Washback of the New Testing Regime Over Time

Phase 3 was used to address the Research Question 3:

What evidence is there of washback in the new testing regime?

Phase 3 focused on the examining the washback of the new testing regime after three semesters of implementation (until August 2017). The purpose of this phase was twofold: first, to explore what impact delayed washback from the new testing regime had on teachers; second, to study how students learning changed in response to washback from the new testing regime, and compare it to their learning under the former testing regime. This phase also helped in establishing washback attributes of the new ISTs, such as direction (positive and negative), intensity (strong or weak), and time (immediate or delayed).

Procedures

This longitudinal case study was conducted over a period of 20 months in three phases. The data collection procedures were guided by the study's research questions. I collected both quantitative and qualitative data from teachers, students and administrators in the EAP program (see also Table 5 for detailed study instruments, their goal and their relevant components in Henrichsen, 1989).

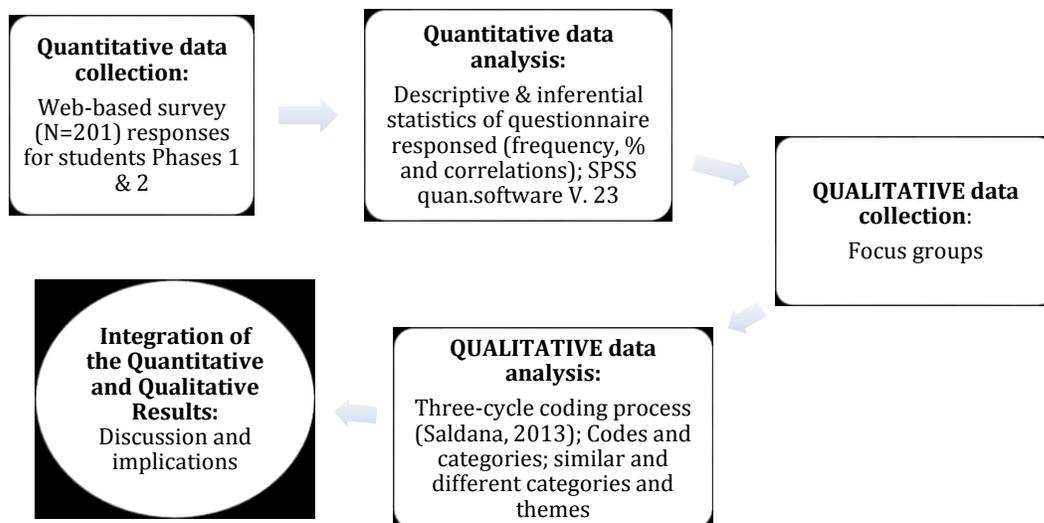
In Phase 1, the main instruments for data collection were:

- a) Teacher interviews (see Appendix B for teacher interview questions).
- b) Student online questionnaire (see Appendix C for student questionnaire).
- c) Two focus groups (1st in April 2016, and 2nd in August 2016) – (see Appendix D for student focus group questions).

- d) Administrator interviews (see Appendix E for administrator interview questions).
- e) Document analysis – Evaluation of test task characteristics of the previous ExitTest.

The teacher interviews were conducted to get a picture of teaching and learning in the EAP program prior to the testing regime change (see next section for interview data). In total, ten teachers were interviewed. The student questionnaires were used for an initial overview of students' accounts on assessment practices in the former testing regime. 137 students filled in an online questionnaire. The 40-item online questionnaires in Phase 1 elicited data on purpose, content, timings, mode, and marking of assessment (Biggs, 1991; MacLellan, 2001) (see Appendix C). A Cronbach's Alpha was run to determine the reliability of the items, with a result of .92 for 40 items. This meant that most items were measuring the same underlying concept of students' account of assessment practices in the EAP program. Items with the highest and lowest means were also identified. The online questionnaires were first piloted in March 2015, and feedback from the pilot was used to make changes in preparation for this study. Figure 6 explains the quantitative and qualitative data collection and analysis of student questionnaire and focus groups in Phase 1 and Phase 2 of the study.

Figure 6 Quantitative and qualitative data collection and analysis for student questionnaires and focus groups



Student focus groups were conducted as a follow-up for deeper insights from the questionnaire, and to explore the phenomenon of washback from the perspectives of students in the program. The administrator interviews in Phase 1 were to understand the intended washback of the new testing regime.

In Phase 2, the main instruments for data collection were:

- a) Teacher interviews (see Appendices F and G for teacher interview questions)
- b) Student online questionnaire (see Appendix H for student questionnaire)
- c) One focus group – (see Appendix I for student focus group questions)
- d) Administrator interviews (see Appendix J for administrator interview questions)
- e) Document analysis – Evaluation of test task characteristics of the new ISTs

Seven teachers were interviewed twice (1st in Oct/November and 2nd in December) during the Fall 2016 semester to understand how they were coping with the implementation dynamics of the testing regime change.

The online questionnaire for students contained questions about the new Reading to Write and Listening to Speak tests and the washback of these on students' test preparation and the type of feedback that they received from their teachers. 64 students filled in the online questionnaire. There were 14 items in this questionnaire. Tavakol and Dennick, (2011) suggest that the value of Cronbach's alpha is affected by the length of a test or scale i.e., "if the test length is too short, the value of alpha is reduced" (p. 53). Therefore, because of the fewer number of items in Phase 2 student questionnaire, Cronbach's alpha test was not performed (Streiner, 2003; Tavakol & Dennick, 2011). Similar to Phase 1, a focus group was conducted with three students to gain deeper insights about the washback of the new testing regime on students. The administrator interviews were conducted to understand their accounts of the change implementation process.

In Phase 3, the main instruments for data collection were:

- a) Teacher interviews (see Appendix K for teacher interview questions)
- b) One focus group – (see Appendix L for student focus group questions)
- c) Administrator interview (see Appendix M for administrator interview questions)

Four teachers and one administrator were interviewed to explore the delayed washback of the new testing regime. One focus group was conducted with six students. Four out of these six students were from the former testing regime and two students were from the new testing regime. These students had previously participated in Phase 1 and Phase 2

focus groups. The purpose of the focus group was to understand if students learning changed in response to washback from the new testing regime, and compare it to their learning under the former testing regime. Table 5 describes the timeline for each phase, instruments used, goals, and the relevant components from Henrichsen's (1989) hybrid model of diffusion/implementation.

Table 5 *Study instruments*

Time Line	Instruments	Goal	Relevant component in Henrichsen (1989)
Phase 1 Winter 2015 and Summer 2016	Interviews, online questionnaire and focus groups: 10 teachers 137 students online questionnaire 6 students focus groups 2 program administrators (curriculum coordinator, and program manager) Document analysis: <ul style="list-style-type: none"> - Previous ExitTest - Official communication - Researcher's field notes 	<ol style="list-style-type: none"> 1. Teacher background, teaching practices, teaching content: academic vs. general English, awareness of the changes, attitudes towards tests etc. 2. Student attitudes towards tests, language learning, goals, classroom practices, abilities, study practices etc. 3. Teaching/learning context, differentiation between different levels of EAP, resourcing, teacher support, planning for future 	Characteristics of users, & traditional pedagogic practices Characteristics of user system
Phase 2 Oct/Nov. 2016 Dec. 2016	Interviews, online questionnaire and focus groups 7 teacher: <ul style="list-style-type: none"> - mid-session interviews - end of the session interviews 64 students online questionnaire 2 Administrator interviews Student focus group Document Analysis <ul style="list-style-type: none"> - Integrated Skills Tests - Official communication - Researcher's field notes 	Attitudes towards new testing regime Classroom conditions Administrative support Descriptions of changes and change implementation such as integration of different skills, testing speaking and support provided Analysis of assessment tasks	Implementation Dynamics Factors which facilitate/hinder change in: <ol style="list-style-type: none"> 1. Innovation- originality, complexity, explicitness, practicality etc. 2. Resource system – capacity, structure, openness etc. 3. In intended user system- the adoption unit, openness, teacher and learner factor, assessment tasks etc.
Phase 3 July/Aug. 2017	Interviews, and focus group 4 teacher interviews 1 administrator interview 1 student focus group	If changes have any lasting effects on teaching/learning	Washback - adoption/rejection Washback – positive/negative Washback intensity- strong/weak

Interviews and focus groups data in Phases 1-3. To gather information from interviews and focus groups, I followed Algozzine and Hancock's (2006) six-step guidelines. I identified knowledgeable and information-rich participants and developed an interview guide with open-ended questions that provided insight into my research questions. I obtained ethics clearance (see Appendix N for ethics certificate) from my home university as well as the university where the study was conducted. For each interview or focus group I booked a private and distraction-free meeting room. I made the format and purpose of the interviews clear to all participants ahead of time, and obtained written consent (see Appendix N for teacher/educator consent form and Appendix O for student consent form) to proceed before audio-recording the interviews for later transcription.

For the teachers' and administrators' interview protocol, I borrowed and adopted Seidman's (2013) three-interview series method. In this method, three separate interviews are conducted to establish better rapport with participants. In turn, this provides more meaning and better understanding of participants' behaviours and words. This series of three interviews can also satisfy the need for triangulation in qualitative research in order to establish the confirmability of the findings (Lincoln and Guba, 1985). However, instead of keeping these meetings as close as possible (1 to 3 weeks), as suggested by Seidman, I met my participants over three to four semesters over the course of my entire study. My rationale for meeting my participants over a longer time period was to better understand the implementation of testing regime change. In the first interview with teachers, my aim was to learn about their pedagogical practices before the change in testing regime. Drawing from Henrichsen's (1989) process element, in the second interview, we talked about how the process of implementation of the new testing regime

was affecting them. In the final interview, we discussed whether the testing regime change had brought about any changes in their pedagogical practices. For students, instead of conducting individual interviews, I conducted focus groups with a different set of students every semester to understand their views and accounts of assessment practices in the EAP program.

As mentioned earlier, my approach to interviewing was semi-structured and I kept my questions fluid rather than rigid (Rubin & Rubin, 2012), thus satisfying the need of my line of inquiry and asking non-threatening questions to my participants. I asked probing questions, for example, taking a salient, specific word used by a participant and asking for elaboration. In focus groups, as a moderator, I made sure that every student in the group got a chance to interact and nobody dominated the floor to prevent any dominating or inhibiting group opinion, sometimes called ‘groupthink,’ by encouraging group members to think critically (Dornyei, 2007). I generally started focus groups with the purposes of the discussion and setting the parameters of interview in terms of length and confidentiality. I also explained the purpose of recording the interviews to students and clarified that discussion was about personal views and experiences and there were no right or wrong answers. I used probes whenever I felt that the students were shy. I always concluded the focus groups by asking for further clarifications and a positive note or feedback.

However, like any other research method, interviews and focus groups have their own weaknesses. First, with regard to interviews, Yin (2014) warns that interview data should be considered as verbal reports and interviewees’ responses could be subject to “common problems of bias, poor recall and poor or inaccurate articulation” (p. 113).

Similarly, Ho (2013) suggests that the data collected through focus groups can be sometimes biased due to group influence and dominant members. These need elaborate preparation to set up and the moderator has to perform several functions simultaneously. Further, it could be susceptible to “moderator bias and manipulations leading respondents to respond to his/her own prejudices” (Ho, 2013, p. 5). Thus, a reasonable approach to conducting research through these methods is to corroborate interview and focus group data with information from other sources. Table 6 presents sample interview and focus group questions used at the various phases of the study.

Table 6 *Sample interview and focus group questions*

Sample Questions	Phases/ Relevant to Henrichsen's (1989) Factors	Participants	Research Questions		
			RQ1	RQ2	RQ3
What are the University's expectations of the EAP program?	P1/ user system	Administrators	×		
What are some of the traditional assessment practices in our program?	P1/pedagogical practices	Administrators	×		
Are you aware of any previous reforms or changes to this program? If yes, please explain them in detail	P1/user system	Administrators and Teachers	×		
What are the current exit criteria for the GL students?	P1& P2/user system	Administrators and teachers	×	×	×
Do you think a change (whether in teaching, learning or assessment) is required in our program? Why?	P1/user and user system	Administrators and teachers	×		
What have you learned about the new tests at GL that you didn't know last year (Fall/Summer/Winter)?	P2/user	Teachers		×	
What factors have facilitated or hindered the implementation and diffusion of ISTs (specifically innovation in testing regime)?	P2 and P3/user system/diffusion process	Administrators Teachers		×	
Tell me about assessment activities that take place in your class. What factors do you keep in mind when you plan these?	P1, P2 & P3/user	Teachers	×	×	×
Do you think that the program learning outcomes prepare students for the final test? If yes, how and if not, why not?	P1, P2 & P3/user system	Teachers and administrators	×	×	×
How satisfied are you with the support provided by the administration regarding the new testing regime?	P2 & P3/users and user system	Teachers		×	×
How do you perceive this change? As a top-down process or a bottom-up one?	P2 & P3/user	Teachers		×	×
On a scale from 1 to 10, where 10 is the best and 1 is the worst, how good is this EAP program for you? How much would you rate it?	P1, P2 & P3/user system	Students	×	×	×

What kind of test related activities, exercises, tests, advice did you teacher provide you for your final exam?	P1 and P2 /user & Process of implementation	Students	×	×	
<p>Please rank order these activities that you think are the most prominent in your class:</p> <ol style="list-style-type: none"> 1. Listening to teacher talking to whole class 2. Reading texts (from books or other materials) 3. Reading and writing short or long answers to questions 4. Memorizing vocabulary and practicing grammar 5. Spending time in practicing summarizing, paraphrasing and referencing 6. Understanding all aspects of the final test (e.g., goals, content, format and rating) 7. Spending time in class practicing writing, such as a 5-paragraph essay 8. Spending time in class practicing practice-tests so that students are familiar with the test format (e.g., structure, vocabulary, and cloze-exercises etc.) 9. Spending time in class practicing group discussion 10. Spending time in class practicing note-taking (reading and listening activities) 	P1, P2 & P3/user, traditional pedagogical practices/Implementation / Immediate Washback	Students	×	×	
How many extended essays or research papers have you worked on this term?	P1 and P2 /user system	Students	×	×	

Document Analysis. Document analysis was another source of data collection for answering the present study's research questions. Yin (2013) advises, "the most important use of document analysis is to corroborate and augment evidence from other sources" (p. 107). Therefore, two types of documents were analyzed to corroborate interview data: a) the official communication regarding the testing regime changes in the EAP program and b) the test tasks evaluations of the old and new tests. These test tasks evaluation, along with interviews, was also helpful to answer research questions about washback, i.e., the effects of test tasks on classroom teaching and learning.

Quantitative Data Analysis

Data analysis in this study consisted of both quantitative and qualitative data analysis. The survey data from students' online questionnaires were automatically saved into Excel spreadsheets as a summary of responses in Google Drive. These Excel spreadsheets were imported into SPSS version 22.0 and Version 23.0. First, each Likert-scale answer for the questionnaire items was given a value. For example Item 5 on the student questionnaire in Phase 2, about understanding the course learning outcomes, were given values as: Definitely Yes – 1, Yes – 2, No opinion – 3, No - 4, Definitely No- 5. To understand students' accounts of assessment practices in the EAP program, a descriptive quantitative analysis (mean, standard deviation and range) was conducted. Also, in the questionnaires, there were a few items with reverse wordings. For example, the questionnaire relating to the former testing regime, item 26, was about students not understanding the assessment criteria in the EAP program. For such items, reverse coding was performed. Also some questions (e.g. questions related to students' uses of English outside the classroom in their native countries) from the questionnaire relating to

implementation phase of the new testing regime were repeated in the questionnaire in Phase 3. To check the quality of the student questionnaire, Cronbach's alpha was used. Correlation coefficients were computed to assess the relationship between students' accounts and assessment practices in the EAP program.

Qualitative Data Analysis

For the qualitative data analysis, the main data sources were individual interviews, focus groups, field notes, and document analysis. Data analysis of these was an iterative process as outlined by Merriam (2009):

Qualitative research is not a linear step-by-step process. Data collection and analysis is a simultaneous activity in qualitative research. Analysis begins with the first interview, the first observation, the first document read. Emerging insights, hunches, and tentative hypotheses direct the next phase of data collection, which in turn leads to the refinement or reformulation of questions and so on. It is an interactive process throughout that allows the investigator to produce believable and trustworthy findings (p. 151).

In my twenty-month long data collection period, I analyzed interviews in Phase 1 to develop subsequent questions for Phase 2 and Phase 3. Similarly, sustained contact, by interviewing two teachers in all the phases, helped me to write with confidence about their experiences over the entire course of my study.

Coding. According to Saldaña (2013), coding is not as precise a science as that used in quantitative analysis, but it is “primarily an interpretive act” (p. 4). Derived from the original Greek meaning “to discover,” coding highlights the constructivist dimension of research. Saldaña (2013) suggests, “the act of coding requires that you wear your

researcher's analytic lens. But how you perceive and interpret what is happening in the data depends on what type of filter covers that lens" (p. 6). Also, the nature of a main research question influences the specific coding choices we make (Saldaña, 2013; Yin, 2014) as research questions "embed the values, world view and direction of an inquiry" (Saldaña, 2013, p. 60). Coding often involves looking for repetitive patterns or consistencies within data and across data sources. A pattern can be characterized by similarity, differences, frequency, sequence, correspondence and causation (Hatch, 2002).

Because of the cyclical nature of data analysis, I adopted a three-cycle coding approach for the interviews and focus groups using mainly deductive data analysis techniques from Merriam (2009), Saldaña (2013), and Schulz (2012). In addition to using computer tools such as Microsoft Word and Excel spreadsheets for data analysis, I used the traditional time-honoured methods of data analysis such as file folders, index cards, sticky notes, and summaries on poster size papers. I also wrote analytical memos along with the first and second cycle coding because of the vast amount of textual data in my study.

Saldaña (2013) classifies an extensive list of First Cycle codes into categories, such as grammatical methods, elemental methods, affective methods etc. Second Cycle coding methods, which are advanced ways of reorganizing and reanalyzing data coded through first cycle coding, include pattern coding, axial coding, and procedural/theoretical coding, among others. For the first cycle coding, I relied on the grammatical, elemental, and affective methods of coding because the exploration of participant actions/process and perceptions found within the data could be answered with these coding methods. These methods got at the phenomenon of washback and

innovation theories at different levels in order to create a richer picture of the implementation process. A brief description of these methods is given next:

Grammatical methods do not refer to the grammar of language, but to the basic grammatical principles of a technique e.g., attribute coding for demographic characteristics; magnitude coding for intensity or frequency of exam related activities; and simultaneous coding where two or more codes overlap for different data sets (Saldaña, 2013). This coding method was prevalent in my coding because I was exploring the process of implementation specifically focusing on the evidence of washback in students' and teachers' interviews.

Saldaña (2013) suggests using *affective methods* for subjective qualities and core motives for human actions. For example, emotions and values codes are used to understand the inner cognitive system of participants; evaluative codes are used to acknowledge conflict and for judging the merits and worth of a program or policies. This coding method was useful for me because emotions and feelings serve as motivation for determining whether to get through a process or address a challenge. My study about the phenomenon of washback focused on how changes in a testing regime impacted stakeholders' beliefs, values, and perceptions regarding teaching and learning. Thus, there were emotions tied to what the participants were going through and, at times, these directly affected the implementation process.

Elemental methods consist of structural, descriptive, in vivo, process, and initial coding. These are mainly used for creating basic, but focused blocks of analysis upon which later cycles of coding are built, e.g., to summarize in a word or phrase a large chunk of data (except in vivo coding, where participants' words are used verbatim).

Descriptive codes identify basic labels of the key topics that are uncovered in the interviews. Structural codes connect a conceptual phase that forms part of the topic under inquiry to particular chunk of data that, in turn, informs a particular research question. Ultimately, structural codes told a story (or the nature) of the phenomenon of washback and theories of diffusion of innovation at a very high level so that whole could be seen.

In second cycle coding I looked at recurrent codes across interviews and/or focus groups to develop categories. I wrote analytical memos to keep track of my interpretations and as a reminder of what to include in the future interview questions in different phases of the study (see Appendix P for examples of sample coding with memos).

Procedural or theoretical methods of coding refer to using empirical, conceptual, and theoretical frameworks with pre-established coding systems. Procedural coding methods are very prescriptive and top down in nature. This method was the most used method in my coding. I used many factors from Henrichsen's (1989) model. In the first cycle coding, I identified all the data that was related to a particular factor in different stakeholder's accounts and brought that together to either compare or contrast whether the washback was positive or negative (see Appendix Q for sample division of categories). I also examined the facilitative or inhibitive roles of different factors (see Henrichsen, 1989, Figure 1) during the change process.

Codes derived from the first and second cycle coding were broadly divided into three main sections: former testing regime, implementation dynamics, and the new testing regime. Each of these was further categorized in such a way that they referred to different factors of Henrichsen's model and to the concepts of washback from my literature

review. Whenever administrators, teachers or students made comments about themselves, the EAP program, or the environment they were working in, the comment would receive a code or codes from the following sections.

In Phase 1, I corresponded the Antecedents of Henrichsen with the influences of former testing regime (discussed in Chapter 4). I used the evaluative criteria of Traditional Pedagogical Practices, characteristics of the user system and the users of innovation i.e. the teachers and students in the EAP program. The washback section contained teachers' and students' comments about the old testing regime and these were coded according to their effect on participants, themselves, the processes of teaching and learning and/or the products of teaching and learning (Hughes, 1993).

In Phase 2, during the implementation (discussed in Chapter 5) process, my goal was to investigate the factors that facilitated or inhibited the implementation process. To do this, I used the evaluative criteria of Henrichsen's (1989) four main factors: within the innovation itself; within the resource system; within the intended user system; and inter-elemental factors (for a detailed description of these see Chapter 2).

For the triangulation and inter-rater reliability of my analysis, I provided before- and after-change excerpts of student focus group transcripts to a class of ten PhD (Applied Linguistics) students at Carleton University. These students first conducted independent first-cycle coding on the provided data and then, in groups, conducted second-cycle coding. We, then, compared all of our second-cycle codes for similarities and differences. As a result of this activity, I became confident of my second cycle codes, their categorization, and the inferences drawn in relation to washback of both testing regimes in the EAP program.

The aim of the final third cycle coding was to arrive at a coherent understanding of the case in its entirety once all the data from all the stakeholders were coded. In this coding cycle, I worked on the final themes from the categories generated in the second cycle coding (see for example, Appendix R for sample division of codes, categories and themes of student focus groups in Phase 1 and Phase 2). These themes are discussed in detail in Chapters 4, 5 and 6. For analytical purposes, I used different types of matrices such as charts and tables with codes for “at-a-glance” format for reflection, verification, conclusion drawing (Miles et al, 2014). For example, I organized the data in a case-level meta-matrix (Table 7) that juxtaposed all of the single-interview data on an Excel spreadsheet.

Table 7 *Meta-analysis of teachers' interview data*

Coding across participants	Groups	Teachers	P1 Antecedents	P2 Process	P3 Consequences	Explanation for data analysis
	1	Stacey Jill	× ×	× ×	× ×	<i>Anchor teachers:</i> Taught in all phases of the study
	2	Derek Ana	× ×		× ×	<i>Anchor teachers:</i> Taught only before and after change
	3	Lisa Alan Mary	×	× × ×		Taught during the implementation period
	4	Raymond Joshua Ashley Maria Kathy Travis Gerald Ruth	× × × ×		× ×	Taught during different phases of the study

From the Excel spreadsheet, I looked for relationships in the data within a single interview, and then checked to see if these relationships existed within and across other interviews. For instance, for meta-analysis, I divided teachers' interview data into four different groups. Groups 1, 2, and 3 were the primary source for data analysis because these teachers were interviewed more than once during different phases of the study. During this process, using the research questions and relevant literature, I compared and contrasted the themes that emerged from their interviews for similarities and differences. Another example of a detailed juxtaposition is students' focus group data (e.g. Appendix R) where I compared the washback of the former and new testing regime on students.

Document Analysis: Evaluation of Test Tasks Characteristics

I analyzed the element of test task characteristics of the ExitTest in Phase 1 and the Integrated Skills Tests in Phase 2 of implementation of regime change as the tasks used in tests can greatly influence how students perform. This was necessary for my study because it was important to understand *what* was going to change. Without understanding the differences between the tasks of two tests (former and new), it would not have been appropriate to discuss the washback from either of them. The tests were thus evaluated using Bachman and Palmer's (1996, 2010) framework of test task characteristics, which represents the "potential relationships between task characteristics and test performance" (Purpura, 2004, p. 113). The framework describes how test tasks can be drawn from Target Language Use (TLU) tasks as samples of TLU domain. The framework includes a set of features that describe five aspects of test tasks: *setting of the test, test rubrics, input, expected response, and relationship between input and response*. I evaluated a sample of the ExitTest and the ISTs in an attempt to objectively analyze the

tests against the standards set by Bachman (1990) and Bachman and Palmer (1996, 2010).

Setting and test rubric. The first two facets of the test task description framework describe testing situations (such as the physical environment) test instructions, and scoring methods. I excluded these two from my analysis of the ExitTest because there were no data in my study regarding these; external testing experts were in charge of the exam. As the ISTs were in-house tests, the following is an explanation of the characteristics of the setting and the rubrics:

1. Both of the new tests (Reading-to-Write & Listening-to-Speak) were administered in computer lab. Each test taker was provided with a PC with a microphone for the speaking test. It was expected that all students are familiar with PC, Internet, and recording devices.
2. Test administrators (teachers, in case of the ISTs) were also expected to be trained in computer-based testing and were expected to have a positive attitude towards the test-takers.
3. The computer labs were booked in advance so that all sections of the GL could take the exam at the same time.
4. Instructions of rubrics⁵.
5. Structure of rubrics- refers to how the parts of the test procedure are put together and presented to the test-takers. Each test explained different tasks e.g., Task 1 in writing stated: summarize in your own words..... according to the article, and Task

⁵ This is discussed in detail in the factors that facilitate/hinder change under 'the complexity of the innovation'

2 stated: “For this task, you MUST draw upon this reading, the readings from class and your experience with the topic to the following prompt...”.

6. Recording method of rubric – A numeric score was given for all the test tasks.

Input facets. The input facets of test tasks consist of the format in which the input is presented and the nature of the language used in the input. The characteristics of the input are critical in all tests and TLU tasks, as test-takers have to process the input in order to appropriately respond to a task. The types of input can be either an item or a prompt. The purpose of an item is to obtain a selected (multiple-choice), limited (gap-filling), or an extended response (e.g., an essay or a dialogue).

From the perspective of an input-facet, the ExitTest contained multiple-choice questions (MCQs, recognition type) on the listening, reading, and cloze tests. Speaking was not assessed in this exam, and so the only extended response on this test was an impromptu essay. The critical part of the input is the language used to deliver the input, which not only involves language knowledge, but also the topical knowledge of the test-takers. The language characteristics include organizational and pragmatic characteristics (Bachman & Palmer, 1996, 2010). The organizational characteristics can be grammar, vocabulary, textual cohesion etc., and pragmatic characteristics, including functional and sociolinguistics features that are linked with the goals and context of the communication.

The strength of the ExitTest was the topical characteristics of the language of the input, as the content of the ExitTest was drawn from various academic situations encountered in typical North American university life. Messick (1996) describes language tests with potentially beneficial washback as those, which include

authentic and direct samples of the communicative behaviours of listening, speaking, reading, and writing of the language being learnt. Ideally, the move from learning exercises to testing exercises should be seamless. As a consequence, for optimal positive washback there should be little difference between activities in learning the language and activities involved in preparing for the test (pp. 241-242).

Aligned with this, the test readings were from authentic academic texts like textbooks or university calendars, while listening tasks were either simulated classroom lectures or discussions between professors and students. The writing prompt elicited an essay of an academic genre, e.g., compare and contrast, cause and effect, or argumentative. Such tasks can invoke not only language-specific skills, but also non-linguistics skills, such as those that are difficult to separate from language abilities, including: foundational skills, such as organizing or planning; cognitive skills, such as problem solving; critical thinking skills, such as comparing, contrasting, identifying, and evaluating main and supporting ideas; and learning methods, such as questioning (Messick, 1996).

From the perspective of an input-facet, neither test contained any selected responses such as MCQs, and/or limited responses such as gap filling. All responses in ISTs were constructed responses, i.e., the test-takers wrote a summary and an extended response to a prompt for the RW exam and orally summarized and responded to a given prompt for the Listening-to-Speaking test. As far as the language of input was concerned, topical knowledge in addition to organizational and pragmatic characteristics of the TLU was required of the test-takers.

Expected response. The fourth facet of test tasks refers to the expected format in which a response is to be produced and, similar to the input, the nature of language used in the response. According to Bachman and Palmer (1996, 2010), the types of responses can be selected response, limited production and extended production. In selected response, no production is needed, whereas in extended production, a response can be in the range from two sentences to a whole composition.

With respect to the facet of the expected response, the ExitTest's listening, reading, and cloze questions were MCQs and all responses provided by the candidates were necessarily in the form of selected responses. However, the writing exam was an extended production of a 300-350 word impromptu essay on a decontextualized prompt. From the TLU point of view, one weakness of this exam was the decontextualized nature of the writing prompt. In academic settings, any written response produced by students is in response to some stimulus presented through readings, lectures or classroom discussions. Also, timed-writing essays generally follow a set form: introduction, body, and conclusion, which can make the "test landscape narrow and artificial because [test-takers] respond to a form [rather] than meaning" (Fox, 2001, p. 268). Such essays do not typically occur in academic settings.

With respect to the facet of the expected response, the ISTs had no selected or limited responses. There were only constructed responses for writing as well as speaking test. Students were expected to have topical knowledge of academic conventions in addition to knowledge of organizational. Pragmatic characteristics for communicative purposes in speaking were required.

Relationship between input and response. The fifth and the last facet of test tasks according to Bachman and Palmer's (1996, 2010) framework refers to the relationships between input and response –i.e., the reactivity, the scope of relationship, and directness of relationship. First, *reactivity of a task* is the effect of an input or response to the subsequent input or response. It can be measured by a) the presence of feedback, and b) the interaction between the two interlocutors. Bachman and Palmer (1996) say “a typical example of a reciprocal test task would be give-and-take that occurs in a face-to-face oral interview” (p. 55). Since speaking is not assessed in the ExitTest, it was considered a non-reciprocal test task. The reactivity of the ISTs (in Phase 2) was externally interactive because a speaking test was conducted in a computer lab, and so there was no directness of relationship like the interaction of an oral interview. However, speaking was assessed through oral interviews in the new format of the ISTs (Phase 3), as observed in Phase 3, so the interaction was reciprocal there.

Second, the *scope of the relationship* is explained as the necessary range or amount of input required of the test taker to respond as required. Bachman and Palmer classify this relationship into two types: broad scope, i.e., requiring a lot of input to process in order to respond to a task; and narrow scope, i.e., requiring a limited amount of input. Broad scope tasks are reading and listening in a second language, while a short stand-alone grammatical item can be an example of a narrow scope. The ExitTest tasks can be categorized under both these categories: reading comprehension and listening tasks are examples of broad scoped tasks (where input requires comprehending main and supporting ideas), while skim and scan questions are examples of narrow scoped tasks (where students look for specific information in an academic brochure or calendar).

Similarly, broad scope of relation was observed in the ISTs because a lot of input was required to process the written and/or oral responses.

Thirdly, the *directness of relationship* refers to the degree to which a response relies on the contextual information available in a task. Messick (1996) explains direct assessments involve open-ended tasks where respondents are free to perform the complex skill, unblocked by structured forms or restrictive response formats. Thus avoiding the respondents' need to guess or make things up. For example, looking at a picture and describing it is an example of a direct relationship in a speaking task, while giving opinion on a subject is an example of an indirect relationship. An indirect relationship requires a test taker to rely on their topical knowledge.

In the ExitTest, the reading and listening test tasks are examples of a direct relationship between the input and response because students either read a piece of discourse or listen to a lecture and then answer comprehension questions that request explicit information in the passage. On the other hand, writing a response on a given prompt in the writing section of an exam is an example of indirect relationship where the students require pragmatic and topical knowledge in order to respond. In the ISTs, a mostly indirect relationship was noted because a lot of topical and pragmatic knowledge was required of the test-takers to complete all the test tasks. Again, in the revised version of ISTs, in Phase 3, speaking tasks were of a direct relationship because students looked at a picture and described it.

In sum, Bachman and Palmer's framework has provided a means of describing the test tasks of both the former and new tests and allowed me to consider individual characteristics of these tasks in order to highlight the potential interaction between the

test method and test performance. Purpura (2004) states this framework is also useful in predicting “the degree to which we are justified in claiming that score-based inferences about language ability can be generalized to non-test-taking instances of language use” (p. 126).

Chapter Summary

This chapter described the methodology adopted in the study. I positioned the study within a case study framework and explained how I selected a single case design with four embedded units of analysis. This was followed by a description and justification of the criteria used in selecting participants and instruments. I then explained the study context, participants, the procedures for data collection and analyses. The next chapter will explain Phase 1 in detail. I will describe the former testing regime, its washback and the intended washback of the Integrated Skills Test (IST) as envisioned by the administrators of the EAP program.

Chapter 4: Results and Discussion of Phase 1: the Antecedents phase (*Former Testing Regime and Intended Washback of ISTs*)

Introduction

In this chapter, I present the Antecedents, which were the conditions in place before the introduction of the Integrated Skills Tests, and the ‘intended changes’ that the administrators wished to bring in the testing regime in the EAP program. This chapter presents and discusses results of Phase 1 of the study, which took place from January 2016 until August 2016. The research question explored in this phase was: What evidence is there of washback in the former testing regime and what is the intended washback of the new testing regime? The results were obtained through analysis of the following data sources:

- Ten teacher interviews,
- Two administrator interviews,
- 137 students questionnaires responses and two focus groups
- Document analysis of the previous ExitTest

Phase 1 of this study looks at the situation on the ground before the introduction of innovation. I will describe the testing situation that existed in the EAP program before the introduction of the new testing regime and then present and discuss the results of the research question. To examine the antecedents⁶ situation, Henrichsen (1989) has suggested examining the following descriptive features of an educational reform context (see also Figure 4, Chapter 3):

⁶ Interestingly, these antecedents roughly correspond to Fullan’s (2015) ‘Initiation’ or Roger’s (2003) ‘Prior Conditions’.

1. *experiences of the previous reforms* (to understand if earlier attempts at innovation had been successful or not),
2. *traditional pedagogical practices* (teachers' classroom assessment practices),
3. *characteristics of the intended user system* (features such as the structure and power relationships in the context which might affect the success of an innovation), and
4. *characteristics of the intended users of innovation* (teachers and students).

In order to understand the data, the three-cycle coding approach (Saldaña, 2013) was carried out as discussed in Chapter 3. Specifically, there was an open coding approach for the first and second cycle codes leading to core categories with conceptual memoing. Then, applying a top-down procedural approach and using my core categories from the second cycle, I identified the themes, which are discussed in this chapter.

Experiences of Previous Reformers

To view how an innovation will be adopted, Henrichsen (1989) suggests looking at the “experiences of earlier reforms efforts in the same socio-cultural context” (p. 81). An understanding of how the previous reforms achieved their successes can provide valuable guidance in subsequent change campaigns. Therefore, it was useful to review the attempts of previous reforms in the EAP program and how stakeholders received these changes. I asked administrators about the assessment-related past reforms in the EAP program and their responses are explained below.

In the few years prior to this study, many positive assessment-related changes had taken place in the EAP program. These included changes in student placement procedures, and an introduction of assessment guidelines. The senior administrator

commented that when he started in 2012, he pushed for new standardized learning outcomes and new tests because teachers were using different outcomes/assessments and there was no “thoroughfare” between levels. In the first week of classes, teachers were giving their own placement tests to move new students from lower to higher levels or visa-versa. By September 2016, there was a standardized on-site placement test. The EAP office claimed to work closely with the university’s Testing Office in the development of this placement test.

Other reforms, mentioned by administrators included the standardization of final tests and scoring rubrics, especially for the lower levels of the program. A test development committee was established for producing all final tests. This committee had also created a test bank that could be used by new teachers in the program. Many of the in-house final tests developed by this committee were multiple-choice reading and listening tests in conjunction with prompts for the writing tests. What differed at the graduating level was that the final tests (e.g., the ExitTest) were not produced in-house; they came from the external Testing Office.

In addition to these reforms, formal student-teacher consultations were established after the midterm tests for standardized feedback. Every semester there were professional development sessions for pedagogical practices for teachers. Workshops, such as marking writing or constructing multiple-choice questions, were part of these professional development sessions. Both administrators that I interviewed mentioned that there was greater stress placed on teamwork and collaboration among teachers. These administrators felt that teachers were taking the lead in shifting learning processes and as

a result administration wanted to introduce further changes in the program, especially in terms of testing.

