The Government of Canada's Second Language Evaluation, Test of Written Expression:
Exploring Validity

by

Ellen Nearing

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

Master of Arts

In

Applied Linguistics and Discourse Studies

Carleton University

Ottawa, Ontario

© Ellen Nearing 2020

Abstract

This study aimed to contribute to the literature on the Canadian government's Second Language Evaluation (SLE) by exploring the concept of validity within the context of the Test of Written Expression (TWE). The case-study approach combined stakeholder perspectives of three employee-participants and two teacher-participants along with a verbal protocol of a shortened practice test. The findings indicate that there may be evidence that threatens the validity of the inferences drawn from the TWE. There is additional evidence of negative consequences in terms of washback and impact of the SLE. These findings shed light on issues with the implementation of the Official Languages Act within the public service, suggesting a bottom-up approach to address these issues.

Dedication

To the Dream Team

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Janna Fox. I feel honored to have had the chance to learn from her and work with her this past year. Secondly, I would like to thank my co-supervisor, Dr. Guillaume Gentil for his expertise on bilingualism and attention to detail. Janna and Guillaume are truly a dynamic duo and I have learned so much from them. Special appreciation is extended to my thesis committee members, Dr. Angel Arias, Dr. Rachelle Vessey, and Dr. Wood for their kind words about my work, useful feedback, and stimulating discussion that offered a variety of perspectives on the topic.

I would like to extend my gratitude to the School of Linguistics and Language Studies at Carleton University; the program has enabled me to reach levels of knowledge, understanding, and practice beyond what I had anticipated. I would like to extend a grateful nod my classmates Hiba Fazl and Sev Smith-Halverson for the insightful discussions, and continuing empathy in getting through our projects. I would also like to extend a special thanks to the participants of this study.

I would like to thank my mother, Grace, and father, Kevin, who have always supported me in whichever road I have decided to follow.

Finally, I would like to thank my husband, Dean, and my dog Bruno, who had to endure endless (and painful) hours of listening to me talk about this project.

Table of Contents

Abstract	ii
Dedication	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	ix
List of Figures	x
Acronyms and Abbreviations	xi
Glossary of Terms	xii
Chapter One: Introduction	1
Chapter Two: Literature Review	6
2.1 Validity	6
<i>Messick</i>	12
<i>Cizek</i>	13
<i>Kane</i>	16
2.1.1 The Construct of Writing in the Public Service Commission	20
<i>Level A</i>	22
<i>Level B</i>	22
<i>Level C</i>	23
2.2 Washback and Impact	24
2.2.1 <i>Washback</i>	24
2.2.2 <i>Impact</i>	30
2.3.3 <i>Washback, Impact and Validity</i>	34
2.3 Threats to Validity	39
Chapter Three: Methods	48

3.1	Research Design.....	48
3.2	Setting	49
3.3	Participants.....	49
3.4	Researcher Positionality.....	51
3.5	Instruments.....	51
3.5.1	<i>Practice Test</i>	51
3.5.2.	<i>Interview - Employees</i>	52
3.5.3.	<i>Interview - Instructor.</i>	52
3.6	Procedures.....	52
3.6.1.	<i>Verbal Protocol</i>	53
3.6.2.	<i>Interviews</i>	54
3.7	Analysis.....	54
Chapter Four: Findings and Discussion.....		56
4.1	Results.....	56
4.1.1.	<i>Verbal Protocol</i>	56
4.1.2.	<i>Interview Data</i>	57
4.2.	Employee Participant Profile	58
4.3	Comparison of VPA Results.....	60
4.3.1.	<i>Correct Answers</i>	61
4.3.2.	<i>Incorrect Answers</i>	64
4.3.3.	<i>Summary of VPA</i>	68
4.4.	Research Question 1: What is the Test of Written Expression (TWE) assessing?	
	Are there any factors undermining the validity of this test? If so, what are they?	69
4.4.1	<i>Summary of RQ1</i>	74