Christopher, the administrator, explained the importance and role of the external testing experts. He said that prior to 2010 teachers were the main gatekeepers and could decide if a student could pass the graduating level. However, in 2010, it was decided by the university policy makers that to make the EAP program more credible, it would be beneficial to administer an external high-stakes standardized proficiency test as an exit criterion for graduating students. The Language Testing Office of the university was to provide this test. Different faculties at the university welcomed this change. The Testing Office was a recognized institution in the university and faculty members had faith in the consistency and fairness of the Office's procedures. Having said that, it was also observed in teacher interviews, in Phase 1, that there was at least one problematic aspect to this change. Tensions arose between the teachers, who considered themselves to be classroom assessment experts, and the testers in the Testing Office, who were external testing experts. This issue will be discussed further later. In the former testing regime, the external testing experts' role seemed to be ascendant. Both administrators, however, claimed that there was collaboration between the testing office and teachers. As Christopher said, "the [testing] office is open to a 'portfolio approach' where teachers can *discuss and defend* the final students' outcomes", but this collaboration was not viewed very positively by teachers. As, one teacher, Raymond, said "as a teacher, you have no real input on the grading criteria.... assessment has been passed on to somebody else...and then you have to end up sort of lobbying on the part of the students if you think an injustice was made". Other teachers expressed similar views as Raymond.

Traditional Pedagogical Practices

As discussed in Chapter 3 (see Table 3), all teachers in the EAP program had either Masters or PhD degrees with significant teaching experience. Most of them also had a degree in Teaching English as a Second Language, so it can be assumed that they had fair knowledge of second language acquisition and assessment. Waters (2014) explains that BANA (British, Australasian and North American) teachers generally have an integrationist educational perspective leading to use of skill-based, discovery oriented and collaborative pedagogy. Similarly, Cheng and Fox (2017), adding on to the list of White's (1988) curriculum philosophies based on implicit beliefs, values and assumptions of teachers, discuss four types of assessments in ESL/EAP classrooms:

- *Classical humanism* where curriculum is traditional and assessments are generally related to memorization, recitation, copying and translation, and performance expectations are that of accuracy.
- *Progressivism* where curriculum focuses on learners' needs and teachers help learners identify language skill areas which need improvement and assessment is dependent on day-to-day classroom work of moving learners to negotiated purposes and goals.
- *Reconstructionism* where curriculum focuses on course learning outcomes which are predefined and these guide classroom activities, and assessment is related achievement or attainment of benchmark criteria.
- *Post-modernism or eclecticism* where curriculum is unique, emergent and developmental based on interactions between teachers and students, and assessments is generally in the form of self-, peer- and group assessments.

Table 8 presents educational philosophies and assessment practices of the participating teachers in Phase 1. It can be seen that most teachers in the EAP program had eclectic and student-centered teaching philosophies and believed in a communicative approach to teaching. Teachers suggested that they planned their lessons according to course learning outcomes and believed in using authentic materials in their classes. Their classroom assessment activities were mostly in the form of projects and presentations.

Table 8 *Educational philosophies and assessment practices of the participating teachers*

Teachers	Curricular and Assessment Philosophies	Excerpts from interview data
<i>Stacey</i>	Progressive	Respectful of students and here to help them get through the program; Consistent with program policy, authentic materials/assessment, but no group/ project mark
<i>Jill</i>	Reconstructionist	Student-centered, authentic materials, and communicative approach; Match course learning outcomes with assessment with peer and group work
<i>Joshua</i>	Progressive	Respectful of course learning outcomes, student-centered; Awareness raising quizzes and tests
<i>Ashley</i>	Post-Modern Eclectic	Pedagogical eclecticism, task-based approach; Group work, academic tasks like summarizing, paraphrasing and presentations
<i>Lisa</i>	Post-Modern Eclectic	Open-ended developmental approach, pedagogical structure as container for learning; Peer and Group assignments, active observable participation
<i>Maria</i>	Progressive	Authentic use of language, student-centered, students needs foremost; Formal & informal, peer-group work, test prep from proficiency tests
<i>Derek</i>	Post-Modern Eclectic	Eclectic, follows student needs and course learning outcomes, communicative approach; Authentic assessment based on skills e.g., note-taking for reading/listening
<i>Gerald</i>	Progressive	Student-centered, students needs foremost, need to get through the text book, communicative approach; Project-based, process writing than product approach
<i>Raymond</i>	Reconstructionist	Student-centered, prefers teacher autonomy; Student-centered in high-stakes, not much group work, high-stakes writing, high-stakes grammar quizzes
<i>Ruth</i>	Reconstructionist	Text book based lessons that cover course learning outcomes; Writing, quizzes, not much skill integration, pair work, or group work

Characteristics of the User and the User System

The characteristics of the intended user system and the characteristics of the intended users are important to examine in the antecedents as these characteristics will influence innovations and will be the basis for the new practices outlined in Phase 2 of this study. In this section, I have combined the context (user system), and administrators', teachers' and students' accounts (users) to answer the research question about the washback from the former testing regime.

Research Question 1: What Evidence is There of Washback in the Former Testing Regime and What is the Intended Washback of the New Testing Regime?

When interviewing administrators, my goal was to understand and get as much information as possible about the user system i.e., the context of the EAP program. On the other hand, my goal for teacher interviews, student questionnaires and focus groups was to understand and get as much information as possible about the characteristics of the users of the EAP program.

The washback literature has suggested that different teachers use different methods to teach towards tests. For example, Alderson and Hamp Lyons (1996), and Watanabe (1996) found that when teachers teach towards the same test or same skill, some adopt a more evident 'teaching to the test' approach and others adopt more innovative and independent approaches. These studies have also confirmed "the variable may not be so much the exam or exam skill as the teacher him/herself" (Spratt, 2005, p.16). Further, a validated proficiency test can be somewhat subverted by test prep practices or coaching that emphasizes "testwiseness" strategies over skill proficiency. This can result in higher test scores without an increased level of skill proficiency or

fluency (Messick, 1996, p. 246). This was evident in teachers' uses of classroom assessments depending on their views about what is important for students to focus on in this level. For instance, Stacey and Raymond, two teachers, who had been with the program for almost a decade, preferred not to use peer- or group-assessments with their students. This wasn't because they did not believe in classroom-based assessments, but because they did not see the value of these in relation to the final test at the GL. For example, Raymond said "I don't do pair work, I do a lot of high-stakes writing....I am not teaching with a theoretical approach, but "what do you need to do to be successful in this test"". Instead, these teachers kept their assessments in accordance with the final test.

Also, because of the high-stakes nature of the ET, most teachers' priorities included making sure students practiced writing essays and passed the GL.

... "what is my priority?" my priority is helping the students meet their goals, which is to bridge into the university. So if I was doing long impromptu speaking or focus on pronunciation it would probably improve their language ability, it would be helpful for them, however it wouldn't fit their immediate need which is to get into the universities (Raymond).

On the other hand, new teachers, such as Ashley, Derek, Maria, and Lisa valued alternative assessment practices, such as project work, self- and peer-assessment and incorporated these in their classrooms frequently. Lam (1993) has suggested that generally experienced teachers pay more attention to examinations than new teachers.

Other potential outcomes of positive washback of former classroom assessment practices were also evident in the teachers' interviews. These were: using academic learning outcomes to prepare students beyond the program, making learning directly

relevant and connecting to what the students' ultimate needs were, preparing students for the standardized proficiency tests, and integrating all skills in testing. For example, Derek, who has taught in the program for 12 years, said "it's not just language that we are dealing with, synthesizing content and regurgitating it for whatever purpose. I keep in mind the authentic academic objectives." Derek's quote brings to mind the idea of learning in situ where skills are used in direct learning in the environment in which it is going to be used. Similarly, Lisa, Joshua, Maria, Gerald and Ashley mentioned that even though not directly relevant to students' needs at the GL, they still used extensive readings, research projects, note-taking and annotating listening and readings, transcribing an audio clip as effective assessment tools in order to prepare students for university life.

Another positive washback of teachers' classroom assessments was that teachers felt confident that once students complete the graduating level, they were better prepared to take any proficiency test such as IELTS or TOEFL. Many teachers used these tests to practice in class, whether reading, listening or writing. For example, Raymond mentioned training his students every week for the writing exam so that students were "able to produce something of quality, quickly, in a single sitting." This section described the washback of EAP teachers' classroom assessment practices. In the next section, I will explain the washback of the ExitTest, first from teachers' and administrator's point of view and later from students' accounts regarding the former testing regime practices in the EAP program.

Washback from the ExitTest.

Teachers' and administrator' accounts. With respect to washback of the former tests, four main themes emerged as a result of teachers' and administrators' interviews: isolation of skills' importance; failure to test speaking and devaluation of academic skills; variability of testing processes; and evolution of testing criteria within the EAP program, generally, and most notably at the graduating level.

Isolation of skills' importance. The final writing task on the ExitTest was an individual, decontextualized writing prompt. Students wrote approximately 300-350 words on the ExitTest and often this piece of writing was what determined the fate of a student as Christopher, the administrator, pointed out:

The tricky thing is that if they fail the final writing portion of the final exam, no matter what, they have to repeat the level. Meaning that students who do well in listening and throughout the course, if they fail the final writing they have to repeat the level.

From this quote, it could be safely assumed that the writing test at the graduating level was high-stakes and exerted a lot of pressure on students and teachers. Writing was considered the core skill in which students must succeed in order to graduate. This became problematic because "passing the writing test" became the be-all and end-all for all graduating level students. Most teachers pointed towards the programmatic tension of assessing decontextualized writing vs. preparing students for academic studies at the university (e.g. see Raymond's comments below). Although, Hughes (2003) has asserted that constructed responses, like an essay, produce more positive washback than multiple-

choice items, such performance assessments can also produce negative washback. For example, Linn, Barker and Dunbar (1991), point out,

...more direct forms of assessment will not automatically produce classroom activities conducive to learning and might “encourage teachers to develop a formulaic approach to teaching generating high-scoring essays within the time limit imposed by the test (cited in Green, 2007, p. 11).

Raymond’s views were very similar when he said:

No, we are assessing right here: “Can I write an essay one page long, approximately 350 words on a topic in 60 minutes?” That is, like it or not, the way we assess, we do other skills but that trumps everything that is a shame. It’s a shame.

Other teachers like Derek also had issues with writing 5-paragraph essays as he said these were not the type of essays that students would write in their “first year and second year university courses” and suggested that tests should be more authentic to match the first year of university. He added:

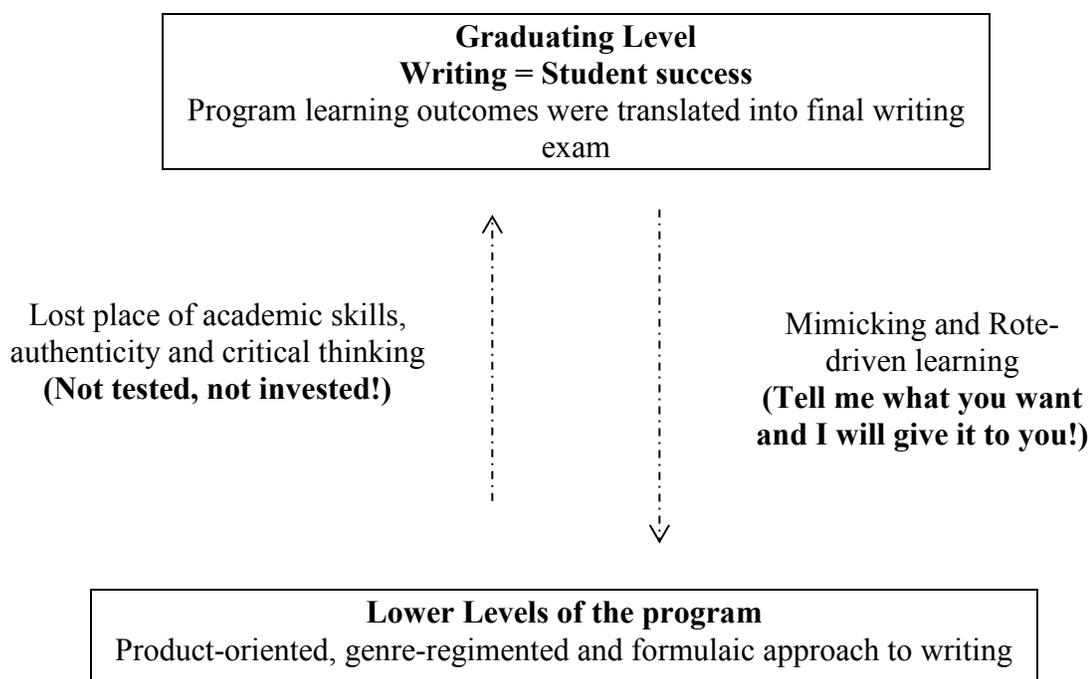
To synthesize information from different sources because that’s the most important academic skill – synthesizing information and making it your own because you have a purpose, you have something to prove. That’s something that our students are definitely grappling with. They are taught to construct this canned 5-paragraph essay, half of it is memorized, and the other half is water.

Students at university are expected to read, research, reflect, and write creatively about various topics, but it seemed that at the graduating level of the EAP program (and at other levels of the program as well), a prescriptive formulaic approach to writing was

advocated, which restricted the teaching of other important skills. Powers (2010) suggests that it is wrong to assume that if students are proficient in writing, they will be proficient in other skills too. The potential consequences of choosing to test some aspects of language proficiency and not others can have serious consequences (Brindley, 2008; Gipps, 1994; Madaus, 1988). Such criticism of skill isolation in the past has motivated changes in high-stakes tests, such as TOEFL PBT or the Test of English for International Communication (TOEIC).

Furthermore, having so much emphasis on writing skill seemed to have washback at the lower levels too. Figure 7, which summarizes teachers and administrators' views, conceptualizes the consequences of the importance afforded to writing skills in the program, namely: if students could write (in isolation of other skills and in isolation of the environment they need to write for), they would be successful.

Figure 7 Consequences of the importance afforded to "writing skills" in the EAP Program



For example, Anderson's, the administrator's, observation was "writing is heavy more so at our advanced courses." He pointed out that although at lower levels of the program teachers focused on grammar, it was still the case that "writing has a heavy emphasis" and that most teachers lamented that the final tests at the lower levels had been developed to reflect the design of the final ExitTest. It seemed that this two-way process (Figure 7) of isolating "writing" skill exerted negative washback, not only at the graduating level, but also at the lower levels of the program. Raymond pointed out that a "product approach and not the process approach" is encouraged in the program because "how can you do the rough draft and the final draft in the 50-minute test."

In the product-oriented teaching approaches, Ferris and Hedgcock (2013) suggest that students' essays are considered a "static representation of students' learning and content knowledge" (p. 64) and students are expected "to produce and master a range of school-based models of rhetorical arrangements, such as description, narration, exposition, comparison, and contrast, process analysis, argumentation and the like" (p. 63). They go on to state that as a result, "little instructional time [is] entailed [in] the planning, drafting, sharing, revising or editing of students' texts" (p. 63). They further suggest that in product-oriented teaching approaches essays are expected to adhere to "rigid rhetorical conventions" (p. 64).

These conventions typically include what Williams (2003) calls examples of rules of "good writing". These included:

- "All paragraphs must have a topic sentence."
- "All essays have an introductory paragraph, three body paragraphs, and a concluding paragraph."

- “All concluding paragraphs reiterated the information in the introduction” (Williams, 2003, p. 100-101).

On the other hand, in a process approach, the emphasis is on an individual writer as “a creator of original ideas and the need to cultivate his or her innately generative predispositions” (Ferris & Hedgcock, 2013, p. 64). In this approach, pedagogies do not focus on “isolated parts of texts or on grammatical features, rather the emphasis of process pedagogies is on identifying and solving problems, discovering new ideas, expressing them in writing and revising emergent texts” (p. 65). Williams (2003) suggests habits of “good” writers are conceptualized in “stages” of the composing process, and translate into “prewriting, planning, drafting, pausing, reading, revising, editing and finally publishing” (p. 101).

Another issue with skill isolation was the rote-memory approach to testing. Although the EAP program had a range of learning outcomes, students had one goal— to pass the test and enter their university program. Many students saw the EAP program, itself, as a hurdle to be gotten through rather than a meaningful and important learning opportunity. Because joining the EAP program was incidental for them, it wasn’t really part of their agenda for studying in Canada. Students were in the program because of their low levels of language proficiency and studying in the EAP program further delayed their actual university degree completion. If intentionally or unintentionally the message was that other skills were of lesser importance than writing, then learning became rote-memory driven. “Tell me what you want and I’ll give it to you” seemed to drive much of the program and classroom interactions. Read and Hayes (2003) suggest that test-taking strategies and using practice tasks are frequent in preparation for high-stakes tests, but

what can get dropped in such a scenario are important (ungraded) skills needed for university success. Raymond acknowledges this in the following comment:

I think for this graduating level, writing is the *be all and end all* of how they will be evaluated. We do reading and listening in class but in reality, and they know this, it all comes down to the writing. And since you're limited in the amount of time you have with them I spend almost entirely my class in terms of writing and I start with sentence level and I make sure they can write grammatical sentences and then we move on from that. It's a writing focused teaching that I do.

The condensed version of an essay and condensed time to write somehow was believed to be a proxy for real life essay writing in the academy. Such academic essays in undergraduate disciplines typically span 10-15 pages of writing and involve extensive research and critical thinking. A hidden lesson seemed to be that fast is right, rather than a slower, more methodical and reasoned academic approaches. Ruth, who didn't teach graduating level often, pointed out the differences between successful students' academic practices and those that are tested:

It's tricky because of the nature of the tests. We have a lot of really smart students who are older *per se*, and they just cannot complete the types of tests we give them; the multiple choice in the given time frame. But I don't know if that really means that they are incapable of surviving a graduate university course.

Davies (2007) argues that although high quality, high-stakes proficiency tests should be based on sound theoretical constructs and are curriculum-free, they nevertheless can appeal to a syllabus and evolve into a syllabus and, hence, become achievement tests. In such scenarios, it becomes difficult to prevent teaching to the test.

This illustrates how washback influences teaching and learning, where the focus becomes high test scores rather than the attainment of domain skill fluency (Messick, 1996). As Green (2007) suggests, “where individual participants come to value success on a test above construct knowledge and understanding, negative effects appear more likely” (p. 17).

Maria, a testing expert, who was teaching in the EAP program at the time of the study, pointed out that although the writing tests at the lower levels were “genre-driven” and “stressed the organizational patterns of writing” (e.g., *compare and contrast essays* or *cause and effect essays*), the final writing for the ExitTest did not test any specific style or organizational pattern. She explained:

I think that our tests, our midterm and finals at lower levels test these organizational patterns more so than having students think critically about a topic and just choose for themselves how to organize a response..... So this is something that bothers me, in our program, that they see these as a type of essay rather than as a strategy for organizing ideas. That’s the difference I see at lower levels. They’re taught “Okay, we’re going to do cause effect, we’re going to do this” and in the midterm/final they’ve trained themselves or they’ve got this in their head and by the time they get to the ExitTest, and sometimes they kind of panic and they do poorly because they haven’t seen these prompts that might combine different things.

Finally, it seems unlikely that a single measure (like the writing on the final ExitTest) could serve as a sufficient proxy for students’ overall proficiency in a) in all modes of communication in English, and b) the TLU, i.e., the academic domain. As

mentioned earlier, there is general agreement (Powers, 2010) that the measurement of one skill cannot substitute for the measurement of others. Similarly, artificially splitting and assessing the skills else could not have provided a good measure of students' abilities to use English for academic purposes in university. What the impact of this practice was at the EAP program will be discussed in the next section.

Failure to test speaking and devaluation of academic skills. According to Green (2007), “washback is grounded in the relationship between preparation for success on a test and preparation for success beyond the test, in the domain to which the test is intended to generalize and to which it may control access” (p. 1), but it is also claimed that “it is testing, not the “official” stated curriculum, that is increasingly determining what is taught, how it is taught, what is learned, and how it is learned” (Madaus, 1988, p. 83). As mentioned in the previous section, true to this assumption, teaching writing was treated as most important in the EAP program. Further, the interview data also suggested that it was done at the cost of teaching other skills, which had repercussions on teaching and learning. For example, Stacey, who taught graduating level every semester, mentioned that she didn't like the idea of teaching to the test, but had to frequently do practice tests in class, which were very different from what was in the textbooks. She liked to “encourage a broader scope in instructing and bring *realia* in to the classroom and teach from that.” She added:

I think that writing summary response and those exercises are very helpful.

Learning how to take notes and distil the important points from a professor. These are very important. But I find that in the graduating level we are teaching to the test and I find that constraining a bit.

While four teachers, Stacey, Derek, Jill, and Joshua had issues with the test content and format, Maria had issues with the test administration. She said she “completely disagree[s]” with playing the listening tasks twice as this had implications for the program practices in general:

I don't think it's something that's authentic....and then that kind of trickles down throughout the program, that they listen to everything twice, but when they get to university, do they really get to listen twice? Not really. And it's different than any other proficiency test. To me, I think that's a little unusual.

All teachers and administrators alike confirmed that speaking was not formally taught or tested in the program and it had been pushed to the afternoons for the teaching assistants (TAs) to engage in. Generally, for students, the relative power position of who teaches a skill gives the impression of the relative value of that skill. If the TAs, alone, assessed speaking and presentation, the suggestion was that these skills were relatively unimportant. Most teachers were aware of this shortcoming in the program and voiced their concern about it as Lisa said:

They [students] might manage in the first year, because classrooms are large, but then the higher up that they go, in the long term, they will need to make use of their speaking skills, otherwise they won't get to speak at all. They will be too shy or intimidated to speak in class. They have to give presentations depending on which program they are in. Speaking is definitely something that they need to do.

This is consistent with Fulcher's (2010) observation that teachers “have to respond to the demands made by testing regimes and students' desire to pass” (p. 277).

In disciplinary classrooms at the university level, the TLU domain of the EAP program, students are not generally assessed on their ability to write spontaneously in response to a given prompt. They have to use language in many other ways, such as listening to lectures, note-taking, conversations/discussions/debate, group work, negotiating problems, thinking critically and creatively, and writing cogently and proficiently, to name a few. Academic communicative competence is a complex construct, which has many facets and to focus exclusively on some facets to the exclusion of others might underrepresent the construct (Messick, 1996) and this may make a test less valid for its intended purpose (Powers, 2010). Further, researchers, such as Nagy and Townsend (2012), question whether the “best assessments for academic language interventions are measures of disciplinary knowledge or measures of those components of academic language that can be isolated for testing purposes” (p.104). Weigle and Malone (2016) explain that there is tension in the second language testing literature between assessing “discrete bits of language and assessing language in a specific context” (p. 669).

Teaching related to academic skills in the EAP program was narrowly defined and generic. It failed to take into account academic practices of critical importance in undergraduate university study. For example, Weigle and Malone (2016) point out the need to develop students’ understanding of academic skills such as citations and avoiding plagiarism.

In Phase 1 of this study, it was the harder skills (such as writing essays), which were narrowly tested in this EAP program. Softer skills like negotiation (through conversation/speech) and teamwork (through conversation/speech) were generally

ignored. The program seemed not to focus or value other essential academic skills. This could be seen in what was graded and tested (and therefore valued) and what was not. Teachers were acutely aware of this gap and often struggled to bridge this gap. As Derek mentioned:

But again, with every lesson, I ask “Is there enough input for the output?”.. “Do students have enough content for the production activity, like vocabulary, conceptual bases?” Especially, in the EAP, I think it’s very important; it’s not just language we are dealing with; we are dealing with synthesizing content and regurgitating it for whatever purposes. I keep in mind the authentic academic objectives. Something students are able to take home with them or their program of study.

He mentioned that once he tried to concentrate more on an extended research paper and on citations than teaching about the test, but hit a roadblock because the students quickly realized “there’s completely different stuff” on the test. Similarly, other teachers commented that missing from the testing was authenticity and critical thinking within the programming, meaning that EAP students were acted upon and not engaged with by the program. Students were added to a ready-made program and they were passive recipients of teaching. Was the program doing a disservice by not creating tests as a learning tool in a continual learning process rather than a completely separate exercise outside of classroom learning? Much of this appeared to be coming from the role of tests especially at the GL, to which we will turn our attention to in the next section.

Variability of testing processes. At the graduating level, there seemed to be a conscious effort to divorce formal testing from classroom teaching. As Derek mentioned

in the previous section, when teachers engaged in teaching universal academic skills, there were various levels of disconnections and students sometimes noticed these. For example, rather than focusing on lessons that were not going to be on the test, the students generally focused on what would be graded. By doing this sorting, many important and necessary skills needed to succeed at university were devalued significantly, disadvantaging the students.

As explained previously, the ExitTest came from an external testing office. However, this test had become so central to the teaching, that it seemed teachers became absent from the equation. Outside testing experts designed the test, without knowing what was going on in the classroom, who the students and teachers were, etc. While this “objectivity” might be seen as a good thing, it did create disconnection of which teachers were very aware. The comment below by Raymond highlights this issue,

One has to take a look at in the end, the final and midterm tests are graded by the Testing Office here. As a teacher, you have no real input on the grading criteria. It is one reason that I don't enjoy teaching at this level. In the past I've always liked to have control over how my students are assessed and after all, I'm the one who has taught them over the past 14 weeks. I think I'm in the best position to know their progress or not or even their proficiency level [I know better] than someone who's looking at one hour of writing... This assessment of this GL has been passed on to somebody else who is not the teacher and then you end up having to lobby on the part of the students if you think an injustice was made. That's insane,.. to have a situation like that where you have to go...

Thus, keeping tests out of teachers' control conveyed the message that teachers couldn't be trusted to test students; testing was important work that only managers or outside experts could do. This view seemed to be related to the diminishing of teacher autonomy by having testing used as an instrument of teacher oversight.

Furthermore, the program became more about getting students past the 70% pass/fail cutoff on the test than on preparing them to undertake study in English as undergraduates in their degree programs. When students did not achieve this benchmark, then it was time for negotiation. The test was essentially the central aspect of the program in the former testing regime as Derek stressed:

I find there's a bit of a disconnect. I understand the difference between objective reference, criteria reference in. So obviously the midterm tests we get from the office is all criteria reference - reference to a certain set of benchmarks, descriptors, it's been tested, the item analysis has been done already, piloted, there's data behind it on performance. I understand all that. So, what seems to be the disconnect is how teachers teach toward something else [academic skills like citations, summaries, paraphrasing etc.] and the students then realize "oops there's completely different stuff on the midterm".

Another point raised by teachers was that they were not sure what some of the parts of the ExitTest actually examined. For example, Ruth thought the cloze test was to test grammar. Maria said that most language proficiency tests did not have cloze tests. She was not sure about the value of it and said that cloze tests measure the understanding of collocations and grammar, but "for some reason, our students don't find it easy". In language testing literature, an integrative test, such as a cloze test is "aimed to assess a

unitary language-proficiency construct that was largely grammatically driven” (Plakans, 2013), but its actual constructs have always been debatable (Fox, 2001). Measuring language this way conceptualized it as four skills, plus grammar and vocabulary. Fox (2001) also claims that the cloze tests elicit much less academic expertise as compared to any English for academic purposed writing tests.

Further, teaching to the test was viewed as very limiting, not only by the teachers, but the administrators as well. While Gerald said in the second half of the term he covered skim and scan techniques with his students, and practiced comprehension and cloze exercises, Maria said she hated teaching to the test, but had to do it and “it ends up taking up instructional time from a lot of other skills”. She further added that because of the high-stakes nature of the final tests, after the midterm, it’s “a kind of wake-up call” and test preparation activities are intensified. However, she reiterated “at the lower levels, it’s less high stakes, so not as much time is spent on test prep”. The washback literature has mixed views regarding the effects of testing on curriculum. While Lam’s (1993) findings were that testing-based classes get more time in curriculum, Shohamy et al’s (1996) study suggests that exams which are high-stakes get more time than low-stakes exams. Similarly, Green (2007) suggests that when the test is challenging for the participants and results are important, the washback associated with that test will be more intense. Anderson, the administrator, also commented on the washback of the ExitTest as:

The receptive skills I’m seeing teachers give practice tests or, practise cloze. You want to ensure that the task types on a test are replicated in the classroom...so teachers have no choice. I think the same is true for writing. They’re given prompts and they practice and they write them, so there’s quite a bit of washback.

I'm not saying that the washback doesn't promote language proficiency, it does. Students certainly benefit from it, but it's our current state and I'm hoping, well not hoping, we will have more to a different assessment practice very soon.

Teachers too, voiced their opinion about the potential of negative washback of preparing students for multiple-choice tests at the graduating level. Gerald was of the opinion that in the case of some students, it was "multiple guess" rather than "multiple choice."

Similarly, Ashley, Jill, Maria and Stacey all explicitly stated that teaching to the test was against their teaching philosophies, but still indulged in test prepping their students as Ashley said she made her students practice many tests because multiple-choice tests require students to learn test-taking strategies and "the more you practice, the more flexible your muscles are, just like exercising". Similarly Lisa lamented about giving in to students' demands about test preparation, but wasn't really sure about the benefit of these exercises for students' university careers.

I can say that at the beginning I did very little [test preparation] especially at lower levels. Then the more I taught at the GL, the more I found myself getting as stressed out about the tests as the students.... so yes, I ended up doing more of that, but still keeping in mind that I still want them to also learn skills that they can make use of in their university careers, transferrable skills.

It appeared that in Phase 1, formal tests were consciously divorced from classroom teaching. Ruth mentioned that at the lower levels, there was some flexibility with the final tests and teachers did get the support of the testing office, but at the GL "the final tests were written in stone".

Furthermore, there were other tensions surrounding the role of the testing office. Firstly, the vision of educators (administrators and teachers alike) was that the testing office was “expert” and so should be relied upon, presumably to the exclusion of teacher expertise. This raised a programmatic tension of “outside expert” versus teacher autonomy to know what was best for teachers’ own particular students. All aspects of the EAP program were tied to this one narrow view of tests.

Maria, who had worked as a test developer for different proficiency tests in Canada, also questioned the results provided by the Testing Office and the validity and reliability of the testing procedures. She was concerned that many students with scores as low as an equivalency to band 5 (modest user) in the IELTS were being allowed to bridge into the university. Her issues were also with the test content and she said, “these tests are missing a lot of learning outcomes” and despite making several suggestions, the testing office remained “hesitant to make any changes to the test”.

The problem is that we might get these reading or listening tests and they only test main ideas and details. There’s no testing of inferences . . . even though it’s a test that’s meant to test the outcomes in our program, it doesn’t test or assess all of the outcomes.

Finally, for teachers like Stacey, the issues were with more practice “testing materials” and she was suspicious, as she said, “I’m not quite sure what adjective to use here. But they’re not . . . I don’t find that they’re very forthcoming with a lot of their materials”.

There was a disconnection between how tests were structured and what went on in classrooms. This was a problem in that tests were considered separate from, instead of

part of, the overall learning process. Testing seemed to be something that was done, in a black box, as a complete, unrelated exercise.

Evolution of testing criteria at the graduating level. All teachers unanimously agreed that a change in testing criteria was required at the GL. Although the rationales given were slightly different, it was agreed that there wasn't much alignment or correspondence between learning outcomes, tests and learning and students' role within that. For example, Stacey wanted changes in textbooks, as there were at least three prescribed books at the GL in any semester. Her concern was that the program had well-qualified teachers who could come up with a complete curriculum:

we have experienced teachers who have created their own material or who have borrowed from other textbooks, you can come up with a complete curriculum “sans text, c'est pas nécessaire.

Stacey's views were in parallel to Wall and Horak's (2006) assertion that in the field of EFL/ESL, a movement called “*Dogme*” is prevalent which:

criticizes ‘material-driven pedagogy’, claiming that the overuse of commercial materials and technology undermines teachers’ ability to truly address students’ needs and promote a one-size-fits all approach to what happens in the classroom (p. 82)

Table 9 describes teachers' views about a need for change in the former testing regime, their rationale for change, and their knowledge about the impending changes as proposed by the administration.

Table 9 *Teachers' views on a need to change former testing regime*

Teachers	Is there a need for change?	Rationale for change	Awareness of impending change/What it will be?
Stacey	Yes	Too much disconnect between what we teach and what we test	Not much, only one comment by an administrator, may be less reliance on text-books, more interaction with the testing office
Jill	Yes	Can't have one size fit all – need to cater for different streams (ESP)	Vague, but moving towards Academic English
Derek	Yes	Exam is proficiency based, but we prepare students for academic readiness	Creating streams (ESP)
Lisa	Yes	Need to test speaking	Project-based learning, take home tests, & portfolios
Raymond	Yes	Limited test-writing environment	No awareness of change
Joshua	Yes	Program is NOT academic, but ESL with academic name	No awareness of change
Ashley	Yes	Multiple-choice tests should be replaced by academic assignments	If starting from top level, not fair for penultimate level
Maria	Yes	Need to prepare students for university	Dense learning outcomes and testing speaking
Gerald	Yes	Testing skills (e.g., skim, scan and infer) but not critical thinking	It is in a state of flux, but more project-based learning
Ruth	Yes	Tests should match learning outcomes	Project-based learning, assignments, take-home tests

On the other hand, Raymond wanted change at a number of levels- from changes in tests to changes in teacher autonomy:

I think that assessment at this level should be changed on a number of levels. I think one thing is that we have too much emphasis on a limited test-writing environment, and some people do well and others don't, and it doesn't reflect their actual language proficiency in a situation like that so I think we should

broaden to evaluate on a more project-based approach.... certain skills are overlooked; especially listening, note-taking, and the kinds of skills that the students require at the university. I think that we could broaden that. I think that having an outside testing office who are going to be, by their very nature, looking for the fastest, sort of easiest to mark grade. Having them make an assessment on a student that they don't know or haven't seen progress, haven't seen development, I think that needs to change.

Similarly, Gerald said he wanted the program to be focused on university preparation for students and not one that feels more like just throw them into university programming like "I've washed my hands, now it's your turn."

Overall, the proposed change to test integrated skills was seen as a very positive and necessary step, but it was also the case that in Phase 1, not all teachers were integrating skills in their classroom. Therefore, changing to skill integration could be seen as an imposed change by such teachers, in turn affecting teachers' and personnel's expectations (House, 1989; Markee, 1997).

As discussed in the above sections, the former testing regime was exerting potentially negative washback on teachers, however, the variability of washback (i.e. the higher the test-stakes, the more the washback, and vice-versa) on students didn't seem to be so strong. In the section that follows below, I discuss students' views regarding the former testing regime and the ExitTest.

Students' views about the former testing regime and classroom assessment. Most students reported positive effects of the pedagogical practices and classroom assessment of teachers. These were clearly evident in students' views about the former testing

regime. According to Bachman and Palmer (1996), three aspects of the testing procedures affect test-takers:

- the experience of taking, and in some cases, preparing for the test,
- the feedback they receive about their performance on the test, and
- the decision that may be made about them on the basis of their test scores (p. 31).

While, conceptually, these three aspects were kept in mind when analyzing students' accounts about assessment practices in the EAP program, methodologically, I first analyzed the data obtained from the quantitative questionnaire using descriptive statistics and then triangulated the results with a three-cycle coding approach (Saldaña, 2013) of the focus group data (see Chapter 3 for details). Two main themes emerged as a result of these analyses:

- 1) positive washback of the former testing regime
- 2) negative washback of the former testing regime

Positive washback of the former testing regime. Watanabe (2004) suggests that it is inconceivable that “test writers intend to cause negative washback” (p. 21) and the issue of value judgment can only be made by referring to the audience -- “who the evaluation is for” (p.21). For instance, teachers may judge an outcome as negative, but students can evaluate the same outcome as positive. Students in the program did see connections between course outlines, learning outcomes and the use of these learning outcomes in learning. Spearman's rho correlation coefficient was computed to assess the relationship between the explanations of the course outlines, learning outcomes and their use in classroom teaching and learning. There was a significant positive correlation between each of the three variables ($\rho = .519, .384, \text{ and } .219, n = 137, p < .01, < .05,$

and .05 respectively). Increase in the use of learning outcomes in classroom teaching was correlated with better learning as perceived by students. Students in the focus groups corroborated that most of their teachers conducted needs analysis and taught according to students' needs, as Sheikha, a potential graduate student, mentioned:

They [teachers] decide if we need to practice more listening, but we are good at reading, So we need to focus more on our listening at this for two weeks and then we return to the regular classes. I think the main thing is that they evaluate our performance and determine which kind of exercises or practices we need to reach the Can Do [course learning outcomes].

Another positive washback of the ExitTest was that students felt confident about preparing for other standardized high-stakes proficiency tests such as IELTS and TOEFL. Practicing multiple-choice items for reading and listening and the product approach to teaching writing skills in the program are useful for students in that they get enough practice to write essays under time constraints. This can be useful when students need to take other high-stakes tests. As Seiko, a student from Japan, commented:

I studied before writing compare essays and cause and effect, but with this term it shown me clearly what is the structure for each kind of writing, so that helped me with the class assessment and other tests like IELTS and TOEFL. When I read the exam I can decide how I will organize my writing and the format and structure.

So it saves a lot of time.

In addition, students mentioned about strong student support provided especially to the weaker students in the program. The Academic Services Coordinator in the program was very helpful and she kept a tab on the weaker students especially at the GL. Free tutoring

was provided to borderline students and this personalized care was one of the hallmarks of the program.

Finally, 59% (81/137) (see Table 10) of students agreed that they were mainly assessed by written essays and most of the time, 71% (100/137) students responded essays were given marks. Rubrics for marking these essays were explained in class, so students were aware of the mistakes they were making in their writing. 60% (82/137) students mentioned that feedback was frequently provided on their written work.

Table 10 *Students' views about assessment practices in the former testing regime*

Overview of student Responses (n= 137)	Cronbach's Alpha =.92			
Perceptions about Assessment	Frequently	Sometimes	Never	Don't Know
Assessed by essay	83 59%	51 36%	2 1%	5 4%
Assessed by multiple-choice questions	59 42%	71 50%	6 4%	5 4%
Assessed by short answer questions	43 30%	90 64%	6 4%	2 1%
Assessed on note-taking skills in listening/reading	32 23%	87 62%	17 12%	5 4%
Essay is given a mark	100 71%	37 26%	1 1%	3 2%
Rubric for marking essay is explained	53 38%	74 52%	3 2%	11 8%
Essay is marked for language accuracy	79 56%	51 36%	2 1%	9 6%
Essay is marked for overall communication	65 46%	57 40%	9 6%	10 7%
Self-assessment is used	31 22%	82 58%	18 13%	10 7%
Peer-assessment is used	28 20%	91 65%	8 6%	14 10%
One-to-one tutorials are provided	40 28%	68 48%	12 9%	21 15%
Mid-term and end-of-term feedback is useful	46 33%	69 49%	12 9%	14 10%

Bachman and Palmer (1996) suggest that when test-takers receive feedback on their work, it contributes toward positive washback of that assessment activity. Focus group students expressed the opinion that if they made progress in writing, then they felt that they actually made progress in this program. This echoed the views expressed by teachers (see Raymond's and Lisa's comments above) that writing was the most

dominant skill in the program and the negative washback of this aspect is explained in the next section.

Negative washback of the former testing regime. Students echoed teachers' thoughts about the negative impact of the ExitTest. First of all, almost 61% (84/137) of students didn't have an understanding of the final tests. Bachman and Palmer (1996) suggest, "the test taker's perception of test task informs or misinforms the test taker" (p. 32). If the TLU domain (here academic English) is unfamiliar to students, it may affect the test taker's perception of the TLU. For example, Jessie, a student who had failed the graduating level once and was clearly frustrated with the exit requirement for the program said:

At graduating level, always the Exit Test..., when we finish, we can't see our answers after the test and we don't know where we have a mistake, so next time when we see the same question, we still don't understand it.

Knowing that tests are powerful and have impact on students' future prospects, it is not surprising that the ExitTest had a deep influence on students' feelings. For better test preparation and positive washback of tests, Hughes (1989) suggests that "the test provider should inform participants about test content, publicize the theoretical basis for the test, and train teachers in effective forms of preparation," (as cited in Green, 2013, p.46).

Although seen as a positive outcome by some students, too much emphasis on writing had effects on learning other academic skills for students. Preparation for the writing exam was the foremost concern of the teachers and students as 71% (100/137) of students said that test preparation was frequently assessed by essays (see Table 10). Students were aware of the importance afforded to the final writing as Tony commented

Midterm [writing] is important, but if you don't pass the final writing, you don't pass. I don't see much difference in midterm and final, but for the final...it's like pressure.

Students also voiced their opinion on other academic skills, such as note taking. For example, only 23% (32/137) of students felt confident in stating that they were frequently assessed on note-taking skills in listening and reading as compared to 59% (83/137) of students who said that they were frequently assessed by essays. Sheikha, who was going into a graduate program, was aware of not preparing enough for her upcoming university studies. She said her friends from her country who were “just learning English” in other language schools were “preparing for academic research” and doing “abstract writing and presentations,” but at the EAP she was learning about 5 paragraph essays and “not speaking enough or learning to give presentations”.

When asked to rank order certain exam-related activities, Seiko and Tony said that the most important activity is *spending time in class practicing sample tests*. Seiko explained,

...I noticed that there are a lot of kinds of English tests. For example, we practiced TOEFL listening however we went into the midterm test and the format is quite different so we don't get used to it and I didn't get a very good mark.

Or, as Tony remarked,

...because when we understand what we are asked to do we can decide an approach to study, how we are to study, or how can we improve our skills that we need in the final test.

Students were more concerned about learning strategies for test-wiseness. Cheng and Fox (2017) define test-wiseness as “the ability to respond advantageously to items or tests formats that contain clues, therefore, to obtain credit without the skill, proficiency, ability or knowledge of the subject matter being tested” (p.230). Tony summed up this concern by commenting “your English abilities are good but if you are not familiar with the format, you don’t get a good mark.”

When asked for the least important activity in terms of test preparation, students mentioned group discussions. Tony explained that although group discussion helps “improve your oral English, it is not important for the final tests”. On the other hand, Sheikha mentioned reading and writing short and long answers were the least important as sometimes, “there is no specific answer for the question (like math and science), and it is your knowledge not answers that matter”; the final reading tests are multiple-choice. For Seiko, the least important skill in terms of test preparation was spending time in practicing summarizing, paraphrasing, and referencing. Although students did practice these skills for some time in class, her teacher said, “it is more time consuming, it is less possible to appear in the test.” In this instance teacher unintentionally highlighted what was important to students and what was not important from the final test’s point of view.

Finally, one of the negative effects of the former testing regime was the secondary place assigned to classroom-based assessments such as peer-, self- and group-assessments. Students had mixed views about self-assessment. Only 58% (82/137) of the students felt that self- assessment was sometimes useful and approximately 20% (28/137) of the students were uncomfortable with self-assessment. Mohammed expressed this frustration as:

Some sentences I write are correct [grammatically], but the teacher says we don't write like this in English and on some topics, when you give opinion, your ideas... lack of background knowledge, so we may sometime give silly idea, that totally doesn't change with self-assessment. Grammar errors, we can find and correct, but this kind of error is hard to correct.

Similarly, students had mixed views about peer-assessment. 65% (91/137) of students felt that peer-assessment is sometimes used in the classroom. For example, Mohammed seemed contented with peer-assessment and said it was "easier for me to get new words or ideas from other students", rather than teachers, as "teachers are native speakers and they don't use our level words [vocabulary]." 16% (22/137) of students were unfamiliar with the concept of peer-assessment in the classroom context.

Intended Washback of the ISTs

Three factors of Henrichsen's (1989) process phase were the best fit with Phase 1 of my study. I explored them before the change implementation in the Fall of 2016. Henrichsen's *source* (innovators) was considered the EAP program administrators in my study. His *message* (innovation) was the new testing regime in the EAP program, and his *plans and strategies* was how administrators in the EAP program intended to integrate the proposed changes. The main data sources used to answer this part of the research question were the administrators' interviews and official internal communication documents of the EAP program. In this section, I will first explain the rationale and the objectives of the intended change as explained by the administrators, and then the explanation of how they intended to integrate these changes into the curriculum. Finally, I

will discuss the potential factors that may facilitate or hinder the process of implementation of the ISTs as explained by the administrators.

Rationale and objectives. Like teachers in the program, the administrators that I interviewed were also aware of the shortcomings of the former testing regime. Anderson said that in the ExitTest tests of receptive skills, listening and reading were multiple-choice and the writing test was a timed impromptu essay; both were traditional methods of testing. He further added that these tests “just didn’t feel right.” The future shift was intended to address this lack of alignment and integrate tests and course work. This move would better align with the teaching practices within the program. Anderson added

I think that for students, when you give them a standardized test that has no relationship to what they’re doing in class, or they don’t see the content validity, you have a problem with face validity where they don’t...even though many of them come from backgrounds that use traditional testing approaches and so we can do that. I want them to come into our program and think, “Wow, this is really novel and this is really interesting and these assessments they really help me with my English.” That’s what I want. “These assessments really help me.”

Similar to teachers, both administrators agreed that there was a need for change in the program’s testing regime. In his theoretical model of innovation, Markee (1997) discusses who adopts what, where, when, why and how. Under ‘what’ he talks about innovation and suggests five kinds of change that can occur in an organization. His notion of *immanent change* or self-motivated change resonates with these participants’ (administrators and teachers) accounts of the former testing regime and their views on changes that were needed to improve the EAP program. Immanent change “occurs when

the persons who recognize a need for change and those who propose solutions to a perceived problem are all part of the same social system” (p. 48).

The two administrators, however, gave slightly different rationales for change. While Anderson stressed that tests should promote learning, have validity and replicate skills and competences required for undergraduate course work such as synthesizing, summarizing, and responding to authentic tasks etc., Christopher maintained that the previous course outcomes were purely language based and there was a need to go beyond linguistic needs of the students, for instance, by developing intercultural awareness, using technological tools, analyzing academic content, and using communicative skills effectively. However, both agreed that the overall aim of the intended changes to the testing regime had to provide students with “more authentic academic experience, improved validity of our tests, and better prepared students for academic life after the EAP” as suggested by Anderson. So the purpose of the proposed changes was to align tests to undergraduate tasks in a valid and guided learning way. Anderson stressed that “if our tests have met that purpose, then we know that our teachers will be doing the same assessments in class and will be working on those same skills in class.” This comment clearly reflects the intended washback of the new Integrated Skills Tests and echoes Messick’s (1996) definition of washback as “the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning” (p. 241).

Integrating ISTs into the curriculum. For the planning, Anderson said that even though he wasn’t directly working on the tests, per se, he was “the project owner” and was involved in bi-weekly meetings to develop the integrated writing and speaking tests

and rubrics for marking these tests. He commented that the “assessments will go beyond the overall comprehension, accuracy, range, and discourse structures to include synthesis” (p. 5). Both administrators stressed the integration between class preparation and final tests, as former tests had a lack of alignment among classroom activities and testing. It seemed that there would be a renewed focus on skill integration, not only of four skills (reading, listening, speaking and writing), but also skills required for academic success. This intention is in keeping with the advice of Green (2007), citing Weir (1993), who proposes that a test of academic English should be made “as realistic and direct as possible so as to reflect the performance conditions and operations that apply in the target language use domain” (p. 13). Christopher said there were new program learning outcomes and changes being made to the course learning outcomes (CLOs) with the introduction of the integrated tests. These CLOs would include “oral production, group work, critical thinking, term projects, study skills, learner autonomy, and academic readiness” (official communication, August 2016). The idea was that receptive skills will be tested indirectly and productive skills will be tested directly. The official communication further stated:

Research in L2 assessment of EAP students lends support of our view of the validity of integrated assessments. Pedagogically, we believe these assessments offer a more advantageous and authentic alternative to the receptive skills and five paragraph essay assessments used in the past (p.1).

The official communication also stressed that although teaching and testing integrated tasks would be a transition for teachers, the administration saw several benefits of these:

1. **Pedagogical Benefits:** To complete an integrated task, students will need to use a variety of skill in transitioning from receptive to productive skills and in working on content previously covered in class. To prepare for tests, students will need to review their course readings, providing them with a sense of control over the outcomes of the assessment and offering them greater exposure to high frequency patterns of academic discourse, syntax and lexis. We believe all of this will translate into greater language proficiencies gains.
2. **Assessment Validity:** In addition to requiring students to communicate clearly and accurately, to complete integrated tasks, students will need to summarize, paraphrase, synthesize information from various sources and cite information. As well, for tests, students will need to review their course readings, This testing of additional skills and the preparation component of the testing process will improve the validity of our assessments (p. 3, EAP official update, August, 2016).

The intended impact of this change would be an evolution in pedagogical processes to mimic undergraduate realities. An expected result would be more integrated teaching, as Christopher stressed:

Basically a shift from just teaching ESL in the more traditional sense toward English for academic purposes where students actually have a purpose behind their writing evolutions in learning processes also including presentations, projects, and group work etc.

To promote positive washback, Bachman and Palmer (1996) suggest, “involving test-takers in the design and development of the test, as well as collecting information from them about their perceptions of the test and test tasks” (p. 32). Similar to this

Christopher said that the ISTs were piloted with a section of graduating level in the Summer 2016 before integrating these tests in the Fall 2016. He said the reason for the pilot was to make sure that “we are doing it right” and to see “what [the] potential risks and challenges were”. These potential risks and challenges will be discussed in the next section.

Potential factors to facilitate or hinder the change. Both administrators were aware of the potential problems they could face while implementing the innovation in assessment as Christopher mentioned:

Part of me is a little bit concerned because with every change comes challenges.

The transition period, I anticipate, is not going to be very easy. But we’re getting ready for that. And we decided the best thing would be to be prepared beforehand and anticipating all sorts of hiccups (p. 12).

In the interview data with the Administrators, the following key factors emerged that could potentially hinder the successful implementation of the assessment innovation:

- Capabilities of students for integrated skills learning
- Teachers’ abilities to teach integrated skills
- Material availability for the new tasks
- Logistics and technology for speaking tests
- Teachers buy-ins and teacher resistance to change

These factors will be discussed below along with the administrators’ proposed strategies to address these issues.

Relating to the *capabilities of students for integrated skills learning*, new students would not have known that there was an evolution; the norm was the situation that new

students entered upon coming to this EAP program regardless of home country learning norms⁷. The concern according to Christopher was that:

...some of our students come from pedagogical backgrounds or educational systems where they heavily rely on memorization. So they memorize the text and it doesn't matter what the topic is, they try to reproduce that same memorized thought, just change words here and there. We're trying to move away from that.

Anderson said “using integrated testing procedures, using authentic projects and group work” can become the new norm and students can take this as “Oh this is Canada; this is how classrooms are here and this is how they assess here.” Also, as mentioned earlier, Christopher had conducted a focus group with the Summer 2016 pilot project's students and they all seemed to be interested in the new testing procedure. However, students who were already in the program and were used to the former testing regime would be going through a transition, and this would need support and lead-time to navigate the new approach.

As for the *teachers' abilities to teach integrated skills*, Christopher said that most teachers were aware of the new course learning outcomes (CLOs) and the administration had already had “the buy in” from them, but Anderson had his own reservations and said “teachers may not know or have trouble instructing” using an integrated skills approach. Previous washback research mentions that various test factors can influence the degree and kind of washback on teachers (Alderson & Wall, 1993; Hughes, 2003; Spratt, 2005). Some of these factors can be the test's “proximity, its stakes, its purpose, the format it employs, weighting of individual papers and how familiar it is to teachers” (Spratt, 2005,

⁷ For this reason, this study did not include Phase 3 student questionnaire results in the analysis (see Chapter 6)

p. 23). These issues will be further taken up in the process of implantation (chapter 5) and the consequences of the testing regime change (Chapters 6) of the study.

Another major challenge, *material availability for the new tasks*, was also highlighted in the interviews with the Administrators. Christopher mentioned that so far the focus of the textbooks and books used in the EAP program was teaching ESL. However, after the change, books will have more “transferrable skills”, irrespective of the majors students declare for their university studies. Such books would include sections on integrated writing, summarizing and citations. The final tests would be based on the themes from a textbook. Christopher mentioned that this would also emphasize the direct role of ‘the effort’ (in class participation), which had been missing in the program.

Right now, because we’re proficiency based, we don’t really have [marks for] effort. For us, effort means nothing. I think this will help that as well. Effort will be appreciated in the program and the final exam. This will make students go and study those two chapters of course and then they are going to study supplementary materials related to that theme.

Anderson’s concerns were, however, with the difficulty of getting suitable textbooks. This problem was compounded as students were asked to review three to four readings prior to the final tests and every exam developed “must be in line with the textbook so that could be a potential challenge.” Again washback research is divided on the issue of the effects of materials on different stakeholders. For example, Lumley and Stoneman’s (2000) study of teachers’ and students’ reactions to a learning package for a newly introduced test showed a mismatch. The teachers saw the potential of the materials “as a teaching package, containing relevant and worthwhile teaching activities, including but

extending beyond test preparation” (p. 75). Students, on the other hand, were concerned with “familiarizing themselves with the format of the test and seemed to be relatively little concerned with the learning strategies proposed and the broader suggestions for improving performance” (p. 75).

Concerns were also expressed for the *logistics and technology for speaking tests*. Anderson said it was not as easy to assess speaking as marking multiple-choice tests, and it required a lot of work. However, Christopher felt that once the rubric was ready, “marking will be easier”: “Once teachers know how to mark an audio passage, it is pretty much the same as a written text. Once you have the rubric it is *completely easy*.” This comment shows an underestimation of the issues relating to rater training and on-going monitoring, etc.

In addition, challenges of the logistics of administering speaking tests in computer labs and a need for tech support were also mentioned. Christopher said at the pilot run of the speaking test only six students could record partial answers. Workshops were needed to show teachers “how to invigilate tests, implement tests, and integrate tests in teaching”. However, he didn’t have much doubt about the technology as he said, “when it comes to recording, RELANPRO [the computer program in the labs] is a solid tool, it is easy, and usable tool”.

Finally the main challenge foreseen was *teachers’ buy-ins and teacher resistance to change*. Christopher said that he anticipated that teachers might not be very comfortable with these new tests because they “will entail an entirely different pedagogical practice”. As previously noted, some teachers were teaching integrated skills, but others were not, so “assessment would be an issue for them”. He added that a

student may have “solid grammar and ideas” but not “content from the reading passages” and some teachers may question why content is being tested. Anderson felt that teachers’ opinions might be divided about the testing regime change as he said,

There are going to be teachers that don’t like that, and there is going to be resistance.... It seems like most of the teachers feel that it’s about time we went this way [but] for many of our teachers, yes, what we’re doing doesn’t feel right.

What didn’t occur during my discussions with the Administrators, however, was whether any kind of needs analysis or follow-up research was planned in order to discover how much buy in was actually happening or how specific challenges were being addressed. There seemed to be many potential risks, but not a lot proactive exploration into mitigation strategies in order to address potential resistance.

From House’s (1981) perspective, the innovation in the EAP program seems to be a Technological, rather than Political or Cultural (see Chapter 2). From the technological perspective, the focal point is the innovation itself, and its effects, rather than the context. The administrators seemed to believe in finding the one best way to accomplish their goal of implementing the ISTs.

Finally, in terms of the factors that would facilitate the diffusion of innovation, both administrators highlighted the importance of communication, support and training. Or, as Anderson noted “getting the right tools, materials, and resources for teachers to work with”.

Chapter Summary

The purpose of this chapter was to get a picture of what teaching and learning were like in the EAP program prior to the testing regime change. This baseline provides a

point of comparison for the later Phase 2, and Phase 3 of the study. The research question explored in this chapter was: What evidence is there of washback in the former testing regime, and what is the intended washback of the new testing regime? In order to answer this question, I referred to the antecedents from Henrichsen's (1989) model and I used these to address the first half of research question 1. The former testing regime had potentially exerted both positive and negative washback on teachers and students. However, consistent with the washback literature, the variability and intensity of this washback was different for different stakeholders.

Positive washback from the former testing regime was in the form of making learning relevant to students' needs, and preparing students' for standardized high-stakes tests. All teachers in the program claimed to teach communicatively and believed in authentic assessment tasks. Most teachers were using both formal and informal types of assessments. While the formal test tasks were to prepare students for the final ExitTest and involved using practice materials from other standardized tests, informal assessment tasks were comprised of various classroom-based assessments. The use of both formal and informal assessment was also evident in students' questionnaire and focus groups data. A few teachers stressed that their classroom assessments already involved assessing integrated skills and felt that changes in the tests would not have any significant washback on their pedagogical practices.

On the other hand, negative washback of the former testing regime was found in the form of skills isolation, diminished value of speaking and other academic skills, and an expressed need for change. All stakeholders agreed that writing skill was the most dominant skill taught in the program and much less attention was paid to the content of

writing as compared to the organization and language. One of the other potential sources of negative washback was the importance given to the external Testing Office in the final outcomes of the program and teachers showed their disapproval towards that. Thus, both teachers and administrators were ready for changes in the testing regime. However, at the time of interviews, teachers were much less informed about the intended changes and there was a lack of communication regarding the impending changes.

The second part of Research Question 1 probed for views regarding the intended washback of the new testing regime. The findings from interviews and official communication described specific changes in the final test formats, i.e. introducing integrated skills testing as oppose to discreet multiple-choice tests, and formally testing speaking skill. This suggested to me that in addition to examining the implementation dynamics through factors that will facilitate or hinder a diffusion process, I should look for evidence of skill integration and a focus on the teaching of speaking in Phase 2 of the study. The next chapter presents the results of Phase 2 of the study, which focused on the process of implementing the new testing regime. The washback factors facilitating and inhibiting the implementation of the new testing regime will be described and discussed.

Chapter 5: Results and Discussion of Phase 2: the Process phase (*Implementation Dynamics and Immediate Washback of the New Testing Regime*)

Introduction

In this chapter, I present and discuss the results from Phases 2 (see Figure 3, Chapter 3). This phase was concerned with the implementation of ISTs and the immediate washback of the ISTs on teaching and learning as accounted by teachers and students in the EAP program. The results were obtained through analysis of the following data sources:

- two sets of teacher interviews (during and at the end of one semester, after the completion of one cycle of implementation of the ISTs)
- administrator interviews at the end of the semester
- student online questionnaire (after the midterm test) responses and a focus group
- document analysis of the new ISTs

For an effective introduction to an educational change, Henrichsen (1989) suggests it is important to know the context (see Chapter 4), and also to understand the implementation factors that are in operation when the innovation takes place. In his hybrid model, Henrichsen (1989) describes the process of implementation as a combination of the source (innovator), message (innovation), plan and strategies, channels of communication, receivers (their awareness, interest and evaluation), and factors that facilitate/hinder change. In my study, the administrators were the sources, the new testing regime was the innovation, and the plan and strategies of the administrators were the diffusion of innovation and all have been discussed previously in Chapter 4.

Henrichsen describes mainly four factors that facilitate or inhibit the implementation process (also see Figure 5, Chapter 3):

- *within the innovation itself* (its originality, complexity etc.)
- *within the resource system* (the management team and the management structure)
- *within the user system* (the classroom, adopting units etc.)
- *inter-elemental factors* (the interaction between the above three factors)

The comments of the teachers, students and administrators regarding the presence or absence of the above factors in their own contexts provided the data that allowed me to perform an analysis (Saldaña, 2013, 45-148, see also Chapter 3) to determine the washback of these implementation factors in the new testing regime on the quality of teaching and learning at the EAP program. To remind the reader, washback across these factors was viewed as positive washback arising from intended actions brought about through the changes to the assessment regime versus negative washback being the unintended consequences or “side effects” of given intended actions.

Research Question 2: What Evidence is There of Washback Factors Facilitating and/or Impeding the Implementation of the New Testing Regime?

The second phase of research explored the immediate washback of the new testing regime and the four main factors that facilitated or hindered the implementation of the ISTs in the EAP program. Each section that follows will contain the results and discussion of the four evaluative criteria of the process of implementation: *the Innovation*, *the Resource system*, *the User system*, and *the Inter-elemental factors* with a special emphasis on teachers and students. These two groups form the key players in any

washback study. I begin with an analysis of the first factor - the characteristics of the innovation i.e. the new Integrated Skills Tests (ISTs).

The Innovation Itself: New Integrated Skills Tests

There were a number of differences between the old ExitTest and the new ISTs. Firstly, the new testing regime was comprised of two tests. One was a Reading-to-Write test with a reading passage and two writing tasks related to the reading passage. The other was a Listening-to-Speak test that included a listening passage(s) with two to four speaking prompts. The texts used in both tests were related to a textbook chapter. One of the strengths of the new ISTs was that the topics drawn were from the TLU domain of academic English in use (also see Chapter 3, test-task characteristics). The new tests assessed speaking skills that were not tested in the former testing regime.

Secondly, the ISTs were an achievement test while the ExitTest was a general proficiency test. A proficiency test assesses general ability and uses that as a measure of global competence in a language rather than any specific content, course or curriculum. An achievement test, on the other hand, aims to measure what has been taught and is, generally, based on a detailed course syllabus or content (Hughes, 2003). Previous washback literature has suggested that one way to produce positive washback is to align curriculum and tests, by matching the content and format of the curriculum with that of the test (Shepard, 1993; Andrews, 1994; Green, 2007).

Finally, the new ISTs were developed in-house within the EAP program, whereas the ExitTest was developed and administered by the Testing Office of the university. What did not change in the new testing regime were the decisions about the final outcomes for students. The Testing Office was still in charge, and looked at the final

achievement test writings of students to make a decision about passing or failing a student. I will discuss the implications of this decision in detail later in this chapter (see *factors within the Resource System*).

Henrichsen (1989) lists eleven characteristics of an innovation, which influence the decisions made by potential users to accept or reject it. These are: *originality, complexity, explicitness, relative advantage, trialability, status, practicality, flexibility, primacy and form* (see also Chapter 2). I looked for these attributes in my data analysis of teacher interviews, administrator interviews, student questionnaire responses, and focus group discussions. I discuss these in the following section.

Originality of the ISTs. Henrichsen (1989), citing Pelz (1985) states, “the degree of originality in an innovation determines the nature of the change process and it can be categorized as: origination – “the innovation is invented locally without benefit of a prior model; adaptation – the innovation is modified from external examples; borrowing – a standardized model is copied with little change” (p. 82).

When asked about the differences between the old and the new tests, there were a number of comments by teachers about how the IST differed from its predecessor, ExitTest. Teachers were of the view that the new tests seemed to be mostly modified from integrated skills proficiency tests, such as TOEFL iBT or CAEL. In this sense, the originality of the ISTs seemed a cross between Henrichsen’s proposed adaptation and borrowing categories and the new tests were not considered original based on teachers’ accounts. However, all teachers remarked that while the old test had global and generic language content, the content of the new tests related to the textbook content covered in class. As Alan commented:

It [the new ISTs] is no longer an isolated proficiency test. The tests are integrated into the content that we are using to facilitate language growth. We're teaching a theme in class and that theme is being reflected in the test (1st interview, Phase 2).

Alan's comment is a reflection of positive washback of the new tests. Green (2007) suggests that "the more closely the characteristics of the test reflect the focal construct, the greater is the potential for positive washback" (p. 14), but at the same time he warns that this overlap of the characteristics of the test and focal construct (or the test design on its own) are not sufficient enough to produce positive washback, other factors such as test use, test stakes and consequences of test scores are equally important.

Complexity, practicality, and explicitness of the ISTs. In my analysis, I decided to combine three aspects of innovation: complexity, practicality and explicitness. The reasons for combining these different factors together were: firstly, the 'complexity of an innovation' is defined in many ways in the innovation literature (see Chapter 2 for details). Secondly, practicality of an innovation can also be related to complexity as either difficulty or ease of understanding an innovation or technical complexity as ease or difficulty of implementing an innovation. Finally, explicitness of an innovation is about whether the demands of an innovation can be met by the intended user system (Henrichsen, 1989). In teachers' data, I looked at the ISTs' content and format, testing ease/difficulty of integrated skills, academic teaching/learning skills and teacher constraints, and in students' data I looked for the content, format and ease/difficulty of learning/testing of the ISTs. Four main themes emerged in my analysis: comfort in using

the new technology, pedagogical issues with the included academic skills, instructional complexity in the new tests, and test bias of the new ISTs.

Regarding *the comfort in using the new technology*, all teachers commented on the complexity of the ISTs for their students, especially the Listening-to-Speak test. Teachers' comments suggested that there was an assumption made by the administration that all teachers and students were comfortable and fluent with technologies, but this was clearly not the case as Lisa's comments suggest:

...if I were to point out any specific hiccups or issues..., it would have more to do with the applications that we were using, the technological applications. I think students weren't used to recording themselves, weren't used to hearing their own voices.

The EAP testing literature suggests that technology can help, impede and confuse instructors and students, who might need help to navigate new technologies (Fox & Cheng, 2015; Weigle & Malone, 2016). This study supports these findings. In the EAP program, it was found that not only teachers, who were light users of technology, but also students who were generally considered more technology-savvy had issues with the technological applications required for the new ISTs.

The Listening-to-Speak test took place in a computer lab, which had around thirty PC stations. In their individual stations, with a headphone and microphone, students listened to two audio recordings and completed two speaking tasks: summarizing the first audio text, and replying to a given speaking prompt. While students recorded their responses, one administrator stood in the lab announcing the time left for students to complete their speaking tasks. To record their responses and pay attention to the

remaining time was very unnerving for students. They were under tremendous pressure to demonstrate their speaking abilities while under strict time constraints. As Valdo, a student, commented in the focus group:

The speaking test was stressful for me because it's weird and nerve racking to record your own speech. At the same time you have time. My ordinary speech when I talk to a real person and when I try to record my own speech by using like RelanPro, or other electronic equipment, it's different. During the test, you have pressure, you focus not only on your grammar and pronunciation. You focus on time limitation, the main ideas which you like to provide, give a complete answer.

Students were also acutely aware of the shortcomings of this method of testing speaking and questioned the validity of such tests. Charlie, another student, commented:

But for the listening, and speaking, I think it's very rare because there is no situation in your daily life where you're listening to a long recording of 15 minutes then someone will give you two question about summarizing it in two minutes and then you need to speak so fast. And you need to pay attention to your grammar and pronunciation. It's not the way people communicate.

Features of a test, such as the title, labels of subtests etc. can also have significant washback on teachers and students (Fulcher, 1999) (also see test-task characteristics, Chapter 3). The complexity of the new tests was also evident in the instructions provided in the tests. For example, the Reading to Write test had a list of at least seven bullet points of instructions for students:

Your responses will be assessed using the following criteria:

- a) the variety and accuracy of your grammar;
- b) the range and effectiveness of your vocabulary;
- c) the clarity between ideas and the ease with which they can be followed throughout your text;
- d) the quantity and / or relevance of the ideas that you have selected from the reading materials to support your answers;
- e) the degree by which you appropriately incorporate and / or paraphrase language from the reading materials into your contributions;
- f) the extent to which you have correctly cited the reading materials;
- g) the overall strength by which you successfully communicate your meaning (Mid-term writing test, Oct, 2016).

These instructions were more complex than the instructions in the former ExitTest. Similarly, the instructions provided for teachers to conduct the Listening to Speak test were equally complex with at least sixteen steps to follow. That is why Anderson, the administrator, commented on the impracticality and complexity of the new tests:

That's the problem. They aren't practical because they are too complex.

There were instructions running in 15 or more bullet points. It's not sustainable. We need to find ways to make it sustainable, accessible for everyone.

He also pointed out that there was a higher failure rate for the midterm test, compared to the midterm failure rate under the former testing regime, because of the new testing procedures and he felt that the fluctuating pass rates could be because of the practicality, complexity and inconsistencies in the test practice. He said one way to stabilize the pass rate could be 'standardization' of test practices in future semesters.

The second set of comments was about the *ease and difficulty of teaching integrated and academic skills*, such as summarizing, paraphrasing and referencing. There were mixed views from teachers regarding teaching these skills. All teachers said they were satisfied to do away with the five-paragraph essay format and genre teaching such as compare and contrast essays. Similarly, in speaking, they considered fluency as more important than absolute accuracy. Just as with the shift in writing from parroting a model, there seemed to be a shift in teaching speaking from mimicking to having a dialogue that was freer and less limiting. This is more in keeping with the academic environment that students will encounter in their future studies. However, in terms of experimenting with teaching speaking, Alan, a fairly new teacher in the program said “It was very hard at the beginning only because I’m figuring it out while they’re [students] figuring it out.”

Three teachers found teaching academic skills difficult, firstly, because of the lack of guidance and standardization from the administration. There were no guidelines provided for using any particular citation management system. Teachers taught whatever citation system (e.g., APA or MLA) they were familiar with and this caused inconsistency among different sections of the graduating level. Also, citation formats are discipline specific and exceedingly challenging to master. During the period of change, when communication is paramount, it seemed, according to most teachers, there was little communication and support with regards to expectations and outcomes.

Secondly, low English proficiency levels and lack of students’ awareness regarding academic skills, such as citation requirements, may have contributed towards difficulty in teaching academic skills. As Mary noted, many students were simply not

aware of academic conventions in North American universities, and a number of the teachers specifically pointed out that their students often considered *verbatim* recall (cutting and pasting as well) to be the highest form of scholarship, because it honoured the original source. There seemed to be tension between teaching discrete, individual skills and more clustered, global skills development because both of these are needed for academic success. It seemed there were issues of timing and progression for these in that there needed to be skill introduction, skill building, skill using and then, ultimately, global skill acquisition, but that didn't seem to happen with most teachers. It was more like a patchwork of skill development as evident in Lisa's comment:

Oh I think that's difficult. It really is. There's the basic principles of summarizing is a shorter version and paraphrasing is the same length, but so many of my students would just take a word, find a synonym, plug it in, and it was really hard to weed them away from that (1st interview).

Or

...very difficult because our students don't understand the concept of paraphrasing and summarizing. The idea of citing is very alien to them and they just want to copy and paste. Also, it was very difficult because they lack the academic vocabulary that they need to be able to do this. It was very challenging in my class (Mary, 1st interview).

The final set of comments from teachers was about *scoring the new tests*. Teachers were provided with analytical rubrics for marking the ISTs. The new tests incorporated the textbook content, so the rubric had criteria related to the use of sources in addition to criteria related to global competence. Table 11 shows a scale for scores interpretations

based on the rubric for the Reading to Write test. The weighting given to each component and the conversion of these components into a numerical number resulted in issues with score interpretation.

Table 11 *Sample scale for score interpretation on rubric for the reading to write test*

Test tasks	Lexical Accuracy and Range	Grammatical Accuracy and Range	Organization	Use of Sources	Overall Effectiveness	Example% Marks
Student 1	✓	✓	✓	✓	✓	70%
Student 2	--	--	✓	--	--	62%
Student 3	+	✓	+	✓	+	80%

Meets the descriptors = ✓ Surpasses the descriptors = + Does not meet the descriptors = --

For example, two check pluses and three check minuses were considered a failure (62%) as compared to all check pluses that meant a pass (70%). This scoring definitely caused confusion and complications for teachers. Because of this, all teachers confirmed that they were marking and scoring the tests with global competence in mind rather than using the sample analytical rubrics provided by the administration. Teachers' dissatisfaction with the complexity of the rubrics was evident in Stacey's extended comment:

We receive a rubric that was check-plus they meet all the requirement, check they meet most, and then check-minus, then you look at the check pluses and that equates to an 80%. Well, then I said what if...I had to design my own: 2 check-pluses, plus, 2 check minuses, plus 1 check equals, a 70. But then it was kind of, I came up with a scale and I checked with the coordinator and I also said to my co-teacher "This is the guideline that I'm going to use" and no names mentioned but she basically, just dismissed it, so she was marking on a completely different scale, or criteria. So we weren't on the same page, but I'm a little reluctant to tell

another teacher how to teacher even though she's my co-teacher. But I think what we need is very definitive guidelines and more specifics in ranges. "So 3 check-pluses and 2 check-minuses means you give the student a 60-70" and more categories than just the three that we were provided with (1st interview).

Similarly, for marking the speaking test, teachers were not sure about what aspect of speaking skill to mark: the fluency or accuracy as evident in Jill's and Alan's comments below:

I think it's quite difficult because when we are talking about listening to speak, it's really unclear how much we test only the speaking skill or the ability of the student to express his or her ideas. Or we are also testing what they're listening to in the text. I think we are doing both, but it seems that in the way we assess the test we mostly focus on the speaking itself. However, there is this part where we see how much they cited or referenced ideas in the text. Well, maybe there we assess the understanding or listening skill, but it's quite difficult or unclear what and how we are assessing (Jill, 1st interview).

or

For me, it was coming down to accuracy. I can teach fluency, but how do you teach accuracy with speaking that so ephemeral? You said it once and then you never go back to it again. Finding a way of marking and for them [students] to realize accuracy was the hardest part (Alan, 1st interview).

The language testing literature has also highlighted the complexity of marking source-based integrated assignments because raters can have different perceptions of evaluating textual borrowing. Weigle and Malone (2016) indicated a need for rater

training on how raters should account for episodes of inappropriate borrowing when assigning scores to essays (p. 669). Next section is about the trialability of the new ISTs.

Trialability and status of the ISTs. Henrichsen (1989) describes ‘trialability as ‘how easily new users can experiment with an innovation on a limited basis’ (p.84) and ‘status’ of an innovation is described as “the association with a higher social level which can impart legitimacy and attract attention to an innovation” (p. 85). The tests were compulsory rather than optional. Teachers did not have any opportunities for first trying it out and then deciding to switch over to it or stay with the former test. Since the new testing regime was implemented in one go, there was no chance to evaluate it in stages. Similarly, the status that users might enjoy if they chose to use this innovation was also not discussed. The ISTs were obligatory rather than a matter of option.

Relative advantage of the ISTs. One way to identify the relative advantage of a new test is to compare it with the old one. This was partly discussed in the ‘Originality of the ISTs’. The ISTs were considered as task- and topic-based integrated tests, designed to elicit academic English language skills from students. Integrated skill tasks first try to create an academic setting by providing tasks related to any of the receptive skills (e.g., reading or listening), which mirror the activities that university students are engaged in on a day-to-day basis. Students then integrate ideas from listening and reading sections of the test, either in reply to a speaking prompt, or by responding to a written prompt.

Teachers’ views about the advantages of ISTs. Teachers spoke positively about the new ISTs as “these tested useful academic skills as compared to receptive skills and 5-paragraph essay of the previous tests” (Jill, p.2). Most teachers believed that the ISTs were having a positive washback on their in-class teaching practices as Stacey (1st

interview) mentioned, “there is more emphasis on citing sources, synthesizing, summarizing and paraphrasing....we were not asking them to do it in test situation or a timed situation earlier for example in the writing lab.”

For optimal positive washback of any test, there should be little if any difference between activities involved in learning the language and activities involved in preparing for the test (Messick, 1996). Similarly, Travis was of the opinion that “students now need to be better at actually expressing their ideas in written and spoken form... and that’s very university related. I think it will help them for more authentic [assessments]” (1st interview, p. 2). He added that students would not only take information in answering multiple-choice tests, but would actually produce information for a purpose. Emphasizing the advantages of integrated skills testing, the official communication from the administration gave the following rationales for introducing the ISTs:

- To assess authentic academic and study skills such as summarizing, paraphrasing, synthesizing information from multiple sources, and citing.
- To encourage critical thinking and the integration of receptive and productive skills.
- To combine proficiency with achievement.
- To use a more effective means of assessment (slide#4).

However, there were a few teachers who believed that the new ISTs were having a negative influence on teaching and learning. Reasons given for these were:

1. Low English proficiency of students,
2. Confusion between teaching ESL vs. EAP, and
3. Lack of student readiness for academic preparation.

Mary's extended comment provides evidence of her perception of negative washback from the new ISTs:

I have a very difficult time with this. In all my years of teaching, this has been the worst group, the one that I've experienced the most challenges with. Not because they're stupid, or ignorant, no, because they really lack the fundamental elements, the basis that they need to enter university. They are not prepared for university and their English is really below the level. I have the GL class, and my students are two levels below. I really think they are, because they cannot even produce a simple sentence accurately. So what do they need? They need grammar. We stopped doing grammar and I think this is a big disservice to our students because no matter what level you are in, if you are in an ESL class, you should do grammar. If you are in an EAP class, that's a different story. But in ESL you need to know your grammar, because when you have good control of your grammar, at least in my opinion, your English proficiency improves along with the acquisition of new vocabulary (1st interview).

Similarly, Kathy's concern with regards to low level of English proficiency of students was that students do not comprehend the instructions so the most important skill for her to teach her students is 'listening' than any other skill.

Students' views about the advantages of ISTs. Bachman and Palmer (1996), suggest "one way to minimize the potential for negative impact on instruction is to change the way we test so that the characteristics of the test and test tasks correspond more closely to the characteristics of the instructional program" (p. 33).

Since the teachers expressed their views that the new test's format now resembled more with classroom activities providing an "evidential Link" (Messick, 1996, p. 247) between the classroom activities and test tasks, I wanted to confirm this with students' accounts of the new tests. Therefore, in the online questionnaire I had three questions related to the relative advantage, practicality and explicitness of the new ISTs (see Table 12).

Table 12 *Student questionnaire results regarding the new ISTs*

<i>Overview of student Responses (n= 64)</i>					
Survey Questions	Definitely Yes	Yes	No opinion	No	Definitely no
Did you like the format of the Midterm test?	3%	27%	28%	25 %	17%
Do you think that the midterm tests were a reflection of the work you did in class?	5%	44%	21%	18%	13%
Do you feel that the extra activities you did in class helped you prepare for the midterm tests?	18%	47%	23%	7%	5%

While around 30% (19/64) of students said that they liked the format of the new tests, around 42% (27/64) of students did not like the format of the new tests. Results also suggest that there was a significant correlation ($r = .653$, $n + 64$, $p < .01$) between students' attitudes towards the new tests and the tests content being similar to the classroom content (see Table 13).

Similarly, there was a significant correlation ($r = .553$, $n + 64$, $p < .01$) between students' attitudes towards the new tests and the extra activities that teachers conducted in class in preparation for the tests. Around 50% (32/64) of students agreed that the content of the midterm tests were a reflection of the work they did in class. 67% (43/64) of students agreed that the extra activities that teachers carried out in class were helpful in preparation for the tests.

Table 13 *Spearman's rho correlation of positive washback of test format*

		Attitude to Test Format	Is the test Content similar to Class Work	Extra Activities Helpful for Test
Attitude to Test Format	Correlation Coefficient	1.000	.653**	.523**
	Sig. (2-tailed)	.	.000	.000
	N	64	64	64
Is the test Content similar to Class Work	Correlation Coefficient	.653**	1.000	.474**
	Sig. (2-tailed)	.000	.	.000
	N	64	64	64
Extra L/S Activities Helpful for Test	Correlation Coefficient	.523**	.474**	1.000
	Sig. (2-tailed)	.000	.000	.
	N	64	64	64

Correlation is significant at the 0.05 level (2-tailed)*

Correlation is significant at the 0.01 level (2-tailed)**

Hughes (2003) has suggested when a test reflects the aims and the syllabus of the course; it is likely to have a beneficial washback. The test tasks on the ISTs were thematically linked and also mirrored the classroom activities. Furthermore, the wider overlap between the focal test construct and the test designs seemed to have exerted a positive washback (Green, 2007). There were two main advantages of the Reading-to-Write test as suggested by students in the focus group: using computers and scaffolded writing tasks. First, student responses indicated that using computers for writing was advantageous. This is evidence of technology as a form of mediation for students. See Valdo's comments below:

For the writing test, for me it's easy to organize my ideas and if you make different grammar mistakes, for example, punctuation mistakes, you can just eradicate them by using the special functions of the Microsoft Word.

Second, the textbook chapters scaffolded writing tasks. Topic familiarity gave students confidence in responding to the writing prompts in the new tests. These skills are also

useful for the future university environment, i.e., the use of previous knowledge to generate new ideas, as student Ali commented:

Because the topic was not randomly chosen for us it was based on what we had studied. And we studied just like 4 chapters and they have 4 topics, and we can do a research before the test and have more of an idea, like knowledge about them. When it comes to writing for the test, we have a lot of idea for the topic so we can write easily. We had our previous topics and list of vocabulary that we could use. I think it was very helpful because we spontaneously could associate these topics with our previous passages and it is very helpful to generate new ideas.

To summarize, the new ISTs seemed to have an advantage over the old ExitTest because ISTs incorporated content familiarity as an important test feature that was also academic in nature. The inclusion of important academic skills in the tests avoided Messick's (1996) notion of construct underrepresentation. These were the evidences that supported the notion that correspondence between test tasks and classroom practices had created a positive washback on most teachers and students in the program. However, there were also issues with the complexity of the Listening-to-Speak test, the scale clarity of the rubrics, and standardization with the new tests, which seemed to be exerting negative washback. In sum, the factors *within the innovation itself* that facilitated or inhibited the diffusion of implementation of the ISTs are summarized in Table 14.

Table 14 *Summary of the characteristics of the ISTs*

Within the Innovation Itself	Manifestation	Facilitative or Inhibitive?
Originality	ISTs were seen as being different than the previous ExitTest (better reflection of the classwork)	Facilitative
	Tests seemed to mirror classroom work	Facilitative
Complexity for teachers	Teaching academic skill were difficult for some teachers	Inhibitive
	The rubrics used to mark the tests were not teacher friendly	Inhibitive
Complexity for students	Most students liked the RW test	Facilitative
	Students were concerned about the complexity of the speaking test	Inhibitive
Explicitness	Official correspondence with regards to the new test was explicit.	Facilitative
	Many test-related materials such as instructions and rubrics were not clear	Inhibitive
Relative Advantage	Most teachers and students felt that tests had a positive effect on teaching	Facilitative
	Content familiarity was considered positive both by teachers and students	Facilitative
Trialability/ Status	No comments	-
Status	No comments	-
Practicality	Reading-to-Write test was practical	Facilitative
	Major issues were with the administration of the speaking test	Inhibitive
Flexibility/Primacy/Form	No comments	-

The Resource System

The second evaluative criterion used for the evidence of washback in Phase 2 of implementation was the resource system. The Resource system relates to the individual or units responsible for planning and organizing the innovation. The main characteristics of the resource system are: Capacity (“to retrieve and marshal diverse resources, and influence opinions”); Structure (“division of labor and coordination of effort”); Openness (“a willingness to help and a willingness to listen and to be influenced by user needs and aspirations”); and Harmony between different groups working within the system (Henrichsen, 1989, p. 86). In my analysis, I did not find much evidence regarding the

Capacity of the resource system, but there were many comments on the structure, openness and harmony in the EAP program by the teachers and administration.

Structure. ‘Structure’ can have different meanings in any resource system. Henrichsen (1987) suggests: it could refer to *division of labour* and *coordination of effort*; it could also be resource system’s views of and relationship with the user system; or it could be the resource system’s ability to plan and implement innovations in a structured order. In some of these respects, the effort of the EAP program was deficient especially in terms of the final outcomes of the graduating level for students. There were a number of comments about the structure of the new testing regime of the EAP program. The main theme that emerged was the centralized testing as weak link.

As explained earlier (see introduction to innovation section), what did not change in the new testing regime were the final student outcomes. The Testing Office was still in charge, and looked at the final achievement test writings of students to make a decision about passing or failing a student.

Regarding *the centralized testing as weak link*, there seemed to be a real division of labor going on with the testing office experts who were given much higher credence than the teachers. All teachers were of the view that there was not enough collaboration among teachers, administration, and testing experts in terms of the final test outcomes for students. Since the Testing Office was outside of teaching, testing and teaching did not become part of the same process, but an add-on to the process. The external testers were unaware of the dynamics inside the classroom, even when the tests materials (i.e., content) were taken from the classrooms. In this EAP program, there seemed to be negative washback in this respect where teachers felt that they had no voice in the

process, as Alan mentioned: “they’re [external testers] the ones who create the questions, they take the theme with which we were teaching from, and then they create the questions. So the teachers have no say.”

The role of testing experts, as outside arbiters of testing in isolation from what teachers did in the classroom created another potential complication. Some argue (Weigle & Melone, 2016), that academic preparation cannot be tested by outside experts who only measure results and not the process. Testers focus predominantly on that, which can be measured: such as writing or writing-based tests of listening and reading. The external testing experts cannot capture the learning and successes that go on in the classroom throughout the semester or the academic preparation evolution that occurs in the classroom, for example, academic readiness, academic confidence, and student maturity etc. These were now part of the new Program Learning Outcomes and Course Learning Outcomes of the EAP program and important aspects of student success. What remained, however, significant at the end was that all importance was given to only ‘writing skill’ on the new ISTs. This was no different from the former testing regime where writing was the dominant skill. Clearly, there seemed to be two solitudes between the external test experts and classroom (frontline) teachers in the EAP program and little collaboration between them. Excluding teachers from the final marking was evidence that teachers’ expertise was neither valued nor trusted as Stacey commented:

They [the Testing office] are marking the tests and if they say a student fails, they fail unless we come up with valid reasons why they shouldn’t fail. I am not happy with that approach. I’ve talked to other teachers who aren’t as well but anyway..... And we were told that we could consider

the listening speaking with reference to the reading and writing if it was a borderline case. I asked admin , “Does think that mean that we can recommend a pass for a student who does not pass the writing test?” no. So, that said...(2 interview).

Three teachers commented that as a back office function, the Testing Office should be supporting what the frontline was trying to do, but it seemed to be more of a barrier or hindrance than a support. That is perhaps why teachers suggested that the roles of the two should be reversed. The testing office, constructed as expert, should be more a check to what the teachers’ marking and tests would be, as Stacey further added:

Quite frankly, I think the roles should be reversed. I think we should mark them and say here are a few questionable papers, what do you think and give a few papers to experts and say what do you think? (2nd interview).

Further, complications also seem to occur because of the centralization of testing. The Testing Office issued only pass/fail decisions with no middle ground or no rationale that could be communicated to students unless it was specifically requested. The following extract from Stacey (2nd interview) conveys all of the teachers’ grievances:

In terms of the final decision? No. They say pass, fail, pass with two PEEL courses, we can contest that but it was the same as before [former testing regime]. The thing that has changed, well obviously the format of the test, but we are still asked to look at the tests but not asked to mark them. In fact, no, each teacher gets a set of tests but I’m really not sure what the point of that is because, especially with what happened between

management this past cycle of tests. I'm just wondering why the testing office is involved at all.

Also, teachers argued that speaking ability was not given enough weighting in the new testing regime to justify the overall students' results as Lisa lamented:

...they [testing office] test reading to write only. Now, I think it would be fair if the listening to speak was given a little bit more weight because ultimately, they [students] could fail everything but pass the writing and be fine (p. 4).

The teachers also mentioned that a number of their students who did well during the semester had failed the final ISTs because of 'the final writing test'. At the time of interview, Lisa's co-teacher (who taught mainly writing skills) was defending the case of two students who did well in class, but failed the final ISTs. Unfortunately, said Lisa, she could not defend these students:

...because I think what the [Testing Office] wants is writing materials. I have tons of recordings. I could have easily sent it to them and say, they're able to be articulate and it's heartbreaking because they've done presentations every week and they're good (2 interview).

Thus, the centralization of the testing, which was evident in former testing regime, continued in the new testing regime. Even though not directly related to the test, this was still an evidence of the negative washback of the new testing regime on the EAP program.

Openness. In relation to the innovation itself, openness means flexibility and adaptability, but for the Resource system Henrichsen (1987) suggests, openness means "*a willingness to help and a willingness to listen* and to be influenced by user needs and

aspirations” (p. 158). The definition incorporates two major characteristics: a willingness to help; and a willingness to listen, I will discuss each separately.

The first one is “*willingness to help*”. The EAP program was willing to help its teachers and students during the implementation period. All teachers commented that the administration was ready to help whenever anyone approached them, but the “open-mechanistic” leadership style (Markee, 1997) was clearly evident in administrator Anderson’s comments:

In terms of curriculum materials, we had to update our testing guidelines, which we provide for our teachers. We had to give teachers new rubrics and new standards for them to work from. That’s a major change (Dec, 2016).

In House’s (1981) technological perspective, teachers were seen as the passive recipients of change being told how things were being amended. An open-mechanistic leadership style “maintains hierarchical authority and central control over decisions but seeks to increase the flow of information about the environment” (Rondinelli et al., as cited in Markee, 1997, p.65). There seemed to be a lack of communication and wraparound support for teachers and/or students with regards to programmatic and testing protocol changes. Lisa commented some help was available, but she would have had to contact the administrator to get information regarding the new tests; it was not readily available. Similarly, Travis (2nd interview) said “there needs to be more bottom-up support.”

The second characteristic is “willingness to listen,” which was also a characteristic of the EAP program, but only during the initial stages of implementation

Travis added to his comments about bottom-up support that “ Personally, I found that at the beginning of the session, there was a good dialogue, good interaction. The office and teachers.... I don’t know that there was enough in total, throughout the whole session. I think we need to work together more effectively” (2nd interview).

While there never seemed to be much doubt from teachers’ comments about problems in the relationship with the external testing office, no proactive solutions seemed to address these problems. The response of the administrator, Anderson, in this regard best illustrates EAP program’s lack of openness:

Definitely we didn’t involve them [teachers] in the process. The final tests were put under lock and key. We want to make sure that all students have the same leveling field. If one teacher knows the test then its hard to not test prep them. Although some may do some may not. So kind of followed the previous practices.

This comment also illustrates that there was no change in practice after the testing regime change. It was still the same as in former testing regime (see also Variability of testing process, chapter 4).

Jill corroborated this as “they [the Testing Office] are still acting as a gate keeper, they do, I don’t know if I should say, they decide.... They do have a huge control over the outcome of the final test.” Other teachers also had a very negative reaction to the way change was being handled so much so that Stacey dreaded the future as she suggested ‘there is no choice in new paradigm’. The implementation dynamics of the new testing regime was exerting positive and negative washback simultaneous. For example, Stacey was “satisfied” with the outcomes of the change, but was unhappy and even fearful of

what the future would hold because of the way change had been handled that semester – centralized and imposed with little teacher buy in or ownership or in Stacey’s views: “Is there a Likert scale? No, I am not very satisfied. Because I don’t think we were consulted, basically that’s it.”

Harmony. Harmonious relationships among different people and elements are of importance in any implementation project. Henrichsen (1989) suggests “planner/managers of implementation projects who ignore the problem of poor relations until it reaches the critical stage may find their efforts crippled by this internal “barrier”” (p.159).

In any project that involves different individuals working together, minor disagreements that require compromise and negotiation are normal. The EAP program was no exception to this rule. Unfortunately, in the important case of making assessments both for continuous assessment and final test pass/fail decisions, disharmony seemed to be beyond the normal level. There seemed no consistency and standardization in classroom assessments across different sections of the GL. Even among the same teaching teams, classroom assessment was widely divergent. There needed to be clear policies followed by clear communication and support protocol for students and teachers to understand. When asked about her satisfaction with the support provided by the main office, Stacey (2nd interview) said she was not satisfied.

Also, with everything centrally located within the Testing Office, there seemed to be a lack of checks and balances within the office, which left frontline teachers fixing Testing Office errors. In essence, under the new testing regime, teachers had the responsibility for carrying out tests, but none of the control. This seemed to cause

resentment at least for teachers like Stacey and Maria. This was quite normal when the two dynamics became divorced from one another. Going back to the silos of teachers, administration and testing, Travis commented:

I think a lot of it is our office that has a role in this. There might have been a disconnect between all three bodies and that would be something to work on for the future. (2nd interview).

Markee (1997) has also pointed out that outside agents normally do not possess any institutional power to force adoption. In such cases, the potential adopters are likely to view imposed change as illegitimate (Kennedy, as cited in Markee, 1997). Markee further suggests, “outside change agents function best as consultants who help end users identify obstacles and who facilitate changes users believe are appropriate to their context of implementation.” (p.45).

Another important reason for disharmony, mentioned by Travis (2nd interview) was that all teachers in the EAP program worked part-time so it was always difficult for management to bring everyone together at the same time. He said:

One session we have these teachers teaching, and the next session we have some of those. Even if you look at this session and the winter, there are a lot of us who taught this Fall who aren't going to be teaching in the Winter and so, if we skip a session, what happens in the meantime? We come back, we have to learn things over again. And that's sort of how I felt sometimes (2nd interview).

In summary, it was observed that the weakest link in the resource system was the centralization of testing in the program. It looked like a classical example of the

‘empirical-rational or power coercive’ strategies of change in the Research, Development and Diffusion (RD&D) model of change (House, 1981; Markee, 1997). Markee suggests the RD&D model “relies on top-down change. Consequently, teachers – who are at the bottom of this expert-driven, decision-making hierarchy - still do not own the products of this approach and so have little or no stake in their success” (p. 65).

In terms of House’s (1989) perspective on innovation, changes within the EAP program seem to be mainly in terms of technological perspective. Although change implementation was systematic and rational the consumers (teachers) were passive recipients of change. The changes seemed “to represent the interests of those ‘who sponsor innovation’ than ‘who are being innovative’ ” (House, 1981, p.40). House, however, warns about the widespread belief that the RD&D model, which is workable in fields such as agriculture, would work in education sector too. He suggests in education it has become less relevant and less workable, which seemed to be the case of the EAP program. Table 15 summarizes the main characteristics of the Resource System in the new testing regime.

Table 15 *Summary of the characteristics of the resource system in the new testing regime*

Within the Resource System	Manifestation	Facilitative or Inhibitive?
Structure	Centralized testing as weak link	Inhibitive
Openness	Support of administration to help teachers in setting up tests, e.g., computer lab etc. Regular updates on new tests from the administration	Facilitative
	Reluctance of administration to listen to teachers’ suggests regarding the final outcomes	Inhibitive
Harmony	Lack of standardization across different sections of the GL, Teachers doing their own assessment Lack of full-time teachers at the EAP to have standardization	--- Inhibitive

The User System

The third evaluative criterion that I used in Phase 2 was *the User system*. Henrichsen (1989) argues, “various characteristics of the target society or organization can be powerful determinants of success in diffusion/implementation” (p. 87). Therefore, it is important to analyze the ‘User System’ in the change process. Other innovation researchers, such as Fullan (2007, 2015), have also given importance to the role of local characteristics and external factors in the implantation process. Although Henrichsen presents eleven characteristics of the user system (see Henrichsen’s model discussion, Chapter 2), when I started analyzing my data, I realized that some factors were either irrelevant for my study (e.g., geographical location and political factors), or there were overlaps with other categories from the Resource system (e.g., Structure and Openness with Centralization and Administration). Therefore, my analysis in this section pertains mainly to four factors: Communication Structure, Educational Philosophy, Teacher Factors and Learner Factors.

Communication structure. As noted in Chapter 2, in the diffusion of innovation, it is important that a clear message about the aim and the nature of the intended change are conveyed through efficient communication (Henrichsen, 1989). Most teachers commented that they were satisfied with the communication regarding the ‘what of innovation’ or the mechanics of the new tests, e.g., term timing of the tests, question formats of the test, and information updates about the tests. What was lacking was the *how of innovation*. Not much information was provided about how to introduce the new CLOs, such as research papers or teaching/marking the newly introduced speaking skill. Although teachers seemed to be self-sufficient in teaching new skills, there was a gap in

the standardization of these skills. For example, Jill said, “Some elements regarding assessment, or the research projects, that’s mandatory, I think some of those things were not sufficiently, clearly communicated from the very beginning.” This communication gap created confusion for teachers. Fullan (2015) suggests many top-down ambitious projects rush users into adopting the innovation during the implementation phase because change agents want to show that “something is happening.” This was evident in Mary’s statement:

There wasn’t really very much direction on the research projects. It’s one of the objectives that listed here, and that’s about all you get..... So I think that we should, I don’t want to say standardize it, but, at least have some sort of even a workshop on it that says some guidelines maybe (2nd interview,).

When planning the assessment processes, Bachman and Palmer (1996) suggest the intended users should get as complete and detailed information as possible. This should include a list of potential consequences, both positive and negative, and possible outcomes in terms of desirability and undesirability of their occurring (p. 35). In the case of the EAP program, sufficient communication regarding the new testing regime was not carried out.

Educational philosophies. Another important factor in any innovation-implementation campaign is the consideration of the “prevailing educational philosophy” (Henrichsen, 1989, p.91). Does it serve culturally enriching, or purely practical ends? Henrichsen suggests if an innovation is not in harmony with the philosophy, it will have little chance of success. As discussed in the previous chapter (see Table 8, Chapter 4), it

was evident that all teachers placed greater stress on the communicative language teaching and the TLU domain of academic English use. Therefore, they considered the ISTs as being transparent and grounded in practice. The teachers suggested that new ISTs would foster more integration of academic skills and generate integrative thinking and critical perspectives of university-level work. As Alan suggested, “the tests are integrated to facilitate academic language growth”. In sum, the educational philosophies of the EAP program matched the new ISTs, so there seemed to be a better prospect of it being successfully implemented with a potential for creating positive washback.

Teacher factors. Teachers are at the grass root of implementing changes at the classroom level. Henrichsen (1989) suggests that changes in teacher behavior require both commitment and capability, and alone, neither commitment nor capability is enough. He further suggests, “Capability determines what *can* be implemented, and commitment determines what *will* be implemented” (p. 90, emphasis in original). Without teachers’ commitments to change, the change would not occur and if teachers do not have the capability, then the commitment to change would not bring much success.

As discussed in the former testing regime (see also *evolution of testing criteria* in Chapter 4) prior to implementation of the ISTs, all of the teacher participants expressed their dissatisfaction with the former testing regime. This dissatisfaction with the status quo was one of the impetuses for change in the regime. Beeby (as cited in Henrichsen, 1989) states two factors are important in moving any educational system from one stage to another: a) educational level of teachers, and b) amount of training teachers receive.

With regards to the educational level of teachers, all teachers in the EAP program were highly educated with at least a Masters’ degree in teaching English as a second

language (see *Participants* in Chapter 3). Many among these held a PhD in Education or Applied Linguistics. With eclectic and student-centered teaching philosophies, all teachers expressed a belief in a communicative approach to teaching academic English. This is in line with BANA ideologies as discussed in chapter 2 (see section on Henrichsen's Hybrid Model). BANA teachers are more likely to adopt skill-based, discovery-oriented and collaborative pedagogy (Waters, 2014).

When asked about changes made in their methodologies and pedagogical beliefs after the introduction of the new ISTs, most teachers admitted on changing very little because they were incorporating academic skills in their lessons before the change too. However, the focus had, now, shifted from genre-based teaching (e.g. compare and contrast essays) to skills teaching, such as summarizing, paraphrasing and synthesizing. The immediate washback of the ISTs were that teachers no longer engaged in test preparation activities, such as practicing for cloze, skim and scan tests, etc. Stacey said such activities had “disappeared from her teaching vocabulary”⁸. Table 16 summarizes teachers' views about changes made to their teaching and the focus of class activities post-ISTs implementation. What was disturbing was that with the exception of Lisa, no teacher mentioned shifting focus to teaching speaking. Their changes were more pertinent to the writing skills, which supported the prevailing perception regarding the dominance of writing in the program (and the influence of the external high-stakes valuing of writing).

⁸ Compare this to her comment in *Failure to test speaking & devaluation of academic skills* in Washback from ExitTest in Chapter 4.

Table 16 *Teachers' accounts of changes in classroom practices in the new testing regime*

Teacher	If made changes to their teaching	Focus of classes in Phase 2
Stacey	Yes	- Did not use any of the previous test prep exercises such as cloze, skim and scan etc. - More emphasis on paraphrasing and summarizing from text book content -Less emphasis on formulating topic sentences, thesis statements and controlling ideas
Jill	No	-Shifted focus from writing 5-paragraph essays to paraphrasing, responding or synthesizing; using more textbook for content
Lisa	No	-Incorporated pronunciation workshops and integrated them into speaking -Focused on knowledge building on topics to include them into speaking and writing from text book themes
Alan	No	-Had always been a content-focused teacher and taught through materials; incorporated textbooks themes
Mary	Partial change	-Still focused on grammar and writing essays.
Kathy	No	-Provided more background knowledge with academic articles. -Did a lot of writings in class
Travis	Partial change	-Added more listening exercises -Moved away from genre writing to synthesize information.

Mary reverted back to teaching grammar and sentence structure in order to focus more attention on writing, so that her students could do well in the final tests, but she said that she did not teach many test-like activities in her class. Thus, the challenge remaining for the EAP program was the difficulty of putting the productive skill of speaking on par with writing. Teachers' accounts of giving different weightings to different skills confirm the views in the washback literature that when teaching towards the same test or skill, some teachers adopt more innovative and independent approaches as compared to a 'teaching to the test approach'. Also, the washback variability may not be so much the

test or test skill as the teacher him/herself (Alderson & Hamp Lyons, 1996; Spratt, 2005; Watanabe, 2004).

Andrews et al (2002) have also commented on the differing responses of the ‘implementers’ (the front-line teachers) and have suggested,

The predictability of the effects of a testing innovation (as of any other educational innovation) is crucially affected by the mediating role played by the teachers. Teachers in general may interpret the innovation in a way that differs from the intentions of the innovators, at the same time, there is likely to be considerable variation among teachers’ interpretations (p. 211).

Beeby’s second factor about implementation takes into account the amount of training that teachers receive towards the change. Since all teachers were part-time, whenever a professional development or training session was arranged by the administration it was difficult for all teachers to attend. However, as I explained in the ‘willingness to help’ section earlier, with regards to the mechanics of the new ISTs, the help was always readily available. Teachers pointed out these sessions lacked training for teaching and/or making the new integrated tests, especially the speaking component. As Alan remarked about the training provided for the new tests as “I think [training for] speaking will be the biggest fundamental change. It just doesn’t exist”. Jill pointed out another issue that “speaking is a softer skill because there is more subjectivity involved in marking and many of the qualities evaluated are not concrete, it is difficult to train for it”.

In terms of immediate washback of the ISTs, all teachers agreed that they were emulating more test-like tasks in their classes. Most teachers said they were incorporating academic skills, such as summaries and paraphrasing, in their weekly lesson plans. As mentioned earlier, Lisa and Jill were the only teachers who mentioned incorporating more speaking in their test preparation activities. Overall, because of the wider overlap between the test tasks and classroom activities, it is safer to assume a positive washback on teaching (Bachman & Palmer, 2010; Hughes, 2003; Messick, 1996). There seemed to be a far stronger link between the new ISTs and its effects on teaching. Alan commented:

Positive [washback]. For the first time, I think since I've done the graduating level, I have seen a strong emphasis on homework and taking the readings and listenings seriously. Because they [students] know that this is going to be material that's going to be on an test. Which is what university is like, so I say that's positive.

However, teachers' comments regarding the issues of standardization, rubrics, guidelines/templates, communication, and policies of the program may beg the question of whether this perception of positive washback is equitable among all teachers in the EAP program.

Learner factors. Students are the most important stakeholders in any high-stakes testing situation (Cheng, 2014; Fox, 2004). Anytime an educational innovation is introduced it affects the social, cognitive, and affective characteristics of students (Henrichsen, 1989, p.90). When investigating learner factors in the implementation process, Wall (2005) has suggested a number of characteristics of students. Prominent among these are students' attitudes: towards classroom teaching and tests, new ideas

(openness) and their goals. I added another category named ‘students’ background information’ to Wall’s categories because I felt that since all students in the EAP program were international students, it was important to know how much exposure they had to the English language before coming to Canada. Based on the factors just stated earlier, I analyzed students’ questionnaire responses and focus group data. In addition, I also looked for teachers’ comments about their students’ responses to the new tests.

Students’ backgrounds. Most students in the EAP program at the time of data collection were from Waters’ (2014) TSEP (tertiary, secondary and primary) ideologies (80% from China and 15% from the Middle East), which meant that the purpose of their previous learning was largely institutionally (educational system needs based) centered on didactic, content-based pedagogy (Waters, 2014). Most students shared the same L1 in their English classes and had very limited exposure to English language beyond their classrooms. This is evident in students’ questionnaire responses as displayed in Table 17 below. Around 57% (36/64) of students said that before coming to Canada, they did not use English outside their schools. Around 52% (33/64) students had not participated in any group work nor had given oral presentations in English. However, around 41% (26/64) percent of students said that they did write assignments, such as essays in English. Therefore, one of the goals of the EAP program was in line with many typical ESL programs, i.e., to increase communicative capability and emphasize social interaction (Fox et al, 2014). Table 17 summarizes the results of students’ questionnaire about their exposure to English outside and in their school context.

Table 17 *Student questionnaire results of English learning experiences in their home country*

<i>Overview of student Responses (n= 64)</i>					
Survey Questions <i>In your home country:</i>	Most of the time	Often	Sometimes	Seldom	Almost Never
- how often did you use English outside of school?	21%	11%	10%	31%	26%
- how often did you write assignments (e.g., essays) in English?	16%	25%	23%	20%	16%
- how often did you participate in group work and oral presentations in English?	8%	16%	25%	21%	31%

Most students entered the EAP program at the lower proficiency levels and by the time they reached the GL, they had a good idea about the test practices of the program. In the ‘relative advantage of innovation’ section above, I mentioned that 42% of students in the graduating level did not like the format of the new ISTs, and a majority among them (25 respondents out of total 64 students) were the students who were with the program for more than two semesters. Time spent in an English-dominant country and previous university experience can also be an indicator of students’ academic experiences.

Students’ attitudes towards classroom teaching and tests. One of the major strengths of the EAP program was the strong sense of community among teachers and students. One student Valdo said “one of the important things in this program is that our teachers try to motivate us. They worry about our results and try to boost our mood and spirit. It’s more like a psychological support. That’s important I think so”. Most students had positive attitudes toward classroom teaching and the feedback they received about their tests. Student Ali said that in terms of feedback on tests:

Apart from marks, we receive personal comment from our teacher and with their recommendations which kind of word, grammar or tense, is

appropriate in this case. And because it's very productive you can memorize it and in the future you can try to avoid the same mistakes. Then if you have difficulties in writing something or in understanding a particular topic, our teachers are always ready to give a hint or to maybe paraphrase this passage and the main idea of this word or passage in simple understandable words for us and I think it's very good when you have mutual teacher-student interaction. It's very good.

Similarly, when asked about understanding the new CLOs, (Table 18), 84% (54/64) of students affirmed that they understood the new learning outcomes. Valdo said, "It is very important to understand what is the academic study? It is the critical thinking. You have to create something new, but your ideas should be rooted in previous information article."

Table 18 *Student questionnaire results about understanding of learning outcomes*

<i>Overview of student Responses (n= 64)</i>					
Survey Questions	Definitely Yes	Yes	No opinion	No	Definitely no
Do you think that you understand the Course Learning Outcomes/Outcomes of GL?	23%	61%	15%	-	-

Also, what stood out in the responses gathered from student focus group was an indication of situated co-production of knowledge through classroom activities and the content of the final tests. Ali said, "I think when we are writing and we find the connection between the chapter and the reading resource it's like the time for preparation is better."

Finally, it was also evident from the student questionnaire responses that students had a good understanding of referencing and citations (Table 19). 78% (50/64) students

responded that they know how to cite different sources as compared to 10% (6/64) who mentioned that they did not know how to cite. In terms of learning academic skills, there, too, seemed to be a positive washback of the ISTs even when different teachers used different citation systems in their teaching.

Table 19 *Student questionnaire results about understanding of referencing and citation*

<i>Overview of student Responses (n= 64)</i>					
Survey Questions	Most of the time	Often	Sometimes	Seldom	Almost Never
Do you know how to cite sources (e.g., According to X.. or Davis-Floyd, R. (1998)... University Press. etc.) in your written assignments like a summary or paraphrasing?	40%	38%	12%	8%	2%
Do you usually know why you received a specific grade on a paper or test, for example because of good content, vocabulary, organization or other criteria?	21.5%	41%	24%	10%	4%

Similarly, 63% (40/64) of students said that they understood the rubric used in grading. When asked an open-ended question about the kinds of feedback that students received on their midterm test, many students reported ‘just a mark’, while others mentioned teachers commented on reviewing mistakes as Charlie said “apart from marks, we receive personal comments from our teachers with recommendations for grammar and vocabulary”.

Students’ goals. International students come to study in Canada in order to secure a degree. Within the EAP program, the key to getting into their chosen program of study is to pass the GL with good results in the final tests.

According to Fox et. al (2014), if the course providers understand the academic needs and strengths of their students, there is a potential for greater language support. Alignment of course activities with students’ goals increases the course impact. Similarly,

within the EAP program, administration and teachers were aware of students' need to continuing their study in Canada and tried to provide as much support as possible.

In summary, Fullan (2007) suggests, there are at least three components at stake in any implementation process: use of new or revised materials, new teaching approaches and possible alteration of beliefs. Within the EAP program, some of these were evident in teachers' and students' comments. Stakeholders suggested it was positive to see the EAP program moving away from "simply parroting a model in writing to students on their own creating ideas and communicating these through fluency of the language" (Alan, 2nd interview). Table 20 summarizes the characteristic of the users.

Table 20 *Summary of the characteristics of the users - teachers and students*

Within the User System	Manifestation	Facilitative or Inhibitive?
Communication Structure	Clear messages received from administration regarding 'what of innovation' (mechanics)	Facilitative
	Teachers' dissatisfaction with messages regarding 'how of innovation' (techniques)	Inhibitive
Education Philosophy	New ISTs were compatible with teachers' administrator' and the program's educational philosophies	Facilitative
Teacher Factors	Capabilities - High educational levels of all teachers	Facilitative
	Commitment- Teachers committed to ready to use the new ISTs	Facilitative
	Not much changes in teachers' pedagogical beliefs	Facilitative
	No test preparation activities for discrete skills	Facilitative
	Classroom tasks mirrored test tasks	Facilitative
	Less stress on teaching or testing speaking skill	Inhibitive
Learner Factors	Different weightage give to different skills (dominance of writing skill)	Inhibitive
	Students' educational background prior to coming to Canada	Inhibitive
	Students' positive attitudes towards the format of the new tests	Facilitative
	Classroom tasks mirrored test tasks	Facilitative
	Students' positive attitudes towards classroom tasks and learning	Facilitative
	Feedback received on test tasks	Facilitative
	Students' understanding of the new CLOs	Facilitative
	Students' understanding of the required academic skills	Facilitative
Students' motivation to achieve their goals of studying in Canadian universities	Facilitative	

Inter-Elemental Factors

The last evaluative criterion used for the evidence of washback in Phase 2 was *the inter-elemental factors*. The Inter-Elemental factors are the factors that "exist 'among' rather than 'within' the elements involved in the diffusion and implementation of innovations" (Henrichsen, 1989, p. 92). These factors are important for my discussion,

because in this section I bring together all the previously analyzed elements. Henrichsen suggests five different inter-elemental factors: Compatibility, Linkage, Reward, Proximity and Synergism.

Compatibility. Under *Compatibility*, Henrichsen points out two types of interactions: between *the Innovation and the intended User System*, and between *the Resource System and the intended User System*. To find out if an innovation is compatible with its users, Wall (2005) has suggested investigating four areas:

1. Can the resource demands of the innovation be met by the structures and facilities of the user system?
2. Does the use envisaged for the innovation fit in with the policies and regulations of the user system?
3. Can the task demands of the innovation be met by the abilities and behavior patterns of the individual?
4. Are the benefits that can be expected of the innovation in line with the attitudes and values of the individuals? (p. 275)

In the case of the EAP program, the answer to Question 1 was a partial 'yes'. The teachers and students was well responded to their needs by the administration. There was enough information provided to teachers regarding the CLOs, textbooks and even some samples of the new tests. However, some teachers lamented the skewed relationship between the management, teachers and the external Testing Office. Therefore, although there seemed to be a chance of successful implementation of the innovation, the roles of external testers and teachers did not create the necessary balance for its complete implementation.

The answer to Question 2 was also a 'yes'. The use of new ISTs, new PLOs and CLOs, and the textbooks were definitely compatible with the policies and regulations of the EAP program and the university in general. In terms of function, the new ISTs took

over the old ExitTest. The innovation was not jarring in the sense that it required either new procedures or policies, or new attitudes concerning what is to be taught or tested. The answer to Question 3 was again ‘yes’. It is clear from the analysis of the new ISTs, teachers’ and students’ abilities, and the teachers’ understanding were needed of the principles underlying the innovation or the techniques needed to develop students’ skills in the way that the innovators intended. However, teachers commented on the inadequacy of training for teaching and marking criteria for the newly introduced integrated skills. This seemed to present a major obstacle to successful implementation.

The answer to Question 4 is a definite ‘yes’. All teachers approved of the introduction of the new tests stating that the new tests were practical and they catered to students’ needs for learning academic skills. The innovation in this respect had a greater chance of successful implementation.

Linkage. The second type of inter-elemental factor is *Linkage*. Linkage refers to “the number, variety, and mutuality of contacts between the Resource System and the User System” (Havelock, as cited in Henrichsen, 1989, p.93). Henrichsen states that linkage reflects “the degree of inter-personal and intergroup connection that exists in a given situation” (p.93). In educational reforms, the important linkage factors are support networks such as professional teacher organizations, journals or other professional agencies. Although teachers did not speak about these subjects, the administrator, Anderson did say that the EAP program was accredited with ‘Languages Canada’ and most teachers in the program were member of organizations, such as ‘TESL Ontario.’ These links are important in the professional development of teachers.

Rewards. Henrichsen (1989) defines *Rewards* as “the frequency, immediacy, amount, mutuality of, planning, and structuring of positive reinforcements” (p.93). Rewards also refer to the need that “teachers have for some form of positive reinforcement, to convince them that they are doing the right thing and they should continue” (Wall, 2005, p. 277). The rewards could be in the form of monetary compensation, recognition from colleagues and/or appreciation from students. Students in the focus group were appreciative of their teachers and suggested that teachers helped them in every possible way whether inside or outside the classroom. Students were especially appreciative of the Teaching Assistants for their social and academic involvement with students. However, some teachers did complain that management was “not appreciative enough of teachers” (Mary) in dealing with change. Similarly, Stacey remarked that teachers had no say in whatsoever in the implementation of the ISTs. House (1981) calls this sad state of affairs “the teacher’s predicament”. House says the rewards for teachers who try innovations are few, and the personal costs are frequently high (as cited in Henrichsen, 1987, p.171).

Proximity. Another predictor of utilization of innovation is the Proximity, which is “the ‘nearness in time, place and context’ of the resource system and the user system (Havelock, 1969, p.20 as cited in Henrichsen, 1989, p.94). It means that if resources are readily available to users for the implementation of innovation, there are better chances of successful implementation. None of the teachers or students complained about any scarcity of resources at the EAP program. The classrooms were equipped with modern multimedia equipment and computer labs had the latest software for audio recordings.

Students had access to the University's library and the department's Resource Center for books and journals in English.

Synergism. The final inter-elemental factor is *Synergism*. Henrichsen (1989) claims synergism is "a working together" (p. 94). Henrichsen (1987) states, "when a variety of forces exert pressure together, in combination, upon the same point, the total effect can be greater than the sum of the parts" (p.171). It was difficult to find examples of synergism in the EAP data as many characteristics of the resource system and the user systems seemed to be working against each other.

In the new testing regime, although administration office had a clear vision for the program, with a lack of trust and a distance between teachers and that office, that vision was not what was getting practiced in the classrooms. The most important and direct aspects in any educational institute are the interactions between teachers and students. The other stakeholders, administrators and the Testing Office were holding much of the power and control did not have a direct influence on the program. The 'third solitude' within the program, and that mattered most, was the one in the classroom because that's where the strongest relationships were created. My analysis of the data pointed out that other outcomes (e.g., classroom assessment) did not have much bearing on students' final outcomes; everything was hinging on the final tests and that seemed to be "the problem." Teachers, in such scenarios, become simply functionaries rather than autonomous agents of teaching, but there was still independence of teachers because they were doing things on their own and despite management – not because of it – they were teaching students. The issue was one of a lack of alignment between teaching and testing and lot of that seemed stem from the disruptive influence of the centralized testing. It is useful,

however, to speculate on how some of the characteristics could have combined with others to enhance the potential for successful washback. An interesting example came up in a discussion about collaboration between the Testing Office and teachers. A teacher suggested that teachers supply a list of questions to the testing office, and from that, the office could select the final tests. This might have helped create a stronger connection between the classroom and the Testing Office.

Consequences

The final component of Henrichsen's framework is the Consequences. Although the focus of this chapter has been mainly on "*the Process*" or "implementation dynamics" in the EAP program (the '*how*' of the implementation of the ISTs), the analysis of the EAP program's effort would be incomplete without a discussion of the immediate washback at the end of first complete cycle of the implementation of the new testing regime. The consequences deal with whether the innovation was adopted or rejected by potential users and if this decision is confirmed or reversed with the passage of time.

In the case of the EAP program, the decision to adopt the innovation (the new ISTs) was not up to individual teachers; the management made an "authority decision" (Henrichsen, 1989, p. 94) that was imposed on teachers. The decision was made regardless of whether or not teachers liked, were ready for, or were in favour of the innovation. This type of technological perspective of innovation, claims House (1981) as well as Henrichsen (1989), generally produce the quickest rate of diffusion and implementation. Such changes, however, sometime can also provoke negative reactions from the users. Murray (2008) is of opinion that:

Often, top-down initiatives meet resistance from teachers; often teacher-led initiatives are appropriated by institutions for their own institutional ends, such as accountability. Sometimes institutions pay lip service to teacher knowledge by consulting teachers in reform initiatives but, as the initiatives are negotiated through the political system, the teachers' voice becomes muted (p. 5).

The administrators, teachers and students all spoke in favor of the new ISTs; however, the positive attitudes that administrators' and many teachers' attitudes professed to have did not always translate into the kind of attitude that was intended in the new testing regime. This is in confirmation of Alderson and Wall's (1993) notion that washback is a complex phenomenon, which should not be "seen as a natural or inevitable consequence of introducing a new examination onto an educational setting" (Wall, 2005, p. 279), and also Shohamy's (1993) advice, "using tests to solve educational problems is a simplistic approach to a complex problem" (p. 19).

Chapter Summary

This chapter presented the results and discussion of the second research question: What evidence is there of washback factors facilitating and/or impeding the implementation of the new testing regime? I used Henrichsen's (1989) four implementation factors as a model to inform my analysis and to structure my coding. These factors were: *the innovation itself*, *the resource system*, *the user system*, and *the inter-elemental factors*, which can facilitate or hinder the success of the diffusion of an innovation.

The *Innovation* (ISTs), in its entirety, seemed to have a potential for a positive washback. The content and format of the new ISTs definitely created a positive washback on teaching and learning. However, there were issues with its implementation, such as the comfort in using the new technology, and difficulty of teaching (and marking) fluency and accuracy in speaking skills. With respect to the *Resource and the User systems*, the fast pace of implementation created negative washback among teachers and students. This negative washback stemmed more from the ‘*how*’ of change rather than ‘*what*’ of change. There was a level of disorientation of change in terms of implementing speaking skills in both teaching and testing. Another potential source of negative washback was the lack of open communication and collaboration among three stakeholders: external testers, management and teachers. Many issues raised in the former testing regime, such as isolation of skill importance, devaluing of academic skills and variability of testing process, still persisted at the end of the first cycle of the implantation of the ISTs in the new testing regime.

Having examined the dynamics of the implementation process and immediate washback of the ISTs, in the next Chapter, I present the results and discussion from Phase 3 of my study. I examine the consequences of the delayed washback (after three semesters of implementation) of the new testing regime on teachers and students in the EAP program.

Chapter 6: Results and Discussion of Phase 3: the Consequences phase (*Washback of the New Testing Regime Over Time*)

Introduction

In this chapter, I present and discuss the results from Phase 3 (see Figure 3, Chapter 3). The purpose of this phase of the study was threefold: first, to explore the consequences of the delayed washback of the diffusion/implementation process of the new testing regime on teachers in the EAP program; second, to compare washback of the former and new testing regimes on students' accounts of their learning; third, to look for "evidential links between the teaching or learning outcomes and new tests' properties" (Messick, 1996, p. 247) that may have influenced the delayed washback of the ISTs in the program. These purposes are addressed through the study's third research question: *What evidence is there of washback in the new testing regime over time?* To answer this question, I used data collected from the following sources:

- four teacher interviews
- one administrator interview
- a focus group with students from both the old and new testing regime.

Although students from the former testing regime were not able to comment on the current testing regime because they had already passed the graduating level by the time changes were implemented in the program, their accounts of learning helped in comparing notes about washback of assessment practices in the EAP program and students' experiences with current university assessments. The purpose of the administrator interview was to gain an administrative perspective on the overall implementation of the new testing regime and its delayed washback on teaching and

learning in the EAP program. In order to understand the data, the three-cycle coding approach (Saldaña, 2013) was carried out as discussed in Chapter 3. Specifically, there was an open coding approach for the first and second cycle codes leading to core categories with conceptual memoing. Then, applying a top-down procedural approach and using my core categories from cycle one and two, I identified the themes, which are discussed in this chapter.

Research Question 3: What Evidence is There of Washback in the New Testing Regime Over Time?

Fullan (2015) suggests changes in three areas are necessary for an educational change to take place: “the possible use of new or revised *materials*, the possible use of new *teaching approaches*, and the possible alteration of *beliefs*” (p. 28). I kept these three in mind while analyzing the data from teacher interviews in Phase 3. First, the washback of the new testing regime on teachers will be presented and discussed, and later the washback of the EAP program’s assessment practices on students’ university studies will be presented and discussed.

Evidence of Washback on Teachers

Before discussing the consequences of the delayed washback of the new testing regime on teachers in Phase 3, I will first briefly summarize the findings about the washback on teachers from Phase 1 and Phase 2 of this study. In Phase 1 (Chapter 4), the positive washback of the ExitTest was evident in: making learning relevant to students’ needs, preparing students for standardized high-stakes tests, and using authentic academic learning outcomes in classroom assessments. However, negative washback was evident

in the form of skills isolation, the diminishing value of speaking and other academic skills, the dominance of the external testing agents, and a perceived need for change. Similarly, in Phase 2 (Chapter 5), positive washback of the new ISTs was evident in the form of more representation of test tasks in classroom activities, while negative washback occurred from not only the *Listening-to-Speaking* test, but also other factors, such as the rushed implementation of the ISTs and the power structure in the EAP program.

In Phase 3, I interviewed four teachers. While Stacey and Jill, my anchor teachers (see meta-analysis in Table 7, Chapter 3) had participated in all phases of the study, Derek and Ana taught in Phase 1 and Phase 3 during this study. Phase 3 results from the analysis of teacher interview data regarding the new testing regime are presented through three themes: teachers' accounts of teaching from Phase 1 to Phase 3, the integrated skills tests, and the Resource System (structure, openness, and harmony) of the EAP program. Each is discussed in turn.

Teachers' accounts of teaching from Phase 1 to Phase 3: As previously discussed, any successful educational innovation requires change on at least three levels (Cheng, 2005; Fullan, 2007, 2015; Markee, 1997; Wall, 2005):

- content or materials,
- methodological skills, and
- pedagogical values

Washback research on the influences of testing on teaching has provided evidence that language tests have a more direct washback effect on teaching content and/or methods than on teachers' beliefs (Cheng, 2005; Wall, 2005). That is why, to generate a more comprehensive picture of washback in the EAP program, it was important to

consider the content, methods, and attitudes of teachers in the new testing regime. The first aspect to consider was whether teachers reported any differences in their classroom practices because of the new tests.

In the EAP program, although teachers confirmed changing teaching content to match the new tests, none of the teachers reported any major change in their classroom teaching or methods as a result of the new tests (Table 21). This supports Alderson and Wall's (1993) Hypotheses # 1, 2, 4, 7 and 8 which are about a test influencing teaching and learning; what and how teachers teach; and the rate and sequence of teaching and learning (see Mechanism of Washback in Chapter 2). In fact, teachers (e.g. Stacey), expressed relief they did not have to suspend their regular classes before the midterm and final tests to prepare their students for the multiple-choice tests.

Derek mentioned that there was no change in his teaching; he had always based his lessons on course learning outcomes rather than on tests. Stacey, however, reported that "it allowed us to continue with the theme-based materials." According to Cheng (2014), "it is the teacher (who s/he is and what s/he brings as a teacher), rather than the testing, that decides how s/he teaches" (p. 6). In terms of change, this resonates with Fullan's (2015) *subjective meaning* of change regarding the potential use of new or revised materials for stakeholders who make use of innovation, and not the *objective reality* of change, where teachers in the EAP program did not change their beliefs about preparing their students for academic studies. Table 21 describes teachers' accounts of changes in their pedagogical practices from Phase 1 to Phase 3.

Table 21 *Teachers' accounts of teaching from Phase 1 to Phase 3*

Characteristics of teaching	Stacey	Jill	Derek	Ana
	<i>Taught in all phases</i>	<i>Taught in all phases</i>	<i>Taught in P1 and P3</i>	<i>Taught in P1 and P3</i>
Difference in teaching in 'delayed consequences' than the former testing regime	Did not suspend teaching for test preparation. Emphasis shifted to theme-based materials in text.	Focused on academic word list and peer-reviewed articles. Focused on summary writing.	No difference. Had always based lessons on course outlines. Always thought of students' needs in the first year of their university studies.	Practiced more short-answer and open-ended questions - not bullet points for MC questions.
Test preparation activities in class	Coordinated writing with listening and reading texts. Summary writing.	Focused on writing, but returned to grammar and sentence level teaching.	Reviewed key vocabulary and expressions from chapters selected for the final tests. Felt the ease of guiding students in terms of content.	No preparation for standardized tests.
Role of research-based projects and other classroom assessments	No evidence of engaging in extended research projects. Passed the responsibility to other team teachers.	Did not engage in research projects every semester. Depended on other team members.	Content based projects part of the course work. Extended essays (2 drafts + 1 presentation).	Had one project where students chose topic of their general interest instead of topics from the textbook.
Feedback provided for midterm and final tests	Detailed feedback on midterm writing tests.	Detailed feedback on midterm writing. Unlike finals, students see their midterm writing papers.	Feedback on content integration and synthesis. Stressed on the needs of students from midterm to the final tests.	Detailed individual feedback on in-class marks, but unsure of the final tests.

The other important theme that emerged in relation to test preparation activities in teachers' accounts was that the focus in the new testing regime seemed to shift from rote

learning to comprehension, thinking, and practicing for open-ended questions, rather than multiple choice questions (which characterized Phase 1). For example when asked about whether the test had made her change her teaching, Ana commented:

A lot of those tests [e.g., former ExitTest] often focused on discrete language items..... They weren't about the kinds of things that professors would assess them [students] on. So now, I do a lot of short-answer questions and have them focus on a whole short answer question, not bullet points.

While Ana stressed practicing short open-ended questions in preparation for the final tests, Stacey remarked that she incorporated writing activities with listening and reading texts in her lessons. Jill, meanwhile, stated that although she brought in more peer-review articles for students to practice with, she had to revert back to the teaching of grammar and sentence level errors because of the low language proficiency of her graduating level students. Derek pointed out that knowing the test chapters in advance helped in reviewing the key vocabularies and concepts for the final writing and speaking tests.

When asked about research projects, one of the new Course Learning Outcomes in the program, there were inconsistencies in teachers' responses. Teachers reported that they assigned research projects as classroom activities, but not for final grades. One possible reason for teachers not assigning a grade to the research project was because of its weighting in students' final outcomes. Even after three terms, teachers were not sure about how much weighting to give the research project as a percentage of overall marks. Only Derek mentioned that he might assign 20% of his post-midterm marks for the

research project, which came to roughly 6/100 marks, but ultimately, the teachers did engage in these final projects; however, no teacher ever assigned marks to these projects.

When asked about feedback on the midterm and final tests, all teachers confirmed that they gave detailed feedback only on the midterm tests, and this feedback was more concentrated on the writing tasks than on the speaking prompt. Also, teachers mentioned that while they were accountable for the assessment of in-class work and mid-term tests, the final tests were non-transparent to students because external raters marked these tests and these tests were not available to the teacher as Ana commented:

What my team agreed [regarding grade distribution] on is that each teacher would give feedback from our in-class marks because it doesn't make sense for me to say "well this is what you got in A's class" and they ask why and I don't know....But the final tests, all we get is a mark so I can't even indicate to them [my students] what they fundamentally did wrong or right. That's a challenge. I think if those marks came back to students with some notes that might be helpful because when you write a final test in other classes, you do know with virtual campus you get that kind of feedback.

This lack of transparency about final tests invoked a lack of trust among the teachers and was a potential source of negative washback on teachers. To promote positive washback and learning, Bailey (1996) has suggested the results of tests should be provided in a detailed fashion to both teachers and students.

The Integrated Skills Tests. In Phase 2, I discussed the main characteristics of the ISTs and the factors of the innovation that could be facilitative or inhibitive in the diffusion process (see Chapter 5, Table 14). I explored these factors again in Phase 3 to

see if they were exerting any washback after the innovation was firmly established in the program. Overall, teachers expressed mixed feelings with regards to their satisfaction with the new tests (Table 22). Generally teachers were positive about the new tests and felt that the ISTs had a positive influence on their teaching, but they also expressed discontent about a number of issues related to the new tests. Table 22 provides an overview of the teachers' responses towards the new testing regime.

Table 22 *Teachers' attitude towards the new testing regime*

Characteristics of the integrated skills tests	Stacey <i>Taught in all phases</i>	Jill <i>Taught in all phases</i>	Derek <i>Taught in P1 and P3</i>	Ana <i>Taught in P1 and P3</i>
Teacher's attitude towards the ISTs	Positive	Positive	Positive	Positive
Strengths of the new tests	Definitely useful in what students will need at university.	Closer to what students will do at university, but over importance given to summary writing.	Felt tests were fairly valid as related to students' needs in university.	More holistic learning focusing on complete integration.
Weaknesses of the new tests	Issues with the choice of questions and the wording of questions.	Topics distant from students' real lives. Complexity of the tasks. Interview questions. needed improvements.	One-size-fits-all approach for different streams. Relevancy of writing tasks only for social sciences and humanities courses.	Lack of resonance between what was thought to be tested versus actual testing Inconsistency in sources and testing.
Priority give to speaking skill and its testing	40% , but depended on the program students were going into. Arbitrariness/ Subjectivity of marking speaking.	May be 40% but speaking was not considered in the final student outcome.	Writing had more weight than other skills. Speaking may be 30%.	—

The first issue with the new test was the complexity of the content of the new tests: the new tests expected students to have a vast knowledge and understanding of

broad Canadian social realities. For example, Jill felt that it was too soon to expect international students to know and comment about complicated topics such as the Canadian health benefits or tax system. Another issue was with the choice of questions and the wording of test questions. Stacey said in the first semester of the implementation, teachers were consulted about the types of questions, but now they were “not in the loop on the choice of questions”. Furthermore, she said teachers did not get to see the tests beforehand, so they could not help in error corrections (grammatical or word choice) in the tests. Her other concern with the change in test format in Phase 3 was that the speaking interviews had a series of pictures for students to describe which she felt were more suitable for the lower levels of the program than the graduating level.

Ana, who had a PhD with a concentration on teaching and learning, was the most critical of the new tests. In addition to raising concerns about the quality of questions being asked, she mentioned that there seemed to be lack of resonance between what was thought to be tested versus what was actually being tested:

It’s an integrated task [the listening prompt in the test], but it’s a comprehension task and a discrete listening task, at that because there’s like a 7-minute listening and only the last 40 seconds was the answer. But I thought the listening was unfair because the students were sitting there for 7 minutes taking notes and the only thing that was relevant was the last 45 seconds. Then you’re asking them to summarize but that’s not what they’re doing. You’re asking them to basically answer a question; they’re integrating information to answer a question. The task doesn’t match.

She felt that there was no communication between the Testing Office, teachers, and students as to *'how'* tests and assessments were being created and how these tied into the program outcomes. The washback literature has consistently repeated that to create beneficial washback testers need to ensure that the test is known and understood by different stakeholders, such as teachers and students (Hughes, 1989; Morrow, 1991; Bailey, 1996).

Christopher, the administrator, on the other hand, commented that the Testing Office was benchmarking the new tests to provide exemplars for teachers.

The reason why we wanted to benchmark [with the Testing Office] is that when we introduce a new tool...our teachers are not all on the same page about marking it. We want still to have an external party to benchmark our marks. We did that. Now we have some exemplars done by the Testing Office and our next step is to have workshops with the teachers based on the exemplars.

However, when asked whether any validation study was conducted regarding the new tests, the administrator replied in the negative, and said, "The team of test developers were thinking of doing that sometime in the future". But he considered that the tests developed in-house were a better reflection of students' proficiencies than a TOEFL test, for example:

Based on the literature, about assessment and integrated skills testing, it's very popular with TOEFL who claims to have integrated tests...but in a very controlled manner of integration. The reading passage is usually a very short passage - I'm talking about 2 paragraphs. The listening passages are very short, and then there is usually some sort of relationship to the listening passage. They

[standardized proficiency tests] are very coaching prone, meaning that the teacher can prepare them. So it's not a real authentic, synthesis process. All they have to do is recognize what the listening passage is giving counter examples. It's very strategy oriented. That's not what we're doing here. We want to have the best of their understanding, that's what makes our tests so difficult.

Here, the administration seemed to have “a different value framework” (Wiliam, 1996, p. 134) for tests than the teachers. Wiliam (1996) suggests that there are differences between the inferential (e.g. what the test means to users) and consequential aspects of validity argument of tests and “the consequences of assessment... are always subjective, personal, and value-laden” (p. 134). Many of the test-related issues that the EAP program's teachers were discussing were considered to be technical questions by the administration. In fact these were the political issues stemming from different value assumptions i.e. Wiliam's consequences of assessment with pedagogical implications.

In terms of new skill testing (i.e., speaking) teachers mentioned that there were issues not only with the weighting given to it towards students' final outcomes, but also with marking it. Some teachers thought the weighting was 40/100 and others were of the opinion that it was 30/100. Stacey said the weighting was more arbitrary and subjective, “What grade do you give to the number of pronunciation errors? It's important to focus on the question being asked ...rather than speak something they've pre-prepared.” The lack of communication that was evident in the implementation phase regarding standardization, continued even after three semesters.

Finally, the most challenging issue with regard to assessing speaking was that it did not contribute towards students' final outcomes. Although the Testing Office still

made the final decisions regarding students' final scores on the ISTs, the oral interviews were not considered in the process. Jill lamented this:

They [the external testing experts] did not listen to the oral interviews; neither for the midterm, nor the final. So the teachers are the ones who are assessing the final and the midterm [speaking test].

The importance afforded to the weighting of different aspects of tests is often more political than technical (Wiliam, 1996). Wiliam, giving the example of the relative weighting used in the three English profile components in Key Stage 3 assessments in the National Curriculum of England and Wales, points out:

The importance of the change in weighting can only be appreciated when viewed in terms of *consequences* rather than *inferences* (emphasis added). Changing the weighting to 30:35:35 sends a clear message that Speaking and Listening are not as important as Reading and Writing (within-domain consequences), which presumably, will have an impact on the actions of teachers, students and others (beyond-domain consequences)..... those who disagree on the weighting do so because they have different value frameworks (p. 134).

Therefore, not having a standardized weighting assigned to tests other than the final test (e.g., teacher made different classroom-based assessments) proved to be problematic in terms of washback within the EAP program.

The characteristics of the resource system. One of the evaluative criteria used in analyzing the implementation process in Phase 2 was *the resource system* (also see Chapter 2). The factors that affect a resource system are: *structure*, *openness*, and *harmony*. In Phase 2, it was discovered that these three were the most inhibitive factors in

the change implementation process (see Table 15, Chapter 5). Therefore, in Phase 3, it was pertinent that I asked teachers about these factors to determine whether (if still persistent) these could be the potential sources of negative influence. Table 23 below is related to the characteristics of the Resource System in the EAP program. I asked teachers if they felt in control of the new tests. None of the teachers affirmed that they had any control over the new tests or the testing regime.

Table 23 *Teachers' accounts of the characteristics of the resource system in Phase 3*

Characteristics of the Resource System	Stacey <i>Taught in all phases</i>	Jill <i>Taught in all phases</i>	Derek <i>Taught in all P1 & P3</i>	Ana <i>Taught in P1 & P3</i>
Control over the new tests	Did not feel in control. Satisfied with the administration, but not the Testing Office.	Did not feel in control and felt students were affected by this feeling.	Felt in control for test task knowledge, but higher position given to 'outside body'.	Teacher powerless and administration powerful.
Support provided by the administration regarding new tests	Not much. The old rubric was recycled so she created her own rubric.	Arbitrariness of the marking rubric. Needed more workshops on both, teaching and marking integrated skills.	Did not need much support. Enough help was available in curriculum documents.	Did not need much support for research essays, had created her own rubric using testing guidelines.
Role of the Testing Office in the new testing regime	Little willingness than before to listen to teachers Would prefer doing away with pass/fail policy Arbitrariness of course grades continued.	Negative because when marking writing, testing office focused on accuracy and not on content.	Marking proficiency is not the same as marking academic preparedness and ability to synthesize.	Felt as a move away from restrictive assessments to <i>really</i> restrictive tests. Skills, not the content, should be tested.
Teachers say in students' final outcomes	Not much. Pass/fail is the policy that prevailed.	Nothing had changed as yet.	It was a negotiation, had to fight for students.	Just as before the change implementation.

In recent times, assessment-led reforms have been one of the most favoured strategies to promote teaching and learning (Hargreaves, Earl, and Schmidt, 2002). Classroom-based assessment approaches that go beyond paper and pencil tests, such as

performance- or portfolios-based approaches (Stiggins, 1999), are gaining more momentum. In the EAP program, there still seemed to be only one “correct” way to assess students. Similar to Phase 1 and Phase 2, the dominance of testing writing skills prevailed in Phase 3 and the vision of the Testing Office as ‘experts’ presumably was perceived by the teachers to entail the exclusion of teacher expertise. The administration invoked a policy that there was only one right way to assess; therefore, the Testing Office oversaw and ensured validity and reliability for students. As a result, teachers were – perceptually, at least – largely absent from the testing process. The implication was that moving forward with this administration strategy (of one way to assess students by a test) might be a source of potential negative washback. Ana’s frustration was evident in this long extract:

I think the changes we intended with our revised curriculum are all really good. I think the learning outcomes we have in place are spot on. If we want to work toward those learning outcomes, I think we are doing well to prepare students for success in university... The problem I have - there seems to be a little disconnect between the point of the learning outcomes and the assessment framework.... I think it’s artificial to say that there’s only one way of testing those skills, and that’s what we’ve done....And I don’t mind that my assessment gets objectively assessed...I don’t mind having to demonstrate this is measuring and what do you want me to measure,. and I don’t want to pigeon-hole my assessment to...*we’re kind of back to the TOEFL where we’re like “there’s only one way to assess your students”...well no there’s not.* Why do I have to make my students write for the 1-hour on a prompt created by someone else who doesn’t know what

we focused on And if I'm testing for proficiency, all I have to know is that they can clearly express their opinion and that they use the rhetorical skills we expect of them. I don't mind sending my prompt and all of my assessments up to the Testing Office and having an objective marker look at them. Someone else to review them so that we can make sure we're all hitting the right mark and I'm not passing students because I like them.... Even if it was the GL teachers that switch papers.

There were a lot of negative feelings coming from the teachers about the role of the external testing that was being done in the program especially at the graduating level. Derek said the EAP program was unique in this regard as he had never heard of this arrangement in other institutions. Teachers did all the testing and assessments. He said most students in the graduating level "sign up for IELTS in the middle of the session just as a back-up". He expressed his frustration as:

So if we have an office...to me I have a hard time articulating this relationship between us and our students. *How we are teachers and there is an office that we report to that can basically tell us that they didn't do a good job.* That's a bit of a strange arrangement. I can understand where it's coming from with standardized test and proficiency assessment. *And that kind of puts them in a higher position.* They know better what skills are expected. *Well, if they know better, why don't they teach those skills?* I'm sure double-blind marking has its benefits to it. We always double mark our papers anyway. I think I know who marked these papers and I think he or she has enough experience and understanding of what skills are expected of first year students. But what if there's another person who's hired and

they have less understanding than some of the teachers. What's the validity and reliability there?

Similarly, Stacey and Jill said they did not really feel in control of the tests. They observed that the testing regime seemed to drive the teaching rather than the reverse; as a result, teachers felt powerless and test held the power even when it was misguided.

Stacey commented about the system's "*willingness to help* and *willingness to listen*" as:

I don't really feel in control because although I said they do listen and appreciate my input and even if there were a number of other teachers who voiced the same view...sometimes it has already been decided and there's this...I don't know if I should say *pretense* of consultation. When things have already been *predetermined*. Look, I don't want to be overly critical. I think what we have is a very good program. It's a program that's responsive to the needs of our students and it's constantly trying to improve and change, so I'm very happy with the relationship that I have with the administration. Maybe a little less so with the Testing Office. Enough said.

The major repercussion of this, said Jill, was that students were noting the limitations teachers were under. She said:

The grading is done by the Testing Office. And the Testing Office, of course they are open-minded, and they listen to our recommendations, but I do not feel that there is a lot of control. I think the students are getting a sense of that too because I'll hear them say "Oh so we heard that the final writing test will actually decide whether I pass or not". And of course we tell the students that that's not the case

and it's for sure not completely, but to a certain extent unfortunately we know that that's the case. So I don't think that has changed.

In terms of support provided by the administration regarding the new tests, teachers' dissatisfaction was evident, but as the teachers whom I interviewed were all experienced and qualified they devised their own system of grading based on the existing rubrics, curriculum documents, and testing guidelines.

Teachers lamented that for all the effort of this program change, it seemed that administration and the Testing Office never really moved away from what they were doing before the change in testing regime. For example, Derek said as in the former testing regime, teachers' relationship with the Testing Office remained that of 'negotiation':

The results we received from the Testing Office are actually quite discouraging. We make our own judgement about our students... But according to the office it's more students who are actually not making it and that's what we'll have to negotiate with them because we're aware of our students' performance and their soft skills and ability to work in groups and engage in subject matter and grow. That to me is a big factor because... I understand about testing and reference, independent graders using a grid and sometimes with teachers who grow attached to students and we project their future successes onto them even though they haven't gotten there yet, we hope that they will. But it's a negotiation... We're going to have to fight for a few students.

In House's (1981) political perspective on innovation, power struggles dominate and cooperation is problematic for the stakeholders. Consensus is possible after negotiation of

interests, but requires open communication. Open communication seemed to be absent in the existing scenario in the EAP program, where the administration was perceived as being absorbed in matters of test development rather than in other issues capsuled around the test. This emphasis demonstrated and brought Wiliam's (1996) "different value implications" to the forefront i.e. consequences rather than inferences of testing. This gap was evident in the administrator, Christopher's, comments when asked how he perceived the communication among teachers, students and the external testing agents:

I think in our very first session that we implemented, the problem was the lack of communication on the part of our teachers with their students. So the students wouldn't know what kind of test they were going to get. I think it is terrible.... here, knowing the structure of the test is very important because it will help give them strategies for doing better on the test. Right now, what we do is, the very first week we have a meeting, I will ask the teachers to go over the structure of the midterm with their students so students are clear.....Teachers now know the test.... The Testing Office, we have meetings with them too, and they know, communication is taking place.

Thus, if the values (inferences and consequences of testing) "are not made explicit then any kind of meaningful debate is impossible" (Wiliam, 1996, p. 134). This was another potential source of negative washback on teachers.

There were other instances of negative washback evident in the execution of testing where teachers mentioned "hyper-marking". Teachers were marking their students harshly to ensure that their students were ready for whatever came. This included and the

potential hard marking from the Testing Office, which has no personal knowledge of the student and only focuses on the material being marked. Ana added:

...what I do, is I'm almost hyper-marking and I tell my students, I think I'm a hard marker... But I don't want you passing my course with a 70 because I don't know who's marking the final. If you get to the final and they're a harder marker, you fail, if they're slightly easier good for you. I'd rather you coming in at 70 and 72 consistently and meet the base requirement.

Other teachers, too, pointed to making students ready for “whatever comes”, i.e. the uncertainty of the final outcomes - especially for the borderline students.

My last question to the teachers was about the overall effect of the new tests on students' learning. In spite of the negative comments summarized above, all teachers unanimously agreed that the overall effect was positive. Stacey said, “Well I think positive in the sense that it's a more integrative approach and that it's tied in with what we're actually teaching them in the class...that's what we're testing them on.” Similarly, Derek was of the view, “I'm leaning more toward positive. We had been talking about it for years and years. And I was always one to mention misalignment. They [students] need to hone in on synthesis skills and application skills. The higher order skills. I think it has potential.” Jill was of the opinion that, “I think we brought the students one step closer to being prepared for university.” Thus, looking at the themes that arose from the analysis of teacher interview data, it can be concluded that the test format was definitely exerting a positive washback on teachers in the EAP program. In the next section, I will present the results of the analysis of the student focus group in Phase 3.

Evidence of Washback on Students

It was observed in Phase 2 that most students in the EAP program had a positive attitude towards the new ISTs. The results of the students' questionnaires and focus group suggested that there was an "evidential link" (Messick, 1996, p. 247) between the classroom activities and the new test tasks. It also seemed that the most important factor in any learning situation – motivation (Alderson and Wall, 1993) was high for these students.

I will now present and discuss the results of the delayed washback of the new testing regime on students. As mentioned earlier, the emphasis of this study was on *'the processes'* of washback rather than the *'products'* of washback. I explained in Chapter 3 that my ethics cover precluded accessing students' final grades. Therefore, the "products" in this study were mainly related to students' accounts of their learning and academic readiness in terms of English language proficiency. After coding the data, three final themes emerged from the focus group data related to the new testing regime: assessments students encounter in their university programs; students' academic readiness after completing the EAP program; and the relevance of the EAP program's assessments to their current needs at university.

Assessment encountered in university courses. All student participants in the focus group were from Science, Technology, Engineering and Mathematics (STEM) studies. Two students were in their masters' programs and four students were in different years of their undergraduate programs. Because their disciplines, levels and workloads were varied, their university experiences were different too. The first year undergraduate students compared their university workloads to their previous high school workloads

with an average of at least six courses. Table 24 shows the contrast between the assessment taking place within the EAP program and the university in general as suggested by the focus group students.

Table 24 *Comparison of assessments at the EAP program and university assessments*

Assessments in new testing regime in the EAP Program		Assessments at University <i>(as suggested by focus group participants)</i>
Reading/listening comprehension Writing Summaries and essays Oral interviews Optional Research projects		<u>MC questions</u> Lab reports <u>General tests</u> Reports <u>Research projects</u> Engineering problems extensive readings <u>Group projects</u> Writing computer programs Computer simulations Online assignments <u>Extended written responses</u> Developing project plans Analyzing different forms of data Online discussion boards Case studies

Students mentioned different types of assessment taking place in their university courses such as reports, case studies, online assignments, lab reports, general tests, and extended responses.

Assessment in the EAP program was more generic/universal. The assessments were a reflection of the idea that all students needed the same skills; this caused a mismatch for students entering diverse fields. Students suggested that some assessment at university like writing computer programs or simulations, analyzing different forms of data, and developing project plans were outside the domain of the EAP program. Other academic domain-specific assessments, such as report writing, group projects, case studies, and extensive readings could have been easily incorporated in the program's

assessment activities. As reflected in the assessment literature (e.g., Cheng & Fox, 2017; Hargreaves, Earl, & Schmidt, 2002; Stiggins, 1999), and in teacher Ana's comments, there is not just 'one correct way' to assess. The EAP program, in its current situation, appeared to have been too restrictive in its assessment practices.

In the new testing regime, tests were better aligned with what some of the students faced in their future studies and they could see a direct connection between, for example, group projects and writing reports in both the EAP program and their degree program courses. As Charlie, a graduate of the new testing regime, said: "I think the writing was useful, the summarizing was the most useful for my criminology course." The potential for positive washback for students, like Charlie, was much higher than other students, who were moving into more technically driven programs and did not write computer programs or undertake data analysis within the EAP program.

Academic readiness after completing the EAP. Another important aspect that I wanted to explore was the academic readiness of EAP students after completing the program. Students' views were varied, but interesting. They were able to shed light on both the positive and negative aspects of the role of the EAP program in their academic readiness. Most students felt confident about their academic readiness. Student Charlie (from the new testing regime) said he felt empowered because he was learning and thinking in English now:

I think *it's very magical* when you need to understand the knowledge in another language and it's not your mother language....It's difficult when you think Chinese and you're learning Math. But when you start Math in English it's like "oh my gosh."

At university, the transfer of knowledge may be based on forms of cross-language knowledge transfer, and Charlie's description of that extraordinary conceptual interchange as "*magical*" was very powerful. This was evidence of very positive washback as students were learning about the world through another's language and were able to bridge their knowledge in their native world to that knowledge constructed in a different one.

Another positive aspect of studying within the EAP program, as expressed by both former and new testing regime students, was learning about research-based assignments. As discussed in Phase 1, most teachers were concerned about the program learning outcomes and prepared students for university studies irrespective of the final tests. Some teachers included activities such as research projects in their classroom assessment. This was also evidence that the final test was a minor player in these students' accounts of washback because teachers were mediating students' experiences. (J. Fox, personal communication, 26 January 2018) commented that teachers play a mediating role in any test washback. Teachers often protect and shelter their students, to the best of their abilities, from test impact by providing students with a moderating influence that supports students' academic language development. Students in the EAP program saw exercises such as research projects, as valuable exercises that mirrored academic realities in university. Messick (1996b) has rightly pointed out:

...a poor test may be associated with positive effects and a good test with negative effects because of other things that are done or not done in the educational system. Technically speaking, such effects should not be viewed as test washback but rather as due to good or bad educational practices apart from the quality of the

test. Furthermore, a test might influence *what* is taught but not *how* it is taught, might influence *teacher* behaviors but not *learner* behaviors, or might influence both with little or no improvement in skills. Hence, washback is a consequence of testing that bears on validity only if it can be evidentially shown to be an effect of the test and not other forces operative on the educational scene (p. 2, emphasis added).

While discussing the research projects, students gave many examples of these activities in their university studies whether writing a chemistry lab report or writing a computer program. However, undergraduate students (from Phase 2 who prepared for the ISTs in their classes) pointed out that in many courses, such as in Organic Chemistry or Math, they did not get many assignments. In these courses students were assessed through midterm and final tests. As a result, there was not much scope for them to use their research skills learned in the EAP program. Similarly, in courses such as electronics and chemistry, the assessments were more hands on than in written formats.

I asked students about the type of assessment activities in their university courses within which thought they were most successful. All students agreed on essay questions as these demonstrated knowledge retention. Sheng, a student from Phase 1 said, “they show how much you understood. And they [teachers] would not take out the mark from you and delete like that with MC; there are always partial marks”. Sheikha, who was a graduate student, reiterated that the most useful assignment in her EAP studies were the research-based projects since she was expected to produce these in her undergraduate program. This was, again, evidence of teachers providing academic preparation in their classrooms.

When students were asked about the assessments that they thought were the most challenging in their university studies, students responded by saying that the ones that required either soft skills like group, or pair work assignments, or that needed technical knowledge of the subject matter. For example, Sheikha said she struggled with group work because the marks depended on ‘the whole assignment’ and sometimes there were delays because of group members. Similarly, for Valdo, the most useful assignment was a ‘self-made assignment’ where doing independent work was more important than team-based assignments. On the other hand, for Johnny, the issues were with assignments that had unclear guidelines that a professor provided to complete an assignment. Mohammed pointed out that assignments that relied on a higher level of English understanding were the most difficult, especially when he had earlier studied the same subject in his native language:

...because of English, my weak understanding of some terms used, not concepts, I had a hard time understanding terms and translating them in my mind to words I know in my home language. Some courses really expect a certain language level, which you cannot obtain by studying the course. Actually, you need to study the course and the English language course, or something. It depends on the terms you didn’t see before... Those who use the language so much to describe what’s happening, like politics or economics... I find difficult.

Within the EAP program, it seemed that English was taught in a more macro or global way, with little regard for specifically disciplinary academic English (Fox & Artemeva, 2017). That seemed to be the gap that Mohammed wanted the EAP program to address.

Students also pointed out some of the drawbacks of studying within the EAP program. Few students, after completing the graduating level, had to take one or two PEEL courses, and this slowed their academic progress. As mentioned earlier, the EAP program had learning outcomes, but students had one goal— to pass the test and enter their university program. For example, Mohammed, who had studied in Phase 1 lamented that because of two PEEL courses (offered in two different terms), he had to take one less course related to his major for those two terms and this was causing a considerable delay in completing his undergraduate degree. Similarly, Sheng bemoaned the fact that even after spending a considerable amount of time within the EAP program, his language skills did not improve as expected, especially his writing, as written assignments in university demanded a great amount of accuracy in grammar. He commented:

...in the technical reports writing courses for engineers, they [teachers] are expecting a very good language. The ESL language you teach, let's say you're writing the same essay that you've written in ESL [EAP program] but you're writing the report and it's the same level of English. Let's say in the ESL [EAP program] you'll take 70, but in the college report you'll take 50. So you need really to write with the least grammatical mistakes.

Relevance of the EAP tests in university studies. As illustrated in Table 25, the current EAP assessments were falling short in representing the demands of university assessment. When asked about the relevance of the EAP program's assessment practices to students' university studies, four themes emerged from the data: effects of skill isolation and the dominance of writing skills, the generic nature of EAP assessments, combining English fluency with disciplinary knowledge, and the ESL versus EAP nature

of the program. While the first two themes were potential sources of negative washback, the last two were more about aligning global English fluency with particular disciplinary use of English and gaining strong grammatical understanding along with writing fluency. Both of these latter themes were strong, potential sources of positive washback.

Students from both Phase 1 and Phase 2 indicated mainly two negative influences of the EAP program's testing practices. First was the skill isolation and the dominant role of writing skills in the program (recurrent themes in both Phase 1 and Phase 2), and the second was the universal assignments in the EAP program. As discussed in Chapter 2, the language testing literature has suggested many constraints related to assessing academic English (Elder, 2017; Weigle & Malone, 2016). Different skill sets are required for different assignments in different disciplines. Similarly, the focus group students expressed their inadequacy in tackling assignments that required extensive reading and critical thinking skills. For example, talking about reading as a critical skill, both Charlie (student in Phase 2) and Mohammed (student in Phase 1) said that they were not prepared for the amount of extensive reading that they had to carry out for their courses. There seemed to be a huge gap in reading volumes and skill between the EAP program and mainstream university as evident in Mohammed's comments:

I think it's the length of the reading. In one chapter in EAP program on a certain topic... you'll read 2-4 pages. But even in one lecture in any course, you'll read 50 pages...you're expected to read 50 pages for every lecture if you want to learn properly..... So you're expecting 150 pages for only one course in one week and that's really a lot. I think we can't compare. But I felt I should have taken certain

courses outside the university [EAP] that would teach me to how to read really fast and in different subjects not just English.

Similarly, pointing out the differences between what was learnt in the EAP writing class and the university writing assignments, Charlie said that the most important thing that he learnt in the EAP academic writing was “you need to delete your emotion and judgment [objective views] but in the PEEL courses we need to think about our audience and get their attention so we need more emotions [subjective views].”

While the EAP program’s constant focus was on writing, focus group students saw teaching of specific reading strategies and techniques as an EAP detriment. Students highlighted a number of academic skills that were needed to be successful at the university. Sheng said he faced difficulty in skimming and scanning, for example, to read through examples on a topic to get at a balanced common point that needed inductive and deductive reasoning. Similarly, Mohammed pointed out that as language was used differently for different purposes, he needed academic skills for understanding discipline-specific knowledge. Many studies (e.g. Cumming, 2014) have also pointed out that inferential understanding or reading between the lines for implied meaning can help students deepen the understanding of ‘English as written’. Mohammed asserted that this was something he did not learn during the EAP program. Absence of teaching these skills was not considered negative per se, but the inclusion of such skills would have deepened the positive washback of the EAP program.

Pointing out the shortcomings of assessment based on a single test, Johnny, a Phase 1 student, said that he considered EAP tests as any other proficiency test, but:

...the preparation of students for their future needs in a university degree is more dependent on the continual and gradual training they receive in class than on a single test.

The second theme that emerged was the relevance of EAP assessment in relation to university studies. The focus group students discussed the generic nature of assignments in the EAP program. They commented that the assessment in the EAP (one standardized test) assumed that all students needed the same skills, but this created a mismatch for students entering diverse fields. Furthermore, even though not explicitly stated (but observed in teachers' accounts in Phase 1, Phase 2 and Phase 3), students picked on the lack of standardization and different assessment practices of different teachers in the program. This seemed to be another potential source of negative washback. The grading and task inconsistency among teachers had dramatic consequences for students. For example, students said that within the graduating level, teachers focused differently on different academic skills, and/or rigor of grading, and that this created a tiered effect of skills within the program itself. To some extent, this also meant that whether or not students learnt certain skills was based on who their teacher was in the EAP program. For example, Charlie, giving an example of his friends in a PEEL course, said "while writing a summary... the full marks is 20 and I got 19 because I learnt it well at EAP, but my friend, who did not learn summary... he just got about 9.5."

The third issue raised with regards to the relevance of EAP program's assessment practices to students' university studies was the scarcity of combining English fluency knowledge with disciplinary knowledge (see Fox & Artemeva, 2017). This was also

evident in teachers, such as Derek's comments earlier about the importance of teaching and testing English for Specific Purposes (ESP). Students said they struggled to combine the two and suggested de-centralizing ESL/EAP teaching to the faculties where there was an ESL center to help students and also to match skills training to skills needed, as evident in Mohammed's comments:

...it's kind of side note, but it's kind of a dream but maybe applicable that every faculty has its own ESL and professors will get to teach these people [students] specifically for sciences.

The idea of team teaching using an EAP teacher and a discipline-specific teacher would deepen the positive washback between the needs of the university and how and what the EAP program teaches.

The final theme, which emerged in the students' focus group, was more theoretical in nature i.e., what is the EAP program's philosophy: An English as a Second Language or English for Academic Proposes? The students felt that the EAP program was more focused on casual English e.g., conversational, rather than academic English. The students wanted more academic English to help them prepare for the university and the academic environment. Students highlighted this gap in principles between students' needs and what the program was focused on - preparing students for social life versus university life. Mohammed iterated this point as:

As I told you I dropped the economics course just because of the English. But what I found is that EAP was more getting you ready for the society. I found myself more successful in the society and how to talk with people because of EAP more than what I found the benefit in the courses themselves.

In addition to this gap in principles, students lamented that many skills needed in academic studies were not being included in the EAP program, and they were left to learn these skills on their own. The level of learning in the EAP program compared to the university classroom was very different according to Sheng:

I had to read a lot... totally different from what I learned in EAP. In the EAP, I learned common words and grammar uses, but in the academic year I have to remember very hard vocabulary. In EAP with my language I would get 70 or 80% but in English [PEEL courses] I just got, maybe 60 or 50... Not friendly to international students.

Cheng and Fox (2017) suggest that learning outcomes are about what students should know and be able to do at the end of a course as a result of classroom activities such as assignments, feedback and tests. The EAP program's new Program Learning Outcomes (PLOs) and Course Learning Outcomes (CLOs) considered the program to be an academic preparation program. When intended learning outcomes, tests and curricular practices are aligned there is a potential for positive washback (Green, 2007).

Unfortunately, in the EAP program, these PLOs and CLOs were not formally measured. Students had pointed out that what was aligning between students' experiences in the EAP program and academy was in the form of research projects. There would have been a greater potential for positive washback in the new testing regime if PLOs and CLOs were aligned with testing because what is tested is what students value (Shohamy, 1993).

The tension between students' wants and needs not being met was another potential source for negative washback. As evident in students' data presented earlier,

students had very specific academic needs regarding the university and, unfortunately, the EAP program did not appear to fulfill all their needs.

In conclusion, when discussing how to promote beneficial washback, Bailey (1996) has suggested four major factors: a) incorporate language learning goals in assessment, b) incorporate authenticity, c) develop learner autonomy and self-assessment, and d) provide detailed score reporting. While the EAP program's new testing regime successfully incorporated the first two of the above factors, it definitely fell short of the last two. The new regime did not promote learning goals and teachers' classroom assessment practices, and the final test provided neither timely nor any feedback to students (see Ana's comments earlier). This also seemed to be the main cause of 'conflicts and resolutions' (House, 1981) among different groups (teachers, administration and external testers) in the EAP program. In House's (1981) political perspective, this type of change, where the focal point is entrenched in power and authority, considers teachers to be a passive recipient of change. Shohamy (1993) discussing the frustration of teachers in similar testing contexts, notes that:

In centralized, "authoritative" educational systems, tests become the major device through which the leadership communicates educational priorities to teachers. The teachers, on the other hand, are reduced to simply "following orders"; they are often frustrated by this role, because their responsibility increases while their authority is taken away (p. 17).

Similarly, teachers in the EAP program have voiced similar concerns that they did not feel in control of the ISTs in the new testing regime.

Previous literature has also suggested (Bachman & Palmer, 1996, 2010; Bailey, 1996) that washback is not only limited to the preparation for taking a test, but also on the decisions taken on the basis of test scores. The disproportionate weighting was another evidence of negative washback in the EAP program. Shohamy (1993) gives valuable advice in such situations:

Tests used for the purpose of improving learning can be effective only if they are connected to the educational system; they are not effective when used in isolation. But using tests to solve educational problems is a simplistic approach to a complex problem. It works on people's fear of authority. It can even be said that the testers themselves are abused by the educational leadership. Testers need to examine the uses that are made of the instruments they so innocently construct (p. 19).

In the EAP program, the disconnection between the final tests (in all three phases) and what went on in classrooms was problematic. For all the efforts of program change, the administration and the Testing Office never really moved too far away from what they were doing before. However, Christopher, the administrator said the program was open to further changes. In the meantime, to conclude, it is safe to say that the changes in the EAP program were not an event, but a process (Fullan, 2007, 2015) that will evolve in further iterative cycles as Shohamy (1993) stresses:

...the role of testers does not end in the development phase of the language tests they employ. Rather, testers need to follow the uses of these tests and examine issues of utility, relevance, ethics, and interpretation (p. 1).

Chapter Summary

In this chapter, I presented the results of Phase 3 of my study by answering the research question: What evidence is there of washback in the new testing regime? The findings suggest that there were not any major changes in teachers' classroom practices because of the new ISTs. Teachers continued to carry out most of their classroom lessons as reported in Phase 1 except that, now, the focus of writing skills had shifted to summary writing instead of five-paragraph essays. The new ISTs in this area had exerted a positive washback on teaching. The newly introduced tests of speaking and research projects were not included in the final outcomes for students. Hence, the washback of testing speaking and research projects in the new testing regime was neutral. Further, teachers suggested that the dominant role of the external testing experts inhibited the potential positive washback of the new ISTs.

The positive washback of EAP teachers' classroom assessment were evident in students' comments regarding their academic readiness in terms of projects and oral presentations. Students, from both the former and new testing regimes, reported that the assessments in the EAP program did not cover the breadth of assessments that students encountered in their academic studies in university. The dominant role of writing skills in the EAP program had its own negative repercussions on the development of students' other academic skills (evidence of negative washback). As well, students reported negative washback in terms of a single test as opposed to continual and gradual preparation for their future needs at university.

In the concluding chapter, which follows, I address the overall purpose of the study and summarize the key findings from the three research questions. I pay particular

attention to stakeholder accounts for this. This will be followed by a discussion of the limitations, implications and contribution of this study: first to the EAP program under study, and then to the understanding of washback literature. Finally, recommendations for future research will conclude the chapter.

Chapter 7: Conclusion, Implications, and Future Research Directions

Introduction

This concluding chapter begins by addressing the overall purpose of the study in light of the findings of all three phases. The purpose of this study was to investigate whether or not a change in testing regime, by introducing a new high-stakes test as innovation, can be used to leverage positive washback on teaching and learning in an EAP program. A summary of the three phases namely: the *Antecedents phase*, the *Process phase* and the *Consequences phase*, and the limitations of the study is presented. This is followed by the implications of this investigation with regards to potential for washback (positive and/or negative), first to the EAP program under study, and then to the language testing literature of washback. The chapter concludes with suggestions for future research.

Stakeholder Accounts and Washback

This study investigated the working and nature of washback as a phenomenon of change, using the models of washback and diffusion of innovation (e.g. Hughes, 1993; Henrichsen, 1989), in teaching and learning in an EAP program. These models were also helpful in highlighting some of the contextual factors, which mediated the implementation process of the new testing regime. Different stakeholders including the EAP program's teachers, students, and administrators defined these factors in the light of positive or negative washback from their own points of view according to how the new tests served their own purposes and uses. Taken together, these stakeholders' reports provided a clearer view of the complexity, direction (positive and negative) and intensity

(strong or weak) of the washback phenomenon as change in the EAP program. There was a perception that the former testing regime was exerting negative washback on teaching and learning, and that the new testing regime, with integrated skills tests, would instead leverage positive washback with increased quality of teaching and learning in the program. However, the new testing regime was characterized by mixed washback based on the accounts of the teachers and students who participated in this study. I will discuss this in the following section.

In Phase 1 the former testing regime showed that there was already evidence of positive washback of the final test, particularly, in the form of preparation of students for high-stakes testing (see Chapter 4 for details). Clearly, students felt that teachers were preparing them during classes for the high-stakes proficiency test. The negative washback during Phase 1 was not rooted in the high-stakes test per se, but in programmatic tensions such as the high value placed on writing skills as compared to the relatively low value of speaking skills and other academic skills. This negative washback was not directly aligned with high-stakes testing, so the former testing regime, described in Phase 1 could be viewed as having relatively neutral washback on teaching and learning. If anything, it could be said that in Phase 1, with regard to high-stakes testing had created a strong presupposition on the part of administrators and teachers that if there were testing regime changes in high-stakes testing, they would create an increased positive washback on teaching and learning.

Then, in Phase 2, the issue of “disorientation” arose, not from the “what” of change, but from the “how” of change. This uncertainty opened up possibilities of accepting change, but in reality was part of the reason why change was slow to be

adopted and teachers were resistant to it early on. The influence of high-stakes testing was off to a shaky start with regard to any leveraging of positive washback (See Chapter 5 for more details of implementation dynamics). This was a pivotal time for the adoption of change, so there needed to be more understanding by the stakeholders in order to be confident and “buy into” the changes that were coming. If this were more evident, a positive testing regime change was both more likely and more easily able to occur.

In Phase 3, some broadly noted issues arose. In particular, with regard to high-stakes testing and its washback, there did appear to be a “braking” effect of this leveraging as a result of the influence of the external testing experts. The lack of recognition of the overall resource systems by people outside of the teaching-learning relationship in the classroom seems to have constrained the influence of any positive influence of the high-stakes testing (See Chapter 6 for details). Also, while the EAP administration saw the change process as “systematic and rational” (House, 1981) from the teachers’ points of view, it was much more complex, messy and snarled (Fullan, 2007, 2015; Markee, 1997). This confirmed the notion that the washback phenomenon in this context was made up of many interrelated factors such as power relationships, the high-stakes nature of the final EAP test, and different stakeholder’s involvement. Together, these factors increased the complexity of the washback and undermined its potential to promote positive curricular innovation

What follows are more summary details of the results of this study to support this discussion about leveraging positive washback during a period of testing regime change and the efficacy of high-stakes testing within an EAP program.

Summary of Results

In Chapter 4, the research question explored was: *What evidence is there of washback from the former testing regime and what is the intended washback of the new testing regime?* Despite the communicative and student-centered teaching philosophies of the EAP program and its teachers, all stakeholders addressed negative washback of the former testing regime, citing two main reasons. The first reason related to the ExitTest itself, primarily because of the test format (MCQs and five-paragraph essays). Teachers disliked the skill dominance (i.e., of writing) and there was evidence in their accounts of construct under-representation (Messick, 1996) of speaking and other academic skills. The second reason related to the context of the EAP program i.e. the asymmetry between the testing processes and classroom assessment practices, which created disconnects between teachers and the external testing experts. For example, the ExitTest focused on decontextualized test content, but many teachers were preparing students for academic studies by giving contextualized assessments. For these reasons, both teachers and administrators proposed a testing regime change. Students reported negative washback of the former testing regime in the form of skill dominance and the undervaluing of speaking and other academic skills. At the same time, students also reported positive washback of the former testing regime, as it prepared them for the ExitTest and primed them for future standardized proficiency tests.

While analyzing the conditions in Phase 1, it was found that teachers and administrators in the EAP program perceived the former testing regime, i.e., the ExitTest, as falling short of preparing students for academic studies. They both seemed to point to high-stakes testing as the potential source of negative washback, and were ready for a

change. However, contrary to their points of view, this study's findings have revealed that:

a) In teachers' views the variability of the testing processes (discussed in Phase 1 and Phase 2 as part of the user and resource system) were the main sources of negative washback;

b) In students' views, the final test was a minor player in terms of test washback. Participating teachers in this study pointed out that the disconnection between the classroom assessment and the final test in the form of 'variability of testing processes' was the main source of negative washback. There were also tensions surrounding the role of the Testing Office (see EAP program's user system in Chapter 4 and its resource system in Chapters 5).

In Chapter 5, during the implementation of the new testing regime the research question explored was: *What evidence is there of washback factors facilitating and/or impeding the implementation of the new testing regime?* It was observed that the new tests were used as a "lever of change" (Pearson, 1988), and did bring positive influences to classroom practices. Teachers and students reported positive washback of the ISTs (i.e., the innovation itself), citing more integrated activities in their classroom instead of practicing MCQs and cloze tests for the final test preparation. This report from both teachers and students provided an "evidential link" (Messick, 1996, p. 247) between test tasks and classroom activities.

However, there was also evidence of negative washback, mainly in teachers' accounts. First, negative washback was related to the characteristics of the innovation itself, namely, the practicality of the new speaking test. Second, negative washback

stemmed from *the user and resource system* (i.e., the context of EAP). For example, as reported in Phase 1, variability in testing processes continued to exert negative washback even in the new testing regime. Furthermore, negative washback was evident in the Resource System (e.g. structure, openness and harmony in the EAP program). These factors inhibited the implementation of the new testing regime, and also the lack of synergism – an Inter-elemental factor (see Chapter 5 for details) certainly appeared to generate negative washback on teachers.

In Chapter 6, the research question that was explored was: *What evidence is there of washback in the new testing regime over time?* Teachers confirmed that the content/materials of the new tests (ISTs) were seen as positive influences on their teaching; at the same time, they pointed out several issues with the new tests. They complained about the validity of the tests citing a lack of resonance between what was thought to be tested versus actual testing. Similarly, they complained about reliability, as there were inconsistencies in sources and testing procedures. They also complained about the task-driven rather than construct-driven (Messick, 1996b) nature of the format of new tests. Teachers reported three kinds of negative washback of the new testing regime: First, they argued there was an imbalance in skill weighting (too much emphasis on writing vs. a lack of speaking) in the final outcomes for students. Second, they perceived a lack of standardization in the weighting of classroom-based assessments and the final learning outcomes. Finally, even after three semesters of implementation, variability in the testing procedure (reported in Phase 1 and Phase 2) continued, causing negative washback for teachers.

On the other hand, student participants who were well into their degree programs when the data were collected in Phase 3, reported that they were adequately academically ready after completing the EAP program. They noted that they had learnt many important academic skills in the program substantiating the earlier claims of Hughes (1993) and Bailey (1996) that the ultimate *product* of positive washback in any study is *improved learning*. In the case of the EAP program, the outcome was enhanced academic language development. Like their teachers, though, students lamented the restrictive assessments and generic nature of teaching English, instead of combining English fluency/proficiency development with disciplinary knowledge. The restrictive assessments thus seemed to be a potential source of negative washback for students.

The double layers of washback and the diffusion of innovation process over time in this study have shown not only the complexity of the phenomenon of washback, but also the intricacy of the implementation process of innovation. This confirms the notion that test washback is a complex phenomenon (Alderson & Wall, 1993).

The design of the new test (ISTs) definitely affected the teachers and students in the EAP program positively, but there were other meta-factors in the context that determined the actual washback for these individual stakeholders over time. Researchers (Cheng, 1997; Wall, 1999, 2005) have also suggested that contextual factors, such as the users and the user system itself interact with each other while changing over time. This makes making predictions of the same washback for all tests in all situations difficult.

What this study has found is that the test was a minor player in students' accounts of washback because of the mediating role of teachers in the program. In other words, teachers were guarding their students from the impact of the high-stakes testing by

providing them with a buffer that supported their academic language development. While going through the program, students (whether in Phase 1 or in Phase 2) did not account for any negative washback from teachers' classroom assessment practices. Most negative washback evident in both students' and teachers' accounts seemed to emanate from the end-of-the-term-final test where students, especially those who succeeded in the coursework, failed the course because of the final test. Therefore, innovations made in the program to support teaching and learning, through a testing regime change, did not shift the larger picture of test washback. The format of the ISTs was an important factor in its washback, but other factors such as the User System and the Resource System in the EAP program's context had an influence on different stakeholders over time. However, this influence was not in the same direction (positive, neutral or negative), or intensity (weak or strong) on all stakeholders or for different participants in the same stakeholder group. For example, in Phase 1 the ExitTest exerted a negative washback on teachers' accounts, but was seen as positive by most students. Such differing accounts are evident in other washback studies, e.g., Cheng (2005), and Andrews and Fullilove (1994). Shohamy et al. (1996) studied washback from two different language tests and pointed out:

...there is a discrepancy between the way teachers, students and bureaucrats view the effects of the test. The gap is mostly evident in the fact that unlike teachers and students, the bureaucrats portray a much more positive picture of the testing event and express satisfaction with the way the test is being administered within the educational system. Minor problems still exist, the inspectors admit, but the advantages of the test are more numerous than its disadvantages. The bureaucrats seem to use the test both

as a means to improve matters and as a device by which they control the system (pp. 313-314).

Similarly, in the EAP program, the administrators, teachers and students also had different accounts of washback.

Finally, this case study's findings indicate that, yes, a high-stakes test could be used to leverage positive washback in an EAP program if all the conditions in the EAP program's context were conducive to it. These conditions were related to the characteristics of: the innovation itself – the complexity, explicitness, practicality and relative advantage of the ISTs; the Resource System – centralized testing, openness and harmony of the new testing regime; the User System – communication structure, EAP program's and its teachers' educational philosophies, and the teachers and students in the program. These factors interacted with each other and definitely changed over time. Shohamy (1993) has rightly pointed out “testing is not an isolated event; rather, it is connected to a whole set of variables that interact in the educational process” (p. 4). Examining washback through innovation theory, particularly Henrichsen's Hybrid model, has pointed out many critical factors that may not have been evident in a cross-sectional study of washback as phenomenon of change. For example, looking at the delayed washback over time helped to clarify what was happening as a result of the change in testing regime i.e. if the negative washback (presumably, that was why the change was proposed) was because of the ExitTest or the other sociocultural factors (e.g. the resource system) in the EAP program.

Alderson and Wall (1996), in the special edition of *Language Testing* on washback, remarked many years ago that “any model of washback must include insights

from the theory of educational innovation, to help to explain why tests do not always have the effect that we desire or fear they will have” (p. 240). Insights from theories of educational innovation have helped the present study to establish a diminishing role of high-stakes testing in teaching and learning in the EAP program. If there were no empirical data collected in Phase 1 and Phase 2 under the two test regimes (former and new) to examine the nuanced differences among them, it would have been superficial to say that the previous ExitTest exerted a negative washback and it had more power than it actually did. However, students’ accounts in pre- and post- implementation suggested that the final tests (in both testing regimes) were not a major player in test washback and teachers played the important role of mediators within the EAP program.

Finally, questioning the role of introducing new tests as educational innovation, to engineer positive washback, Hamp-Lyons (2016) states:

...the debate over whether it is desirable or possible to design and implement tests specifically with the purpose of creating beneficial washback onto teaching and learning is one that still has some way to run..... we must question whether it would be of any value to make beneficial washback itself a test purpose (p. 19).

Her question is valuable in the present study where policy makers in the EAP program introduced a high-stakes test to promote curricular innovation and improve the quality of teaching and learning. The results confirm that the context (i.e., the testing regime as a procedural system) in which the test resided was equally responsible for test washback. Tests used as mediators to promote desirable innovation in learning over time are not necessarily always efficient, and they may not have the desired consequences as predicted (Cheng, 2014). In the end, I would like to reiterate Rea-Dickins and Scott’s (2007)

assertion that “washback can be viewed as a *context-specific, shifting process, unstable, involving changing behaviors* in ways, which are difficult to predict (p. 5).

Limitations of the Study

Although investigating washback in one EAP program has been fruitful, there are limitations that are to be considered while interpreting the results of this study. First, the washback and innovation literature have given ample empirical evidence that it takes longer than a year to document washback and collect evidence of tangible changes from any implementation process (Frederiksen & Collins, 1989; Fullan, 2015; Markee, 1997; Shohamy et al, 1996; Wall, 2005; Wall & Horak, 2008, 2011). In the present study, though longitudinal data were collected over a 20-month period, this time frame may still have been insufficient. Since the study was conducted just before and after testing regime changes were made, and only at one level of the EAP program, it was not possible to get a complete and comprehensive picture of washback to the EAP program as a whole within the time frame of this PhD research. Collecting data while further changes were to take place more gradually at the lower levels of the EAP program was beyond the scope of this research, but evidence of washback at other levels of the program would have greatly added to my understanding of its impact overall.

The second limitation of this study was my limited use of students’ accounts in addressing my research questions. Further, the focus group student participants in Phase 3 were all from Science, Technology, Engineering and Mathematics (STEM) studies, disciplines that are not necessarily indicative of other university programs. Also, it was not possible to administer the questionnaire from Phase 1 to students in the EAP program

at the end of the study. Had this been possible, it would have allowed for a useful comparison of overall student responses to the former and new testing regimes.

I relied mainly on teachers' accounts in interviews in this research study. All of the participants (teachers, students, and administrators) volunteered to share their insights, and although arguably they reflected a range of opinion, which was indicative of the general view of the testing regime change, it is important to acknowledge this limitation in interpreting the results of the study. Additional quantified measures of successful/unsuccessful implementation of innovation in the EAP program might have added to my understanding, but it was difficult to predict in advance which factors would facilitate or inhibit the success or failure in the adoption of innovation. Such measures would need to be based on "individual's statements and the patterns of evidence which support them" (Henrichsen, 1987, p. 362). Also, as stated earlier, the complexity of both the washback phenomenon and the diffusion/implementation process involved a myriad of variables, and only some can be measured quantitatively and objectively.

One of the strengths of this study was that I was able to look at washback over time primarily in relation to teachers. Because the student population within the EAP program was transient, I could only recruit volunteers to talk about their more immediate learning experiences within the EAP program. A washback study dedicated to more direct linkages between student test results and longer term learning legacies of those results would be important to further understand the effects of washback more deeply than I was afforded here.

Finally, the study was embedded in one EAP program, so the findings from this study cannot be generalized to other contexts. However, like many other qualitative

research findings though not generalizable, can still inform in meaningful ways other EAP programs which may see something useful in the findings of this study.

Implications and Contributions

I will present the implications and contributions of this study in three sections: the first relating to the EAP program directly, the second more external benefits, and the third as contributions to the relevant washback literature.

The EAP program. One of the purposes of this study was to investigate the nature and processes of washback, or the ways in which washback in the context of testing regime change occurred over time. Also explored was the potential to better understand said processes in order to make the process of implementation “more efficient in the future, thus leading to fuller and more fruitful outcomes” (Wall & Horak, 2011, p. 136). This study has not only produced evidence of positive and negative washback for teachers and students in the EAP program under study, but it has also led to several ideas of how the newly-implemented ISTs might have been used to produce more positive than negative washback. The first among these ideas was that a detailed baseline study prior to introducing the ISTs would have been useful. Such a study would have identified important characteristics of the user system (i.e., testing context of the EAP program) and users (i.e., teachers and students) to find out what kinds of changes would have been desirable and how those changes could have been attempted. Innovation theory also would have provided the EAP program administrators and test designers with a better understanding of the risks they were taking when they rapidly introduced testing regime change, as teachers were evidently displaced by the speed of the implementation process.

Secondly, the frontline implementers in this study, i.e., the teachers, suggested that not all stakeholders were included in the different stages of the change process, i.e., initiation, implementation, and institutionalization (Fullan, 2015). Fullan describes these as critical stages in an innovation when it is put into practice and attention is directed towards its sustainability. Teachers implied that they were superficially consulted in the initiation and implementation stages, but were left out of the institutionalization stage of the ISTs' implementation. As noted in Phase 3 of this study, this exclusion caused considerable negative washback on teachers. In my view, the administration should have considered all stakeholders' voices in the change process, including teachers, students, faculty representatives of different programs in the university, and external testers (Fullan, 2007, 2015; Markee, 1993, 1997; Wall, 2005).

Importantly, it might not have been enough for the test reform team to solicit the concerns of different stakeholders alone; the team should have also included a key member from each of the different stakeholder groups throughout to enable better cooperation, communication, and shared understanding of the goals of the change process. More thorough inclusion of key stakeholders would have ensured the criteria were met for judging assessment systems, such as validity, reliability, and practicality of the ISTs. Insight gained from such a process could have helped in scrutinizing the test specifications before they were adopted as assessment guidelines and/or test blueprints (Wall, 2005).

Teachers in the program suggested having standardized practical guidelines related to the final test instead of needing to constantly negotiate with external testing experts at the end of every semester (see variability of the testing process in Chapters 4,

5, and 6). Their suggestion was to give precedence to teachers' voices who, after spending 14 weeks with students, were in a better position to make a high-stakes judgment about their students' academic language proficiency and readiness to engage with academic work in degree programs. Sloane and Kelly (2003) have pointed out that without an active teachers' voice, "it is highly unlikely that tests of future will fully serve the dual goals of cognition and instruction, while, at the same time, responding to legitimate calls for accountability in the educational system" (p. 16).

Most importantly, the new ISTs were not stand-alone tests. The content of the textbooks and classwork were to be incorporated in the tests, so that the tests would not be judged in a decontextualized manner by outside experts as was done in the former testing regime in decontextualized five-paragraph essays. In the new testing regime, other voices (such as teachers) were essential to make judgments about students' language proficiency and academic readiness.

Other standardized proficiency tests (e.g., IELTS, TOEFL) assess all four language skills (speaking, listening, reading, and writing) and weight the scores across skills in arriving at an overall proficiency score. Judgment based on the two stand-alone writing tasks in the ISTs could not stand shoulder-to-shoulder with these other standardized high-stakes tests, thus creating for the ISTs "a case of invalidity of the testing protocol" (J. Fox, personal communication, February 2018). This 'protocol' ultimately contributed to negative washback on the program as a whole.

What's more, the curriculum/assessment documents in the program had not enough clarity regarding classroom-based assessments. The result, as reported by teachers and students, was inconsistency across different sections of the same level.

Finally, this study appeared to be one of the first studies to investigate the influences of the new ISTs on teaching and learning at one level in the EAP program. Thus, findings from this study could be useful in future iterations of test change at the other levels of the program. As stated earlier, washback is not a simplistic or direct phenomenon; therefore, some areas (e.g., validity and/reliability of the test) need further research, especially given the role of socio-cultural and other contextual factors in the EAP program. This and other future directions for research are discussed later in this chapter.

External Benefits. The direct benefit to other programs considering adoption of a new testing regime, or to programs which who have already started to engage in one, is to read the lessons learned from this EAP testing regime change experience. There is much to inform the thinking and decision making that goes into the planning and implementation of change. The results of this study reveal positive outcomes and potential pitfalls, both of which should be noted and understood by any program before attempting to proceed with a testing regime change.

Within the broader research of washback, this study will be of direct benefit to other researchers who are interested in washback and who want to better understand the complex and highly contextual nature of washback studies. They may want to draw on examples such as Cheng (1997, 2005), Henrichsen (1989), and Wall (2005) who looked at washback as a change phenomenon typical of reform agendas in curriculum.

Relevant Literature. The present study can contribute to both the general education and language testing literature as the study combined insights from both (e.g., Cheng, 2005; Henrichsen, 1989; Hughes, 1993; Markee, 1997; Wall, 1999, 2005). These

provided valuable insights into why certain attempts to introduce intended positive washback were not as effective as their implementers had hoped they would be. Similarly, an understanding of the change process (Fullan, 2015) also played a part in understanding the factors that were responsible for unintended negative washback from the test. The present study reiterates the importance of investigating contextual factors along with the investigation of washback from any revised or new test.

From an educational point of view, well-constructed tests are undoubtedly a prerequisite in improving educational standards. Equally important, however, are the roles of clearly defined standards, well-explained exemplars, adequate training of raters, and opportunities for different stakeholders, such as teachers and students, to receive meaningful feedback from assessment practices (Hughes, 1993; Bailey, 1999; Cheng, 1997; Wall, 2005, 2012).

This study also provided evidence that one test can have either positive or negative washback according to stakeholders. In other words, “the distinction between positive and negative could usefully be made only by referring to the audience” (Watanabe, 2004, p. 21). For example, while students may evaluate one type of outcome as positive, teachers may evaluate that same outcome as negative.

From a washback perspective, high-stakes tests –especially when used as an exit requirement in any EAP program– are a “differentiating ritual” (Bernstein et al., 1966, as cited in Wall, 2005). The results used for international students’ university placements impose a great deal of pressure on students to succeed, and to make high stakes decisions from one test is extremely limiting as general proficiency tests can be insensitive to

students' growing mastery of academic skills (Elder, 2017). Madaus (1989) has rightly suggested:

...it is time we began to work towards lowering the stakes associated with the test use. Test results can provide valuable information, but only one piece of information, and only when interpreted with wisdom, in conjunction with many other indicators and factors (p. 86).

Having one-size-fits-all standardized tests, especially in EAP contexts, can counteract the richness of these programs where a specialized curriculum can provide opportunities to assess EAP knowledge in different disciplines through integrated skills and a variety of topics and tasks (Schmitt & Hamp-Lyons, 2015).

It is my observation that the current washback literature has not explored the full potential of the connections between key theories of validity (e.g., Messick, 1989), principles of high-stakes testing (e.g., Madaus, 1989), and practices in the diffusion of innovations (e.g., Henrichsen, 1989) in a synergistic way. To my knowledge, only a few researchers (Cheng, 1997; Wall, 1997; Wall & Horak, 2006, 2011) have studied test washback by using the literatures of curricular innovation along with measurement and language testing literatures. There is thus a great deal of potential in drawing more comprehensively on these literatures to inform washback studies, which have traditionally concentrated mainly on the relationship between test properties, stakeholders (e.g., teachers and students) and teaching/learning process. Rigorously studied together, the above-stated literatures have a powerful potential for depicting a more stable (although dynamic) picture of washback pointing to specific contextual factors in play.

Recommendations for Future Directions

According to Schmitt and Hamp-Lyons (2015), there is a need for additional research in the “under-defined and under-theorized” field of EAP assessment. They suggest that so far there has been very little research “from *within* ‘live’ EAP programs” (p. 3, emphasis added). Cheng and Fox (2013) also identified a lack of innovation studies in EAP programs in their review of doctoral studies in Canada. Wall (2012) expressed frustration regarding the different directions research on washback has taken, even when researchers are working on similar types of questions. She laments the lost opportunity that comes from the fact that “it is rare in the field of washback research to truly build on previous work” (p. 88).

There are several potential future research directions stemming from the results of the current study. This project studied one level of an EAP program while the implementation process of a test reform at the lower levels was still underway. Therefore, this study could be considered a baseline study for other test-related studies in the EAP program, such as psychometric validation studies of high-stakes tests or social consequences of high-stakes testing in the EAP program. For example, one study could consider the new ISTs in terms of their construct validity and reliability. Alternatively, to add to the under-theorized area of EAP assessment (Schmitt & Hamp-Lyons, 2015), the classroom assessment practices of the EAP teachers could be further explored. Moreover, the program could benefit from the lessons learnt about the contextual factors, especially how the resource system and other inter-elemental factors facilitated or hindered the potential positive washback of the ISTs in the EAP program.

Another potential extension of this study could be to involve the testing experts and different faculty representatives in studying the impact of the macro-context on the EAP program's assessment practices. Studying washback through both micro- and macro-contexts would build up "a more complete and comprehensive picture of the washback operating therein (Wall, 2012, p. 89). It will also be helpful in studying the testing context through different methodologies: quantitative, qualitative, and mixed-methods (Cheng, 2014).

Finally, since the field of EAP assessment is under-researched (Fox, 2009; Schmitt & Hamp-Lyons, 2015), and currently there are more than 80 Pathways programs in Canada, another future ambitious project in relation to EAP assessment could be to focus on how these Pathways programs are evolving in terms of assessment practices, both internal-to-program classroom assessment and external-to-program high-stakes language proficiency testing. Mixed-methods research could yield evidence of washback from these different forms of assessment on teaching and learning in these programs.

To conclude, this research was conducted at one university EAP program and this study's implications and recommendations will hopefully have beneficial outcomes for the test practices of this program. The study pointed out many contextual factors that could contribute towards the positive washback of assessment innovation. Studying the washback effects of a testing regime change, while being a teacher in the same EAP program, has provided me time and space to reflect on my own practice as a teacher with regard to testing strategies to minimize negative influences, while increasing positive influences. This study will better prepare me by providing me with a greater understanding of test demands and how my fluency in assessment, generally, will play a

role in my effectively supporting future students with confidence. The study may, also, encourage the relevant authorities to develop and implement policies to enhance the satisfaction of the two most important stakeholders in the program: frontline teachers and, most critically, future students.

References

- Alderson, J. C. (2004). Foreword. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: research contexts and methods* (pp. ix–xii). Mahwah, NJ: Lawrence Erlbaum.
- Alderson, J.C. & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing*, 13(3), 280–97.
- Alderson, J. C. & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115–129.
- Alderson, J. C. & Wall, D. (1996). Editorial. *Language Testing*, 13(3), 239–240.
- Algozzine, B. & Hancock, D. R. (2006). *Doing case study research: A practical guide for beginning researchers*. New York: Teachers College.
- Andrews, S. (1995). Washback or washout? The relationship between exam reform and curriculum innovation in English language teaching. 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. (2004). Washback and curriculum innovation. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: research contexts and methods* (pp. 37–50.). Mahwah, NJ: Lawrence Erlbaum.
- Andrews, S. & Fullilove, J. (1994). Assessing spoken English in public examinations—Why and how? In J. Boyle & P. Falvey (Eds.), *English language testing in Hong Kong* (pp. 57–86). Hong Kong: The Chinese University Press.
- Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback—a case-study. *System*, 30(2), 207–223.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C.E. Turner & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–71). Ottawa, ON: University of Ottawa Press.
- Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.
- Bachman, L. F., & Palmer, A.S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.

- Bailey, K.M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13, 257–79.
- Bailey, K. M. (1999). *Washback in language testing*. Princeton, NJ: Educational Testing Service.
- Banerjee, J. & Wall, D. (2006). Assessing and reporting performances on pre-sessional EAP courses: Developing a final assessment checklist and investigating its validity. *Journal of English for Academic Purposes*, 5(1), 50-69.
- Baxter, P. & Jack, S. (2008). Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. *The Qualitative Report*, 13(4), 544-559. Retrieved from <http://nsuworks.nova.edu/tqr/vol13/iss4/2>
- Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher Education Research & Development*, 18(1), 57-75.
- Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74.
- Brindley, G. (2008) Educational reform and language testing. In E. Shohamy & N. H. Hornberger (Ed.), *Encyclopaedia of language and education, language testing and assessment* (2nd Ed.). (pp. 365-378.) New York: Springer.
- Broadfoot, P. M. (2005). Dark alleys and blind bends: Testing the language of learning. *Language Testing*, 22(2), 123-141.
- Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *Jalt Journal*, 10(1), 15-42.
- Burrows C. (2004). Washback in classroom-based assessment: A study of the washback effect in the Australian adult migrant English program. In L. Cheng & Y. Watanabe. (Eds.), *Washback in language testing* (pp. 113-128). Mahwah, NJ: Lawrence Erlbaum.
- Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, 1(1), 1-47.
- Casanave, C.P. (2015). Case Studies. In B. Paltridge, & A. Phakiti (Eds.). *Research methods in applied linguistics: A practical resource* (pp. 120-135). New York: Bloomsbury Publishing.
- Castillo-Montoya, M. (2016). Preparing for interview research: The interview protocol refinement framework. *The Qualitative Report*, 21(5), 811-831.
- Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing*, 14(1), 3-22.

- Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, 20(4), 369-383.
- Chalhoub-Deville, M. & Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (ed.), *Handbook of research in second language teaching and learning* (pp. 815-832). New York: Routledge.
- Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272.
- Chapelle, C.A., & Plakans, L. (2013). Assessment and testing: Overview. In C.A.Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 240-244). Oxford, England: Blackwell/Wiley.
- Chapman, D. W. & Snyder Jr, C. W. (2000). Can high stakes national testing improve instruction? Re-examining conventional wisdom. *International Journal of Educational Development*, 20(6), 457-474.
- Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education*, 11, 38-54.
- Cheng, L. (1998). Impact of a public English examination change on students' perceptions and attitudes towards their English learning. *Studies in Educational Evaluation*, 24(3), 279-301.
- Cheng, L. (2005). Changing language teaching through language testing: A washback study. In M. Milanovic & C. Weir (Series Eds.), *Studies in Language Testing* (Vol. 21). Cambridge, England: Cambridge University Press.
- Cheng, L. (2014). Consequences, Impact, and Washback. In J. Kunnan (Ed.), *The companion to language assessment*, Hoboken, NJ: Wiley-Blackwell. DOI: 10.1002/9781118411360.wbcla071.
- Cheng, L. (2018). Geopolitics of assessment. In J.I. Liantas (Ed.), *TESOL encyclopedia of English language teaching*. Hoboken, NJ: John Wiley. DOI: 10.1002/9781118784235.eelt0814
- Cheng, L. & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng & Y. Watanabe (Eds.), *Washback in language testing: research contexts and methods* (pp. 3-17). Mahwah, NJ: Lawrence Erlbaum.
- Cheng, L. & Fox, J. (2013). Review of doctoral research in language assessment in Canada (2006-2011). *Language Teaching*, 46(4), 518-544.
- Cheng, L., & Fox, J. (2017). *Assessment in the language classroom: Teachers supporting student learning*. London: Palgrave.

- 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(4), 436-470.
- Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing: Research contexts and methods*. Mahwah: Lawrence Erlbaum Associates.
- Cohen, A. D. (1994). *Assessing language ability in the classroom* (2nd ed.). New York: Heinle & Heinle.
- Crabtree, B. F. & Miller, W. L. (Eds.)(1999). *Doing qualitative research*. Thousand Oaks, CA: Sage.
- Creswell, J. W. (2012). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage.
- Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, 10(1), 5-43.
- Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, 10(1), 1-8.
- Cumming, A. (2014). Assessing integrated skills. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 216-229). Hoboken, NJ: John Wiley. DOI: 10.1002/9781118411360
- Damankesh, M. & Babaii, E. (2015). The washback effect of Iranian high school final examinations on students' test-taking and test-preparation strategies. *Studies in Educational Evaluation*, 45, 62-69.
- Davidson, F. & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CN: Yale University Press.
- Davies, A. (2007). Assessing academic English language proficiency: 40 years of UK language tests. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner & C. Doe (Eds.), *Language testing reconsidered* (pp.73-86). Ottawa, ON: University of Ottawa Press.
- DeMarrais, K. (2004). Qualitative interview studies: Learning through experience. *Foundations for research: Methods of inquiry in education and the social sciences*, 1(1), 51-68.

- Denzin, N.K. & Lincoln, Y.S. (2011). *The sage handbook of qualitative research* (4th ed.). Thousand Oaks, CA: Sage.
- Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford, England: Oxford University Press.
- Douglas, D. (2013). ESP and assessment. In B. Paltridge & S. Starfield (Eds.) *The handbook of English for specific purposes* (pp. 367-383). Hoboken, NJ: John Wiley.
- Duff, P. A. (2012). Case Study. In, C.A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. DOI: 10.1002/9781405198431.wbeal0121
- Elder, C. (2017). Language assessment in higher education. In E. Shohamy & S. May (Eds.) *Language testing and assessment: Encyclopedia of language and education* (3rd ed.). (pp.271-286). Cham, SW: Springer.
- Ferman, I. (2004). The washback of an EFL national oral matriculation test to teaching and learning. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contents and methods* (pp. 190-210). Mahwah, NJ: Lawrence Erlbaum.
- Ferris, D. R., & Hedgcock, J. (2013). *Teaching L2 composition: Purpose, process, and practice*. New York: Routledge.
- Fox, J. (2001). *It's all about meaning: L2 test validation in and through the landscape of an evolving construct*. [Unpublished doctoral dissertation]. McGill University: Montreal, Canada.
- Fox, J. (2004). Test decisions over time: Tracking validity. *Language Testing*, 21(4), 437-465.
- Fox, J.D. (2009). Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes*, 8 (1), 26-42
- Fox, J., & Artemeva, N. (2017). From diagnosis toward academic support: developing a disciplinary, esp-based writing task and rubric to identify the needs of entering undergraduate engineering students. *ESP today-journal of English for specific purposes at tertiary level*, 5(2), 148-171.
- 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*, 14(1), 9- 26.
- Fox, J., & Cheng, L. (2015). Walk a mile in my shoes: Stakeholder accounts of testing experience with a computer-administered test. *TESL Canada Journal*, 32, 65-86.

- Fox, J., Cheng, L. & Zumbo, B. D. (2014). Do they make a difference? The impact of English language programs on second language students in Canadian universities. *TESOL Quarterly*, 48(1), 57-85.
- Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18(9), 27-32.
- Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics*, 20(2), 221-236.
- Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics*, 29, 3-20.
- Fulcher, G. (2013). *Practical language testing*. New York: Routledge.
- Fullan, M. (2007). *The new meaning of educational change*. New York: Routledge.
- Fullan, M. (2015). *The new meaning of educational change*. New York; Routledge.
- Fullilove, J. (1992). The tail that wags. *Institute of Language in Education Journal*, 17(2), 205-217.
- Gipps, C. (1994). Developments in educational assessment: What makes a good test? *Assessment in Education: Principles, Policy & Practice*, 1(3), 283-292.
- Gorsuch, G.J. (2000) Educational policies and educational cultures: influences on teachers' approval of communicative activities. *TESOL Quarterly*, 34(4), 675-710.
- Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education* (Vol. 25). Cambridge, England: Cambridge University Press.
- Green, A. (2013). Washback in language assessment. *International Journal of English Studies*, 13(2), 39-51
- Hamp-Lyons, L. (2016) Purposes of assessment In J. Banerjee & D. Tsagari (Eds.), *Handbook of second language assessment* (Vol. 12). Berlin: De Gruyter
- Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. *College ESL*, 6(1), 52-72.
- Hargreaves, A., Earl, L., & Schmidt, M. (2002). Perspectives on alternative assessment reform. *American Educational Research Journal*, 39(1), 69-95.
- Hatch, J. A. (2002). *Doing qualitative research in education settings*. Albany, NY: SUNY Press.

- Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 97-112). Mahwah, NJ: Lawrence Erlbaum.
- Heaton, J.B. (1990). *Classroom testing*. Pearson PTR.
- Heyneman, S. P., & Ransom, A. W. (1990). Using examinations and testing to improve educational quality. *Educational Policy*, 4(3), 177-192.
- Henrichsen, L. E. (1987). *Diffusion of innovations in English language teacher: The English language exploratory committee's promotion of C.C. Fries' oral approach in Japan, 1956-1968*. Honolulu, HI: University of Hawaii.
- Henrichsen, L. E. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956-1968*. New York: Greenwood Press.
- Ho, D. G. E. (2012). Focus groups. *The Encyclopedia of Applied Linguistics*. DOI:10.1002/9781405198431.wbeal0418
- House, E. R. (1979). Technology versus craft: A ten year perspective on innovation. *Journal of Curriculum Studies*, 11(1), 1-15.
- House, E. R. (1981). Three perspectives on innovation. In R. Lehming & M. Kane, (Eds.). *Improving schools: Using what we know* (pp. 89-111). Thousand Oaks, CA: Sage
- Hughes, A. (1989). *Testing for language teachers*. Cambridge, England: Cambridge University Press.
- Hughes, A. (1993). *Backwash and TOEFL 2000*. [Unpublished manuscript]. University of Reading: Reading, England.
- Hughes, A. (2003). *Testing English for language teachers*. Cambridge, England: University of Cambridge.
- Jamieson, J. (2014). Defining constructs and assessment design. *The companion to language assessment*. DOI: 10.1002/9781118411360
- Kennedy, C. (1988). Evaluation of the management of change in ELT projects. *Applied Linguistics*, 9(4), 329-342.
- Kramersch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, 70(4), 366-372.
- Lam, H. P. (1993). *Washback: can it be quantified? A study on the impact of English examinations in Hong Kong*. [Unpublished MA dissertation]. University of Leeds: Leeds, England.

- Lamie, J. M. (2004). Presenting a model of change. *Language Teaching Research*, 8(2), 115-142.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry* (Vol. 75). Thousand Oaks: Sage.
- Linn, R. L., Baker, E. L. & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15-21.
- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29(2), 4-16.
- Lumley, T. & Stoneman, B. (2000). Conflicting perspectives on the role of test preparation in relation to learning? *Hong Kong Journal of Applied Linguistics*, 5, 50-80.
- Maclellan, E. (2001). Assessment for learning: The differing perceptions of tutors and students. *Assessment & Evaluation in Higher Education*, 26(4), 307-318.
- Macnaghten, P. & Myers, G. (2006). Focus groups. In C. Seale, G. Gobo, J.F. Gubrium & D. Silverman (Eds.), *Qualitative research practice: Concise paperback edition* (pp. 65-79), London: Sage.
- Madaus, G. F. (1985). Public policy and the testing profession—You've never had it so good? *Educational Measurement: Issues and Practice*, 4(4), 5-11.
- Madaus, G.F. (1988). The influence of testing on the curriculum. In Tanner, L.N. (Ed.), *Critical issues in curriculum: eighty-seventh yearbook of the National Society for the Study of Education* (pp. 83-121). Chicago, IL: University of Chicago Press.
- Madaus, G.F. (1989) The Irish study revisited. In B. R. Gifford (Ed.). *Test policy and test performance: Education, language, and culture* (pp. 63-89) [Vol. 23 in *Evaluation in Education and Human Services*.] Rotterdam: Kluwer Academic.
- Markee, N. (1993). The diffusion of innovation in language teaching. *Annual Review of Applied Linguistics*, 13, 229-243.
- Markee, N. (1997). *Managing curricular innovation*. Cambridge, England: Cambridge University Press.
- McNamara, T. F. (1996). *Measuring second language performance*. Boston, MA: Addison Wesley Longman.
- McNamara, T.F. (2000). *Language testing*. Oxford, England: Oxford University Press
- Merriam, S. B. (1998). *Qualitative research and case study Applications in education. Revised and expanded*. San Francisco, CA: Jossey-Bass.

- Merriam, S. B. (2009). *Qualitative research: A guide to design and interpretation*. San Francisco, CA: Jossey-Bass.
- Messick, S. (1989), 'Validity.' In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.
- Messick, S. (1996a). Validity and washback in language testing. *Language Testing*, 13 (3) p. 241-256
- Messick, S. (1996b). Validity and washback in language testing. *ETS Research Report Series*, (pp.1-18). DOI:10.1002/j.2333-8504.1996.tb01695
- Miles, M. B. & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage.
- Miles, M. B., Huberman, A. M. & Saldana, J. (2014). *Qualitative data analysis: A method sourcebook*. Thousand Oaks, CA: Sage.
- Moore, T. & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes*, 4(1), 43-66.
- Morrow, K. (1991). Evaluating communicative tests. In S. Anivan (Ed.), *Current developments in language testing* (pp. 111-118). Singapore: SEAMEO Regional Language Centre.
- Murray, D. E. (2008). *Planning change, changing plans: Innovations in second language teaching*. Ann Arbor, MI: University of Michigan Press.
- Nagy, W. & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, 47(1), 91-108.
- Nolen, S. B., Haladyna, T. M. & Haas, N. S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice*, 11(2), 9-15.
- Patton, M. Q. (2002). *Qualitative research*. Hoboken, NJ: John Wiley.
- Pawlikowska-Smith, G. (2002). *Canadian language benchmarks 2000: Theoretical framework*. Ottawa, ON: Centre for Canadian Language Benchmarks.
- Pearson, I. (1988). Tests as levers for change. In D. Chamberlain & R. J. Baumgardner, (Eds.), *ESP in the classroom: practice and evaluation* (pp. 98-107). [EDLT Documents, 128]. Oxford, England: Modern English.
- Peat, J. (2001). *Health science research: A handbook of quantitative methods*. Sage.
- Plakans, L. (2013). Assessment of integrated skills. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. DOI: 10.1002/9781405198431

- Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan*, 68, 679–82.
- Powers, D. E. (2010). Validity: What does it mean for the TOEIC tests. *TOEIC Compendium 1.10* Princeton, NJ: Educational Testing Service.
- Purpura, J. E. (2004). *Assessing grammar* [Cambridge Language Assessment Series]. Cambridge, England: Cambridge University Press.
- Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, 22(2), 142-173.
- Qi, L. (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), 51-71.
- Rea-Dickins, P. & Scott, C. (2007). Washback from language tests on teaching, learning and policy: Evidence from diverse settings. *Assessment in Education: Principles, Policy & Practice*, 14(1), 1-7. Available at: http://ecommons.aku.edu/eastafrica_ied/46
- Read, J. & Hayes, B. (2003). The impact of IELTS on preparation for academic study in New Zealand. *IELTS International English Language Testing System Research Reports*, 4, 153-206.
- Resnick, L. B., & Schantz, F. (2017). Testing, teaching, learning: Who is in charge? *Assessment in Education: Principles, Policy & Practice*, 24(3), 424-432.
- Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.
- Roulston, K. (2013). Interview in Qualitative Research. *The encyclopedia of applied linguistics*. DOI: 10.1002/9781405198431
- Rubin, H. J., & Rubin, I. S. (2012). *Qualitative interviewing: The art of hearing data*. Thousand Oaks, CA: Sage.
- Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Language Testing*, 23(1), 1-34.
- Saldana, J. (2013). *The coding manual for qualitative researchers*. Thousand Oaks, CA: Sage.
- Schmitt, D. & Hamp-Lyons, L. (2015). The need for EAP teacher knowledge in assessment. *Journal of English for Academic Purposes*, 18, 3-8.
- Schulz, J. (2012). *Analysing your interviews* [Youtube Video], Southampton, England: University of Southampton. Retrieved from <http://www.southampton.ac.uk/education>

- Scott, C. (2007). Stakeholder perceptions of test impact. *Assessment in Education*, 14(1), 27-49.
- Seidman, I. (2013). *Interviewing as qualitative research: A guide for researchers in education and the social sciences*. New York: Teachers College Press.
- Shepard, L. A. (1993). Chapter 9: Evaluating Test Validity. *Review of Research in Education*, 19(1), 405-450.
- Shih, C. (2007). A new washback model of students' learning. In A. Cumming & M. Laurier (Eds.), *Language assessment*, [Special Issue of *Canadian Modern Language Review*], 64(1), 133-160.
- Shohamy, E. (1993). *The power of tests: the impact of language tests on teaching and learning*. [NFLC Occasional Paper]. College Park, MD: National Foreign Language Center, University of Maryland.
- Shohamy, E. (2000). The relationship between language testing and second language acquisition, revisited. *System*, 28(4), 541-553.
- Shohamy, E. (2001). *The power of tests: a critical perspective on the uses of language tests*. London: Pearson Education.
- Shohamy, E. (2017). Critical language testing. In E. Shohamy & S. May (Eds.) *Language testing and assessment: Encyclopedia of language and education (3rd ed.)*. Cham, SW: Springer.
- Shohamy, E., Donitsa-Schmidt, S. & Ferman, I. (1996). Test impact revisited: washback effect over time. *Language Testing*, 13, 299-317.
- Sloane, F. C. & Kelly, A. E. (2003). Issues in high-stakes testing programs. *Theory into Practice*, 42(1), 12-17.
- 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. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.
- Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, 18(1), 23-27.
- Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. *Input in Second Language Acquisition*, 15, 165-179.
- Stoller, F. L. (1994). The diffusion of innovations in intensive ESL programs. *Applied Linguistics*, 15(3), 300-327.

- Stoyhoff, S. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching*, 42(1), 1-40.
- Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99-103.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53.
- Tsagari, D. (2009). Revisiting the concept of test washback: Investigating FCE in Greek language schools. *Cambridge ESOL. Research Notes*, 35, 5-9.
- Tsagari, D., & Cheng, L. (2017) Washback, impact, and consequences revisited. In: E. Shohamy, I. Or & S. May (Eds) *Language Testing and Assessment. Encyclopedia of language and education (3rd ed.)*. Cham, SW: Springer.
- Turner, C. E. (2006). Professionalism and high-stakes tests: Teachers' perspectives when dealing with educational change introduced through provincial exams. *TESL Canada Journal*, 23(2), 54-76.
- 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.
- Wall, D., & Alderson, J. C. (1993). Examining washback: the Sri Lankan impact study. *Language Testing*, 10(1), 41-69.
- Wall, D. (1997). Impact and washback in language testing. In S. May (Ed.), *Encyclopedia of language and education* (pp. 291-302). New York: Springer.
- Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System*, 28, 499-509.
- Wall, D. (2005). The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory. *Studies in Language Testing*, 22. Cambridge, England: Cambridge University Press.
- Wall, D., & Horák, T. (2006). Using baseline studies in the investigation of test impact. *Assessment in Education*, 14(1), 99-116.
- Wall, D., & Horák, T. (2008). 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*, (pp. i-105). DOI: 10.1002/j.2333-8504.2008.tb02123.x

- Wall, D., & Horák, T. (2011). The impact of changes in the TOEFL exam on teaching in a sample of countries in Europe: Phase 3, The role of the coursebook phase 4, describing change. *ETS Research Report Series* (pp. i-181). DOI: 10.1002/j.2333-8504.2011.tb02277.x
- Wall, D. (2012). Washback. *The Routledge handbook of language testing* (pp. 79-92). New York: Routledge.
- Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 129-146). Mahwah, NJ: Lawrence Erlbaum.
- Waters, A. (2009). Managing innovation in English language education. *Language Teaching*, 42(4), 421-458.
- Waters, A. (2014). Managing innovation in English language education: A research agenda. *Language Teaching*, 47(1), 92.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge, England: Cambridge University Press.
- Weigle, S. C. & Malone, M. M. (2016). Assessment of English for academic purposes. *The Routledge Handbook of English for Academic Purposes* (pp. 608-620). New York: Routledge.
- Weir, C. J. & Roberts, J. (1994). *Evaluation in ELT*. Oxford, England: Blackwell.
- White, R. (1988). *The ELT curriculum: Design, management and innovation*.
- White, R. (1993). Innovation in Curriculum Planning and Program Development. *Annual Review of Applied Linguistics*, 13, 244-259. DOI:10.1017/S026719050000249X
- Wiliam, D. (1996). National Curriculum Assessments and Programmes of Study: Validity and Impact. *British Educational Research Journal*, 22(1), 129-141. Retrieved from <http://www.jstor.org.proxy.library.carleton.ca/stable/1501582>
- Williams, J. D. (2003). *Preparing to teach writing: Research, theory, and practice*. London: Routledge.
- Winke, P. (2011). Evaluating the Validity of a High- Stakes ESL Test: Why Teachers' Perceptions Matter. *Tesol Quarterly*, 45(4), 628-660
- Xi, X., Bridgeman, B., & Wendler, C. (2014). Tests of English for academic purposes in university admissions. In J. A. Kunnan (Ed.) *The companion to language assessment* (pp. 318-337). Hoboken, NJ: John Wiley. DOI: 10.1002/9781118411360

Yin, R. K. (2009). *Case study research: Design & methods* (4th ed.). Thousand Oaks, CA: Sage.

Yin, R. K. (2014). *Case study research: Design and methods*. Thousand Oaks, CA: Sage.

Appendices

Appendix A: Research focus and methods of prominent empirical studies on washback

STUDY	TEACHING LEARNING CONTEXT	PURPOSE/ MAIN ISSUES ADDRESSED	METHODS	RESEARCH FOCUS
Alderson and Wall, 1993; Wall and Alderson, 1993	Sri-Lankan Secondary Schools – O-Level evaluation project	To observe effects of changing the O-Level examinations	Class room observation Teacher and student interviews	-Curriculum -English teachers -Teacher advisors -Teaching process
Lam, 1993	Hong Kong secondary schools –revised Use of English Exam	To examine how examination system is used to bring about a positive washback on English language classrooms.	Teacher questionnaire Analysis of textbooks	-English teachers -Teaching materials, textbooks
Shohamy, Donitsa, & Ferman, 1996	Secondary schools in Israel – ASL and ESL tests	To examine the impact of two national tests in and beyond classroom settings	Student questionnaire Structured interviews with teachers and inspectors Analysis of inspectorate bulletins	-Arabic and English teachers -Head teachers -School administrators -Inspectors - Parents
Alderson & Hamp- Lyons, 1996	TOEFL and language schools for university entrants	To determine the influence of TOEFL on classroom teaching	Student and teacher interviews Classroom observations	-Teaching materials -Test takers -Language teachers – what and how teachers taught
Cheng, 1997; 2005	Hong Kong Secondary schools – old and new HK certificate of English Exam	To compare teachers’ perceptions towards both old and new exams when an examination was used as a change agent	Teacher questionnaire Student questionnaire Classroom observation Structured interviews with teachers	-Test takers -Language teachers’ teaching style and materials -Material developers

Wall, 1999; 2005	Sri-Lankan Secondary Schools – British O-Level evaluation project	To examine the effects of changing the O-Level examinations on teaching – use of examination as a change agent	1. Individual and group interviews with teachers 2. Teacher questionnaires and teacher advisors questionnaire 3. Observations 4. Material and test analysis	-Teacher training and background -Role of publishers in material design -General social and political context
Hamp-Lyons, 1998	TOEFL studies	To examine the the role of textbooks in test washback	Analysis of 5 TOEFL preparation textbooks	-Textbook analysis
Watanabe, 1996; 2000; 2004	Private extra-curricular institutions in Japan preparing students for university entrance exams	To investigate the relationship between university entrance examinations and grammar-translation approach to teaching	Analysis of entrance exam papers Classroom observation Teacher interviews	- English teachers' background and training -Teaching methods and use of materials -Teachers' beliefs about teaching and attitude towards exams
Lummley and Stoneman, 2000	Hong Kong Poly technique University – Graduating students' Language Proficiency Assessment	To understand students and teachers' reactions to exam preparation materials. To understand student motivation in relation to test stakes – IELTS vs. Hong Kong's Graduating Students Language Proficiency Assessment (GSLAP).	Interviews with teachers Student questionnaire	- High-stakes test (IELTS) -Low-stakes test- Hong Kong's Graduating Students Language Proficiency Assessment (GSLAP) -Students' and teachers' beliefs
Andrew et al. 2002	Hong Kong secondary schools – revised Use of English Exam	To investigate the washback effect from a revised exam on the teaching of English in Hong Kong secondary schools.	Video recordings of mock oral tests. Discourse analysis Grading of oral tests	-Test takers -Discourse markers such as transition words
Turner, 2006, 2009	Canadian-French speaking primary and secondary schools – ESL exams	To examine the development of rating scales and their consequential effects on the teachers involved in the	Feedback from teachers involved in developing rating scales	-Teachers as assessors and evaluators -Development of rating scales

		development of assessment instruments		-Classroom assessment
Read and Hayes, 2003	Tertiary institutions and private schools in New Zealand conducting IELTS or other English classes	To investigate the impact of IELTS on preparation for academic study in New Zealand	Questionnaires to schools, teachers and students Interviews with teachers Observation of classes Pre- and post-tests	-Test taking practices by students and teachers -Relationship between academic skills and language proficiency
Green, 2007	IELTS preparation courses, pre-sessional EAP courses, and combined courses in the UK	To examine how preparation classes impact score gains	Two IELTS writing tests Two questionnaires with participant and process variables respectively	-Test features -Learning outcomes -Academic skills and language proficiency
Shih, 2007	Private technical colleges in Taiwan – General English Proficiency Test (GEPT)	To explore the effects of GEPT exit requirements on learning.	Interviews with department heads, teachers, students, and family members Classroom observation	-Test takers and raters -Test preparation classes -Comprehensive list of intrinsic and extrinsic and test factors. -Policies
Qi, 2004, 2005, 2007	CTE in China	To explore the effects of communicative tests on communicative teaching	Survey Questionnaires	Teacher beliefs
Tsagari (2009)	First Certificate in English (FCE) in Greece	To examine the washback of FCE in intermediate EFL classes	Interviews, exam preparation materials, student diaries	-Teachers, students Teacher trainers, test developers
Saville, 2009	Cambridge ESOL, high-stakes language examinations and their impact	To investigate the test impact at micro and macro level To propose a new expanded model of impact to provide Cambridge ESOL with a theory of action for test validations	Meta-analysis of three case studies test impact as part of test development and validation procedures	Improvements to the examination system by investigating the test impact at micro and macro levels
Wall & Horak (2006, 2008, 2011)	TOEFL iBT in Central and Eastern Europe	To examine if the new TOEFL iBT brought any changes in teaching and learning after the change in exam format	Longitudinal study in three phases. Computer-mediated contact with teachers	Teachers beliefs and their reactions to tasks and test construct Classroom practices

Appendix B: Phase 1 - Teacher interview questions

1. Tell me little bit about your background and how long you have worked within the EAP program?
2. How many times have you taught the Graduating Level?
3. Tell me a bit about your teaching philosophy.
4. What factors (such as course objectives, textbooks, etc.) do you keep in mind when you plan your lessons?
5. Tell me about assessment activities that take place in your class. What factors do you keep in mind when you plan these?
6. Do you assess all skills equally? If not, how much weighting do you give to each skill?
7. Tell me about the mid-term and final tests at the Graduating Level. Are they the same or different than tests at other levels in the EAP program? How?
8. Do you think that the intended program learning outcomes prepare students for the final ExitTest? If yes, how and if not, why not?
9. How important are activities like pair-work and group-work in your classes? What kind of feedback do you provide for such activities?
10. Are you aware of the curricular changes that are going to take place at the EAP program? If yes, tell me more about that.
11. Do you think these changes will affect your teaching practice? If yes, how?
12. Do you think these changes will affect the teaching materials that you use? If yes, how?
13. Finally, do you think these changes will affect your assessment practices? If yes, how?

Appendix C: Phase 1 - Student questionnaire

- a. What is your first language?
- b. Do you have a conditional offer of admission at this university? If so, what program?
- c. What ExitTest score do you need to do this course?
 - a. 3.5 3.5+ 4- 4 4.5 Don't know
- d. How long have you been with the EAP program?

1 semester	2 semesters	More than 3
------------	-------------	-------------

Read the following statements related to assessment practices at the EAP program. The following questions are about your understanding of the purpose, content, timing, mode, making, feedback and markers of assessment. Choose what you think is the most appropriate response.

(* (F)Frequently (S)Sometimes (N)Never (DN)Don't know)

- | | | | | |
|---|---|---|---|-----|
| 1. Assessment motivates learning | F | S | N | DN* |
| 2. Assessment is used for grades/ranks | F | S | N | DN |
| 3. Assessment is used for diagnosis | F | S | N | DN |
| 4. Assessment is used to evaluate teaching | F | S | N | DN |
| 5. Development of knowledge is assessed | F | S | N | DN |
| 6. Application of knowledge is assessed | F | S | N | DN |
| 7. Course outline is discussed in class | F | S | N | DN |
| 8. Can-do statements are explained in class | F | S | N | DN |
| 9. Can-do statements are used in teaching | F | S | N | DN |
| 10. Self-assessment is used in classes | F | S | N | DN |

11. Peer assessment is used in classes	F	S	N	DN
12. Portfolios are used for writing assignments	F	S	N	DN
13. Assessed during a term	F	S	N	DN
14. Assessed at the end of a term	F	S	N	DN
15. Assessed when student feels ready	F	S	N	DN
16. Assessed through audio/video presentations	F	S	N	DN
17. Assessed in computer labs	F	S	N	DN
18. Assessed in what is taught in class	F	S	N	DN
19. Assessed by essay	F	S	N	DN
20. Assessed by multiple-choice questions	F	S	N	DN
21. Assessed by short answer questions	F	S	N	DN
22. Assessed on note-taking skills in listening/reading	F	S	N	DN
23. Assessed through writing reflections	F	S	N	DN
24. Assessed by projects and tasks	F	S	N	DN
25. Students understand assessment criteria	F	S	N	DN
26. Students do not understand assessment criteria	F	S	N	DN
27. Marking strengthens knowledge	F	S	N	DN
28. Making develops thinking	F	S	N	DN
29. Making improves presentation	F	S	N	DN
30. Essay is given a mark	F	S	N	DN
31. Rubric for making writing essay is explained	F	S	N	DN
32. Essay is marked for language accuracy	F	S	N	DN
33. Essay is marked for overall communication	F	S	N	DN
34. Essay is second-marked by other teachers	F	S	N	DN
35. One-to one tutorials are provided	F	S	N	DN
36. Mid-term and end-of-term feedback is useful	F	S	N	DN
37. Feedback is provided for essays	F	S	N	DN
38. Feedback prompts discussion with tutor	F	S	N	DN
39. Feedback helps understand assessment	F	S	N	DN
40. Feedback improves learning	F	S	N	DN

Students from your class are invited to participate in a focus group. If you are willing to participate, please enter your email below:

(Items 1-6, 10 -29 and 36-40 in this questionnaire were drawn from Maclellan, 2001 with her permission)

Appendix D: Phase 1 – Student focus group questions

Welcome everyone... what I have here is a list of activities, mostly test related rank order what you think is the most important and which one you think is least important or you don't really need to spend time on from the point of view of passing the graduating level:

1. Teacher's lecture in class
2. Reading different types of texts (from textbooks or other materials)
3. Reading and writing short or long answers to questions
4. Remembering vocabulary and practicing grammar
5. Spending time in practicing summarizing, paraphrasing and referencing
6. Understanding all aspects of the final test (e.g. goals, content, format and rating process)
7. Spending time in class practicing writing, such as a 5-paragraph essay

8. Spending time in class practicing practice-tests so that you are familiar with the test (e.g. structure, vocabulary, and cloze-exercises etc.)
9. Spending time in class practicing presentations and group discussion
10. Spending time in class practicing note-taking while reading and listening activities

- A. Why do you think teachers choose these activities?
- B. What do you think is the link between those activities, your course objectives and can-do statements in your course outlines?
- C. Do you think you have made progress in learning academic skills this semester? If yes, can you give me some examples of where you think you have made the most progress?
- D. This might be different for each group member, but with what area do you think you need more practice? Or something that you thought you will be able to improve but with which you didn't make as much progress? So something that you need more practice with and you think you did not getting enough of it.
- E. What could your teachers do to improve that?

Appendix E: Phase 1- Administrator interview questions

The interview was divided into three sections: EAP Program – present and past, graduating level, and future changes in the program

Program – Present and Past

1. Tell me little bit about your background and how long you have been with the EAP program?
2. What are the university's expectations of the EAP program?
3. Students:
 - a. Demography – mostly where from –levels etc.
 - b. Placement procedure -How are students placed in different levels? Who decides?
 - c. What do students, in general, expect from this program?
 - d. What are the program's expectations of students?
 - e. How are students supported in the EAP?
 - f. What are the exit criteria for students (in general now, but in detail later)?
4. Teachers:
 - a. Teacher demography
 - b. General criteria for teacher selection (not specifications)
 - c. What are the program's expectations of teachers?
 - d. What are teachers' expectations of this program?
 - e. What have you noticed about the pedagogical practices of our teachers? Are all skills given equal importance? If not, which is the most targeted skill? Why?
 - f. How are teachers supported in the EAP program?
5. What are some of the traditional assessment practices in our program?
6. Are you aware of any previous reforms or changes to this program? If yes, please explain them in detail.

Graduating level

7. What are program expectations of graduating level students?
8. What are the program expectations of graduating level teachers?

9. Is there any additional support provided for teachers (e.g. writing rubric workshop etc.) and students at the graduating level?
10. How important are prescribed textbooks for the graduating level?
11. What are the current exit criteria for graduating level students?
12. What is the role of the Testing Office in the exit procedures of our students?

Intended Changes

13. Do you think a change (whether in teaching, learning or assessment) is required in our program? Why?
14. What's your vision of the change(s) and what will be your role in this change process?
15. What's the goal of these changes? What do you hope will be the outcome of these changes?
16. What are some of the classroom pedagogical practices that you envision with this change?
17. Will students' expectations change?
18. Will teachers' expectations change?
19. How will these changes be implemented?
20. Will the exit criteria change at the graduating level? How? What will the assessment procedures be?
21. What will the role of the Testing Office in the new scenario be?
22. What support will be provided to the teachers and students at the graduating level?
23. What factors will facilitate or hinder the implementation and diffusion of change (specifically innovation in assessment)?

Appendix F: Phase 2 - Teacher interview questions set 1

This interview was divided into three parts: awareness, interest and evaluation of the new testing regime.

Awareness

1. What do you think are the main differences between the revised GL and the former GL tests?
2. Do you think it is easy/difficult to test integrated skills in the new L/S and R/W tests? Why and How?
3. In your opinion, what knowledge or skills do students need to get a good grade at the graduating level? How are these skills/this knowledge different than those required in the previous semesters?
4. Did these changes make you change your teaching methods? If yes, how? If not, why not?
5. Have you worked on any extended essays or research projects in this term so far?
6. Did you do any special activities to prepare your students for the midterm test? If yes, what materials did you use? How did you use this material or what techniques do you use?
7. What differences do you see in marking the new GL tests? What do you look for when marking writing and speaking?

Interest

1. Do you think that the new GL test has positive, negative or neutral on your teaching, or both? In what ways?

2. Do you personally enjoy teaching at the graduating level? Why/why not?
3. What has most influenced your teaching this semester?

Evaluation

1. Overall, what do you think of the changes introduced at the graduating level?
2. Do you think the objectives of the GL Tests overlap those of the new PLOs - Program learning outcomes? (Present a copy)
3. Which sections/aspects of the new PLOs are hardest to teach?
4. Which sections/aspects the new PLOs are easiest to teach?
5. Has the role of the Teaching Assistants changed because of the recent changes? If yes, how?
6. How useful are the sample tests and rubrics provided by the administration?
7. What were students' responses towards the midterm grading? Were they satisfied?
8. Finally, how will this change impact teaching at the lower levels of the program?

Appendix G: Phase 2- Teacher interview questions set 2

1. Before the assessment change, writing was the dominant skill. How about now, has skill importance changed?
2. Over the length of this semester, (doesn't have to be 100% accurate), what percentage of time did you spend in class on the following language elements: Reading, listening, writing, speaking, and grammar. Why?
3. How was the test preparation different this semester than the previous semesters?
4. In terms of teaching speaking, because teachers earlier pointed out the tensions between simple language instruction and speaking for academic purposes, what do you think is important for students to learn?
5. Can you tell me some of your testing and evaluating strategies for speaking?
6. One of our course learning outcomes is to write a well-structured report, or a short research project about 800-1000 words. How many research papers/reports did you assign?
7. What guidelines or templates or rubrics, were you provided by the office regarding this?
8. How have students benefited by doing a research project in terms of assessment? Were there any marks assigned for this counted in the final assessment?
9. How easy or difficult was it to teach referencing and citation?
10. What guidelines or citation-management templates did the office provide?
11. What were students' reactions toward learning referencing and citations?
12. Using criteria other than tests and assessments in the final grades, how much of classroom assessment activities were incorporated in the final grade. For example, peer work or self-assessment or portfolios if you had any.
13. We have new program learning outcomes and course learning outcomes. One of the reasons we have outcomes is so that we can measure them. Were these incorporated in assessment?
14. What role did the textbooks play in the midterm and the final test?
15. What were students' responses to using this textbook?
16. Going back to the issue of teacher autonomy versus the testing office control noted in previous teacher interviews. What is the role of the testing office in the final tests and how is that different from the previous semesters?
17. Were you satisfied with the final tests and did you contribute toward the final test?
18. How did students feel about the final test?

19. I have a couple of statements here. If you want you can use a Likert scale, 1-10 or, if you want to discuss them in general, that's ok too.
- How satisfied are you with communication regarding the what, how, and why of how assessment changed.
 - Support provided by the main office regarding sample papers, rubrics, or training to teach or assess especially speaking.
 - The time commitment regarding extra preparation for teaching or assessment, like looking for appropriate materials or preparation for speaking activities.
 - How satisfied are you with your say in the assessment process or change process?
 - Your satisfaction with the role of external testing experts?
 - Satisfaction with teacher collaboration with the testing office for the final test. Were teachers contacted or did teachers collaborate on the final test?
 - How do you perceive these changes? Top-down or bottom-up approach?
 - Are you satisfied to embrace these new changes?

Appendix H: Phase 2 – Student questionnaire

- How often did you use English outside of school in your home country?
Most of the time – 80% - 100% Often – 60% - 80% Sometimes – 40% - 60%
Seldom - 20% - 40% Almost never - > 20%
- How often did you write assignments in English in your home country?
Most of the time – 80% - 100% Often – 60% - 80% Sometimes – 40% - 60%
Seldom - 20% - 40% Almost never - > 20%
- How often did you participate in group work and oral presentations in English in your home country?
Most of the time – 80% - 100% Often – 60% - 80% Sometimes – 40% - 60%
Seldom - 20% - 40% Almost never - > 20%
- How many semesters have you studied in this EAP program?
One two more than two
- Do you think you understand the course learning outcomes (or course objectives) of graduating level?
Definitely yes yes No opinion No Definitely no
- Do you usually know why you received a specific grade on a paper or test, for example because of good content, vocabulary, organization or other criteria?
Most of the time – 80% - 100% Often – 60% - 80% Sometimes – 40% - 60%
Seldom - 20% - 40% Almost never - > 20%
- Have you worked on any research projects (finding sources and using them to write a paper of 800 – 100 words) in your class?
Yes No
- If yes, how many
One Two More than two
- Do you know how to cite sources (e.g. According to X... or Davis-Floyd, P (2015)..... University Press. etc.) in your written assignments like a summary or paraphrasing?

Most of the time – 80% - 100% Often – 60% - 80% Sometimes – 40% - 60%
 Seldom - 20% - 40% Almost never - > 20%

10. Did you like the format of the midterm test?
 Definitely yes yes No opinion No Definitely no

11. Do you think that the midterm tests were a reflection of the work you did in class?
 Definitely yes yes No opinion No Definitely no

12. Did you do any extra listening and speaking activities other than the activities in the textbook?
 Yes No

13. Do you feel that the extra activities that you did in class helped you prepare for the midterm tests?
 Definitely yes yes No opinion No Definitely no

14. What kind of feedback did you get on your midterm tests?

Students from your class are invited to participate in a focus group. If you are willing to participate, please enter your email below

Appendix I: Phase 2 – Student focus group questions

Welcome everyone. In the questionnaire that I administered to your class, almost 47% students Only 26% students liked the tests.

1. What aspects of the final/midterm tests supported/enhanced your ability to illustrate what you've learned? What aspects of the final/midterm tests blocked or did not support what you have learned? Be specific by providing an example.
2. Did you see evidence of classroom work/exercises being included in the final/midterm? Provide examples of what you mean.
3. What kinds of test-related activities, exercises, tests, or advice did your teachers provide to you leading up the final/midterm?
4. What aspects of the textbooks helped you learn the material? Did anything about the textbooks (format, layout, examples, language, etc.) challenge or block the usefulness of the book for your learning?
5. Did you feel that the criteria used for marking your tests were reflective of your learning? In what? Or why not? Give examples.
6. Did you do any group/pair-work in your classroom activities? If yes, did you find them useful? Why/why not?
7. Describe the feedback that you received (or received for the midterm) on the classroom assessment activities like assignments and quizzes. What was the most useful feedback? What feedback was not very useful? Give examples.

Appendix J: Phase 2 - Administrator interview questions

1. In our previous interview, we discussed the planning part of the new assessment. Today we will discuss the implementation and evaluation part of the change process in that we will discuss the attributes of change, challenges in implementation process and overall impact of the assessment change. My first question is what has changed?

2. What were the curriculum changes in terms of teaching or testing materials that were made for this change?
3. I have a couple of attributes of innovation and change. Some of these you had identified last time when we discussed planning to implement these changes. Let's discuss these one by one. If you wish, you could use the Likert scale of 0 to 4 describe these:
 - a. First one is 'practicality'
 - b. The feasibility of the innovation
 - c. Usefulness of the innovation
 - d. Compatibility of this innovation with the past practices.
 - e. Complexity of these tests?
4. Do you see the new tests as an improvement over past practices?
5. What were the challenges faced by your office in implementing these changes?
6. Do you think teachers were ready for this change?
7. Were students ready for this change?
8. In our previous interview, you mentioned that writing was the dominant skill before change. Has the skill importance changed?
9. The tests are different now than the ExitTest, so how much input we have from the Testing Office in construction of these tests given that earlier the Testing Office was in charge of the construction and administration of the tests.
10. How much was teacher collaboration in the midterm and final test? Were teachers involved in test construction?
11. Has this change in assessment changed the exit criteria for the students?
12. What were the intended outcomes you anticipated from this assessment change?
13. To what extent have you observed these intended outcomes to have actually occurred?
14. Overall, in your opinion, is this change successful?
15. Would you like to add anything else to this discussion?

Appendix K: Phase 3 -Teacher interview questions

We'll continue our discussion from three semesters ago about the changes made in the curriculum and assessment in the EAP program.

1. How is your teaching different now than before the changes in the graduating level?
2. How is the test prep different now than earlier?
3. Has this improved your teaching in anyway?
4. How much research-based or project-based assessment is incorporated in your teaching now?
5. Are students happy about these changes?
6. Going back to the test, what kind of feedback do you provide on the new tests to your students?
7. How realistic do you think these tests are and how are they not? The test itself and what's expected of students. Are these tests more relevant to students' needs?
8. According to you, what further development is needed in these tests?
9. If you had concerns with the new tests, did you raise them with the administration?
10. What priority is given to the speaking skill in the overall weightage?
11. How do you see the role of the Testing Office in the new testing regime? Has it changed from the previous semesters?
12. We have 8 or 9 CLOs, softer skills like negotiation and teamwork are given priority but how much weighting is given to them in students' final outcomes?
13. How engaged are students in their assessment as compared to previous exams?

14. How is this emphasis on global skills and integrated tests, as opposed to discrete skills on MC exams beneficial for students?
15. The speed of change - Has it been something that you can handle while addressing your day-to-day needs, especially with assessment?
16. Do you feel in control of the pace of change or do you feel that it's all imposed upon you?
17. Would you see this new assessment as having a positive or negative impact on students' learning, or perhaps both?
18. Would you like to add anything else to your comments?

Appendix L: Phase 3 - Student focus group questions

1. What are you studying now? How many courses have you taken so far?
2. You must've done a number of different (types of) assignments this past year. Can you tell a bit about them?
3. Which were most successful for you? Why do you think you did better in these?
4. What about the ones you did less well in? Why do you think that?
5. What are some of the reading and writing expectations of different subjects?
6. How are these similar or different than reading/writing expectations that you encountered at the EAP program?
7. How do you feel about your academic (all four) skills now?
8. Were you prepared for the university studies? Did EAP program prepare you well for the university courses? If yes, how? If not, why not?
9. Have your academic skills improved? If yes, what contributed to the improvement? If not, what is lacking?
10. How would you describe a successful research-based project?
11. Was there anything that you learnt at the EAP program that is helpful in your research-based assignments? Can you give specific examples?
12. Do you remember the midterm and final tests at the EAP program? If yes, do you think they were relevant to your current needs in the university studies?
13. What are your views about the feedback (written or verbal) from your teachers in your degree courses?
14. Being an L2 student, what are some of challenges that you face in completing your assignments?
15. What do you wish you had learnt more at the EAP program to make you confident about your university success?
16. What advice would you give to L2 students about how to succeed at the university?
17. Would you like to add anything else to your comments?

Appendix M: Phase 3 - Administrator interview questions

This is our 3rd interview and I would like to discuss the following: the in-house tests, how has assessment change taken place and what are teachers' reactions to it, and the role of the Testing Office in the students' final outcomes.

1. Until last December, there were two integrated tests: Listening to Speak and Reading to Write. Has this changed?
2. Who provides interview topics for teachers in the new format of oral interviews?
3. During the last interview, you mentioned about benchmarking these tests with the Testing Office. How is the process going?
4. One of the issues, teachers pointed out was about the score interpretation in the rubric provided. What do you say about that?

5. Are there any validation studies carried out for the new tests?
6. Who is developing the tests at the lower levels of the program?
7. Have you received any feedback from the previous students regarding the new testing procedures?
8. What about the new PLOs and CLOs? Have any changes been made to these?
9. My impression from teacher interviews (with the new PLOs, CLOs) was that teachers were teaching/assessing these differently. For example, one teacher may do two or three research projects per semester while the other might do one, yet another may not include a research project. Have you thought of standardizing these? If yes, how? If not, why not?
10. Teachers mentioned that, overall, the influences of the ISTs were positive in the classroom teaching and learning. However, they also pointed out tensions in the current testing regime. For example, we have integrated skills tests, but writing is still the most dominant skill and secondly these tests are still marked by the outside testers who concentrate mainly on proficiency than content.
11. How do you see the communication between students, teachers and the Testing Office?
12. Do you think there will be further changes in these integrated skills tests?
13. Would you like to add anything else to your comments?

Appendix N: Ethics Clearance and Teacher Consent Forms



Research Compliance Office
 511 Tory | 1125 Colonel By Drive
 | Ottawa, Ontario K1S 5B6
 613-520-2600 Ext: 4085
ethics@carleton.ca

CERTIFICATION OF INSTITUTIONAL ETHICS CLEARANCE

A Change to Protocol for the following research has been **cleared** by the Carleton University Research Ethics Board-A (CUREB-A) at Carleton University. The researcher may proceed with their research. CUREB-A is constituted and operates in compliance with the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (TCPS2).

Ethics Protocol Clearance ID: Project # 100686
 Principal Investigator: Janna Dorothy Fox, Carleton University
 Research Team (and roles) (If applicable): Janna Dorothy Fox (Primary Investigator); Mrs. Poonam Anand (Student Research: Ph.D. Student)
 Study Title: Students' and Teachers' Views on the Assessment Practices in an English for Academic Purposes Program [Poonam Anand]
 Funding Source (If applicable):

Effective: **October 28, 2016**

Expires: **May 31, 2017.**

Please email the Ethics Coordinators at ethics@carleton.ca if you have any questions or if you require a copy with a signature.

CLEARED BY:

Date:

Andy Adler, PhD, Chair, CUREB-A

October 28, 2016

Shelley Brown, PhD, Vice-Chair, CUREB-A



Teacher/Educator Consent Form

Title: Students' and Teachers' Views on the Assessment Practices in an English for Academic Purposes Program

Date of ethics clearance: 8th April 2015

Ethics Clearance for the Collection of Data Expires: 26th January 2018

The researcher for this study is Poonam Anand, an instructor at the [REDACTED] program and a PhD student at the School of Linguistics and Applied Linguistics, at Carleton University. This study is a part of her PhD dissertation. The main goal of this study is to paint a vivid portrait of classroom assessment practices as experienced by students and teachers in the [REDACTED] program. The aim is not to generalize results to entire assessment practices; rather, to listen to [REDACTED] students' and teachers' voices, and to validate classroom assessment tasks by examining the matches and mismatches between students' and teachers' views.

I, _____, am invited to participate in a study on assessment practices in the [REDACTED] program. This purpose of this study is to help teachers to support students in classroom-based language assessment.

I have the right to end my participation in the study at any time. I can withdraw either by a written note or emailing the researcher or the research supervisor.

As a token of appreciation the researcher, Poonam Anand, will offer me a \$ 10 Tim Horton's gift card at the end of the study.

All information I share will remain strictly confidential and my identity will be protected. To protect my anonymity, I will be assigned an alphabetical code. Therefore my name will not be used in the dissemination of data. The information I share will be used only for the research purposes.

All research data, including audio-recordings and any notes will be encrypted and password-protected. Any hard copies of data (including any handwritten notes or USB keys) will be kept in a locked cabinet at Carleton University. Research data will only be accessible by the researcher and the research supervisor.

Once the dissertation is completed, all research data will be kept for five years and potentially used for other research projects on this same topic. At the end of five years, all research data will be securely destroyed. (Electronic data will be erased and hard copies will be shredded)

If I would like a copy of the finished dissertation, I will contact the researcher for an electronic copy. The Carleton University Research Ethics Board and the Office of Research Ethics and Integrity, University of [REDACTED] reviewed this project and provided clearance to carry out the research. Questions or concerns related to my involvement in this research may be addressed to:

REB contact information:
Professor Andy Adler, Chair
Professor Louise Heslop, Vice-Chair
Research Ethics Board
Carleton University
1325 Dunton Tower
1125 Colonel By Drive
Ottawa, ON K1S 5B6
Tel: 613-520-2517
ethics@carleton.ca

Researcher contact information:

Name : Poonam Anand
Department: SLALS
Carleton University
Tel: 613-614-7024

Supervisor contact information:

Name: Dr. Janna Fox
Department: SLALS
Carleton University
Tel: 613- 520-2600x 2046

Email: poonamanand@cmail.carleton.ca Email: JannaFox@Carleton.Ca

Signature of participant

Date

Signature of researcher

Date

Appendix O: Student Consent Form



Information Letter

Title of the study: "Students' Views on the Assessment Practices in the [REDACTED] Program"

Researcher contact information:

Name: Poonam Anand
Department: SLALS
Carleton University
Tel: 613-614-7024

Supervisor contact information:

Name: Dr. Janna Fox
Department: SLALS
Carleton University
Tel: 613- 520-2600x 2046

Email: poonamanand@cmail.carleton.ca

Email: Janna-Fox@carleton.ca

Dear Participant:

The study: You are invited to participate in a research study, "Students' Views on the Assessment Practices in the [REDACTED] program".

The researcher of this study is Poonam Anand, a PhD candidate at Carleton University. Using a questionnaire and focus group, I would like to explore students' understanding of the assessment practices at the [REDACTED] program.

There are two parts in this study. For part 1, I hope to administer the questionnaire to all Graduating Level students and for part 2 of the study, I hope to conduct a focus group with 6 – 8 students. Completing and submitting the questionnaire will take around 10 – 12 minutes and the focus group will be of 45 minutes duration.

Your participation

Your decision to complete this questionnaire will be interpreted as an indication of your consent to participate. The information that you will share will remain strictly confidential and will be used solely for the purposes of this research. Only my supervisor and I will have access to the research data. Anonymity is guaranteed since you are not being asked to provide your name or any personal information.

Students, who choose to fill in the questionnaire, will be asked if they are willing to participate in a focus group. This focus group will be audio recorded. To indicate their willingness to participate in a focus group, those students will provide their email contact address at the end of the questionnaire.

The information collected through this online questionnaire will be with the researcher in the digital format. Once the project is completed, all research data will be kept for five years and potentially used for other research projects on this same topic. At the end of five years, all research data will be securely destroyed.

There are no anticipated risks related to this research. Participating in this study is completely voluntary and if you wish, you may not fill in the questionnaire. Refusal to participate will involve no penalty or loss of benefits.

Also, filling in the questionnaire does not have bearing on who will be selected for the focus group. The focus group will consist of students from diverse cultural and linguistics backgrounds. If you would like a copy of the results from this study, you can reach me at poonamanand@cmail.carleton.ca.

If you have any questions with regards to the ethical conduct of this study, you may contact:
Professor Andy Adler, Chair
Professor Louise Heslop, Vice-Chair
Research Ethics Board
Carleton University
1325 Dunton Tower
1125 Colonel By Drive
Ottawa, ON K1S 5B6
Tel: 613-520-2517
ethics@carleton.ca

Thank you for your time and consideration.
Please print and retain a copy of this document for your records.

If you are interested in participating in the study, please click on the following link to complete the questionnaire

https://docs.google.com/forms/d/e/1FAIpQLSdgHPtabOQWFVdylydU33oOr_yYRjivCRsE5suCz-_cgoggsA/viewform

Appendix P: Examples of first and second cycle coding sheets with analytical memos

Example 1: Phase 1 student focus group

<p>Student 1: I think number 8, <i>Spending time in class practicing practice-exams</i> because practicing exams is most important.</p> <p>Group Leader: And why?</p> <p>Student 1: Because I noticed that there are a lot of kinds of English tests. For example, we practiced TOEFL listening however we went into the midterm exam and the format is quite different so we don't get used to it and I didn't get a very good mark. But I don't think that the articles I read (in the exam) were difficult, but maybe they were not in order, I was not familiar with them. I think this is the question.</p>	<p>Importance of exam format knowledge²</p> <p>Problem of content knowledge in decontextualized exams³</p>	<p>(Test context is key) Contextualized vs decontextualized exams¹</p> <p>Disconnect between classroom activities and testing exercises²</p>
<p>Memos</p> <p>2 & 3 – Very valid point. In case of other high-stakes exams like TOEFL/IELTS etc. practice material is readily available, but there isn't enough practice material available for the external high-stakes exam used at the EAP. This student is an experienced teacher in Japan and understands the importance of knowledge of the exam format in exam success. Clear exam format/context critical – need to explore how this interaction in the exam shifts after the exam change.....</p> <p>1 & 2 – This is one of the main differences between discrete point multiple choice exams and integrated skills tests. Also there is a disconnect between the classroom activities and the test tasks. May be in-class activities are more contextualized and content-based as oppose to test prep stand-alone reading or listening texts. Need to understand how any shift in context in the exam might help with test results and exam experience....</p>		

Example 2: Phase 1 teacher interview

<p>Leader: How much teacher input is there in these tests?</p> <p>Person 2: Uh.... Very little.</p> <p>Leader: Very little <u>because....?</u></p> <p>Person 2: Because the testing office sets the exams and not everybody is in favour of that but I am because they're the experts; They know what we're looking for, they know how to set fair, valid exam questions.</p> <p>Leader: Do you think these tests assess the program learning outcomes? I have an example of them here. If we talk about the final exam, do we see the learning outcomes?</p> <p>Person 2: I would have to say not entirely. I mean, <u>I would</u> have to go through these point by point.</p> <p>Leader: Well for example speaking would be completely out of this</p> <p>Person 2: Right, we don't test speaking. If we're looking at this "I can follow along conversations up to 7 minutes" the listening exams that the testing office sets are not that long. I mean they do ask them to make logical inferences so that's true, and yes [are supposed to find] main ideas. We do not ask them to take notes, we don't</p>	<p>Disagreement about role of testing office - some believe testing office is fair, valid (unbiased?), yet tests don't assess learning outcomes¹³</p> <p>Examples: speaking not tested; listening exams not as long as learning outcomes¹⁴</p>	<p>Role of Testing office = experts⁹</p> <p>Programmatic Tension resides within testing office versus the broader program¹⁰</p> <p>Programmatic tensions: academic prep, but not taught to take notes¹¹</p>
--	--	--

Example 3: Phase 2 student focus group

<p>Student 1: May I start?</p> <p>Interviewer: Yes</p> <p>Student 1: I think in my opinion the writing midterm exam <i>was very good</i>, it was easy, but in comparison with the speaking exam. The speaking exam was a little stressful for me because it's weird and (nerve racking) to record your own speech. At the same time you have time limitations and that's why I would prefer to talk to a person like face-to-face. I like to participate in real conversations. My ordinary speech when I talk to a real person and when I try to record my own speech by using like <u>RelanPro</u>, or other electronic equipment, it's different. During the exam you have pressure, you focus not only on your grammar and pronunciation. You focus on time limitation, the main ideas which you like to provide, give a complete answer. So that's why I would recommend to develop or to improve the system of speaking midterm exams.</p> <p>Interviewer: What about you?</p> <p>Student 2: I would agree with him. But the speaking was my biggest problem. As long as I know I can speak fast or good or better, but when it comes to the recording the person whom I'm talking with I don't see any reaction from him. I don't know what I'm saying. So that's why I think the speaking was one of the main problems on this test. Writing was good because it depends on us what we have studied. We can just take the ideas from our books and write it by our own words and it will help us to avoid plagiarisms for university. Writing was better compared to the speaking.</p>	<p>Very good – magnitude coding Positive views about the new RtW exams¹</p> <p>Issues with the speaking exam – machine talk vs human talk²</p> <p>Pressures of taking LtS exam³</p> <p>Speaking exam focus – not only language proficiency demands, but coherence and time limitations⁴</p> <p>Student recommendations to improve the LtS practicality⁵</p> <p>Students' focus shift from writing to speaking⁶</p> <p>Taking ideas from textbook and preparing for the exam⁷</p>	<p>Positive students' views with respect to RtW exam¹</p> <p>Human interaction vs. machine interaction in speaking²</p> <p>Value of academic skills such as knowledge of paraphrasing and plagiarism³</p> <p>Face and content validity of the writing exam⁴</p>
<p>Memos</p> <p>1- This student is new to the EAP and liked the new exam format, which is an integrated skills test. The previous GL mid-term exams were discreet point MC reading and listening exam.</p> <p>2 -. Check administrator's account for the speaking exam.</p> <p>4- Text book supports the exam content which helps in better test preparation and also the student feels now the onus</p>		

Example 4 Phase 3 teacher interview

<p>A: Well the research project was an individual work. The curriculum says that of our in class mark, 20% of it has to go to the individual project. But there's nothing in the curriculum that says how much the group work should be. So for the first maybe half of the term I had one assignment that was a marked group assignment, they had to write a summary together. And they did presentation but I didn't mark them on it because it was informal. I didn't mark the speaking, they did lots of practice with me and my colleague but we didn't spend time marking it, we would give them a general feedback. The group work is a little bit harder to fit in and we have to be a little bit careful about the percentage we assign to it because you have to look at actual course credits and how much of their mark would ever be assigned to group related speaking. I think group work is really important to have and encourage but not always for marks. And sometimes its even more productive when it's not for marks. The challenge I had this term is I had 23 Chinese students, so their</p>	<p>16. Lots of skills practice not being marked</p> <p>17. Teacher picks what work to be marked</p> <p>18. Limited by the curriculum assignation of term marks</p> <p>19. Juggling various externally imposed limitations</p>	<p>13. Unknown as to which assignments to be marked and which not</p> <p>14. External factors shaping internal assessment reliability and validity</p>
<p>MEMOS</p> <p>The testing regime is trying to "manufacture" simulated ways to assess students and teachers are wanting to provide tasks that students would likely see in the university classroom – how is it that the testing regime has created this level of disconnect with the teachers? What effect would this have on <u>washback</u> moving forward on teachers? Students? Assessors? </p>		

Appendix Q: Sample division of categories

(Procedural coding example)

Teaching/Learning Curriculum

Issues with the new exam + are workload.

Positive workload issues of with the new exam.

First-Cycle Codes	Second-Cycle Codes		
Very good – magnitude coding	Positive students' views with respect to RW exam ¹	Time constraints and unrealistic demands on the speaking test.	Summary writing important
Positive views about the new RW exams ¹	Human interaction vs. machine interaction in speaking ²	Exam tasks mirrored classroom tasks	Thematic learning vs. decontextualized learning
Issues with the speaking exam – machine talk vs human talk ²	Value of academic skills such as knowledge of paraphrasing and plagiarism ³	Use of previous knowledge to generate new ideas	Situated cognition
Pressures of taking LIS exam ³	Face and content validity of the writing exam ⁴	Situated cognition	Role of memorizing text book content to pass the exam
Speaking exam focus – not only language proficiency demands, but coherence and time limitations ⁴	No predictive validity of the speaking exam	Use of outside sources useful vs. previous session teachers' and students' perceptions of irrelevance of outside materials.	Task demands vs. rubric knowledge
Student recommendations to improve the LIS practicality ⁵	Speaking to a wall – mechanical aspect of speaking exam	Students value classroom learning	Paraphrasing and citation vs. rubric knowledge
Students' focus shift from writing to speaking ⁵	Role of mediation – technology as ZPD	Role of memorization vs. knowledge of content/critical thinking	Incomplete citations in the exams – exam content issue
Taking ideas from textbook and preparing for the exam ⁷	Thematically- linked test tasks	Writing summaries were regularly practiced in class	Problem/Project-based learning of summary
Writing exam tasks mirror school life writing but not the listening example of situated learning	Situated cognition	Usefulness of peer-feedback and group-discussion	Textbooks helped in promoting academic skills
Topic familiarity of writing tasks	Inclusion of summary writing as an important test-taking skill	Student culture of memorizing text book chunks	Issues with textbook – one sided view
Content knowledge useful in writing exam	Indirect learning of paraphrasing skills	Academic vs. general English – more from a	Rubric knowledge important/helpful for

Appendix R: First and second cycle coding with emerging categories and themes

Student focus groups (Phase 1 & Phase 2)

Phase 1			Phase 2		
First-Cycle Codes	Second-Cycle Codes	Emerging Categories and Themes	First-Cycle Codes	Second-Cycle Codes	Emerging Categories and Themes
Rank ordering different exam-related activities ¹	(Test context is key) Contextualized vs decontextualized exams ¹	Positive/Neutral washback of the previous test	Very good – magnitude coding	Positive students' views with respect to RtW test ¹	Positive Washback of the new test
In class work given priority	Disconnect between classroom activities and testing exercises ²	Preparing for the ExitExam is same as preparing for any standardized High-Stakes proficiency exam	Positive views about the new RtW exams ¹	Human interaction vs. machine interaction in speaking ⁷	Positive students' views with respect to RtW exam ¹
Importance of exam format knowledge ²	Whole versus incremental learning	Out-of class support for weaker students ¹¹	Issues with the speaking exam – machine talk vs human talk ²	Value of academic skills such as knowledge of paraphrasing and plagiarism ³	Face and content validity of the writing exam ⁴
Problem of content knowledge in decontextualized exams ³	test driven material priority	Peer-feedback, self-correction and importance of 'modeling'	Pressures of taking LtS exam ³	Face and content validity of the writing exam ⁴	Inclusion of summary writing as an important test-taking skill
(whole versus component aspect of learning)	Undervalue of speaking skill ³		Speaking exam focus – not only language proficiency demands, but coherence and time limitations ⁴	No predictive validity of the speaking test	Importance of speaking skills in class
Developing language proficiency ⁵	Context between test/classroom matters	Negative Washback of the previous test	Student recommendations to improve the LtS practicality ⁵	<i>Speaking to a wall – mechanical aspect of speaking exam</i>	Thematic learning vs. decontextualized learning
Developing exam construct knowledge ⁵	Hierarchy of skills promoted in program	(Test context is key) Contextualized vs decontextualized exams ¹	<i>Students' focus shift from writing to speaking⁶</i>	Role of mediation – technology as ZPD	Task demands vs. rubric knowledge help in summarizing
Speaking not given importance because of the exam format even though group work is helpful ⁶	Hierarchy of skills promoted in program	Disconnect between classroom activities and testing exercises ²	Taking ideas from textbook and preparing for the exam ⁷	Thematically- linked test tasks	Paraphrasing and citation vs. rubric knowledge
Reading texts in class aren't similar to the reading texts in exam – in length or in content ⁷	Exam format matters	test driven material priority	Writing exam tasks mirror school life writing but not the listening example of situated learning	Situated cognition	Promotion of Problem/Project-based learning
Short answers aren't important because of the exam format ⁸	Undervalue of summarizing, paraphrasing and referencing skill in terms of exam preparation ⁴	Undervalue of speaking skill ³	Topic familiarity of writing tasks	Inclusion of summary writing as an important test-taking skill	Rubric knowledge important and helpful for performance tests
Short question answers in lower levels of ELP ⁹	Hidden curriculum: ease and efficiency of getting the work done	Hidden curriculum: ease and efficiency of getting the work done	Content knowledge useful in writing exam	Indirect learning of paraphrasing skills	Positive and continuous feedback from teachers
Irrelevance of clozed passages in reading ¹⁰	Continuity of skill development	Exam format matters - test items format given priority e.g. non-existence of short-answers	Use of technology	Importance of speaking skills in class	Student satisfaction with the program

Legend: Themes
 Categories
 Codes