4.5. Research question 2: do the test tasks on the TWE correspond to those in the TLU domain?	75
4.5.2. Rodney.....	79
4.5.3. Abigail.....	81
4.5.4. The teacher's perspective	82
4.5.5. Summary of Research Question 2	83
4.6. Research question 3: what, if any, are the washback effects of the test?	85
4.6.1. Holly.....	86
4.6.2. Rodney.....	89
4.6.3. Abigail.....	91
4.6.4. Bruno.....	92
4.6.5. Willow	95
4.6.6. Summary of research question 3.....	98
4.7 Research question 4: what is the impact of the SLE?	100
Theme 1: SLE is a Business	100
Theme 2: SLE is an investment	101
Theme 3: SLE is a gatekeeper.....	103
Theme 4: SLE is a game.....	104
4.8 Are there any factors undermining the validity of the test? If so, what are they?	
107	
Chapter 5: Conclusions	111
5.1 Summary and Implications of Findings.....	111
5.2 Limitations	113
5.3 Future Research	113
References.....	115
Appendix A: Ethics Approval Certificate.....	121
Appendix B: Practice Test	122
Appendix C: Semi-Structured Interview Questions – Employees.....	139

Appendix D: Semi-Structured Interview Questions – Teachers	140
Appendix E: Consent Form – Employees.....	141
Appendix F: Consent Form - Teachers	144
Appendix G: Qualitative Coding Scheme.....	147
Appendix H: Primary Researcher Initial Coding Sample.....	150
Appendix I: Second Coder Initial Coding Sample	151

List of Tables

Table 1: Facets of validity (Messick, 1987, p. 17).....	12
Table 2: Table 2: Qualification Standards – Official Languages (TBS, 2016, sec. 4).....	21
Table 3: Characteristics of employee participants	50
Table 4: Participants' scores out of 30	61

List of Figures

Figure 1: Cizek's (2012) model (p. 36)	15
Figure 2: Kane's Chain of Inferences (Bachman, 2005)	17
Figure 3: Levels involved in hierarchical decisions (Markee, 1993, p. 234)	31
Figure 4: Question 12	61
Figure 5. Question 15	62
Figure 6. Question 21	62
Figure 7. Question 13	63
Figure 8. Question 20	64
Figure 9. Question 22	65
Figure 10. Question 1	66
Figure 11. Question 5	67

Acronyms and Abbreviations

CH	Canadian Heritage
CSPS	Canadian School of Public Service
IUA	Interpretation/Use Argument
L1	First Language
L2	Second Language
NCR	National Capital Region (Ottawa/Gatineau, Canada)
OCOL	Office of the Commissioner of Official Languages
OLA	Official Languages Act
PSOLAR	Public Service Official Languages Appointment Regulation
PSC	Public Service Commission
PSPC	Public Services and Procurement Canada
SLE	Second Language Evaluation
TWE	Test of Written Expression
TLU	Target Language Use Domain

Glossary of Terms

Construct: the ability a test purports to or intends to measure through its design (Cronbach & Meehl, 1955).

Construct Irrelevant Variance: factors outside of language ability that negatively affect the score (Messick, 1994, 1996; Cizek, 2012; McNamara & Ryan, 2011).

Construct Underrepresentation: a lack of clarity or explicitness of the construct (Messick, 1994; 1996)

Construct validity: based on three qualities: the theories that define the construct, the justification the construct is measured through test design, and the value of this construct as represented in the test (Bachman & Palmer, 1996; Cronbach & Meehl, 1955; Messick, 1987).

Content Validity: the correspondence between the test tasks and the TLU domain (Bachman & Palmer, 1996).

Direct Assessment: a test that requires the test-taker to use the language productively (i.e., measuring speaking by asking test takers to speak; measuring writing by asking test-takers to write.).

Face Validity: the test-taker's perspective on the efficacy of the test's design to measure the skill it purports to measure or as representative of the domain to which it applies (Stiggins, 1982).

Fairness: features related to the technical qualities of test design (McNamara & Ryan, 2011).

Impact: the effects of testing on the larger societal level, such as in educational institutions and employment (Bachman & Palmer, 1996; Cheng, 2014; McNamara, 2006).

Indirect Assessment: a test that requires the test-taker to judge the appropriateness of the language (e.g., a multiple-choice test to assess writing) (Stiggins, 1982).

Interpretation/Use Argument: “a network of inferences and assumptions leading from the test performances to the conclusions to be drawn and to any decisions based on these conclusions” (Kane, 2013, p. 8).

Justice: features related to qualities of test use (McNamara & Ryan, 2011).

Language profiles: describes the employees’ level of bilingual proficiency in Reading, Writing and Oral Interaction. From highest to lowest, the levels range from X-A-B-C-E (TBS, 2016).

Nomological Network: a set of laws that guide construct validity (Cronbach & Meehl, 1955) (see Construct Validity).

Official Languages Act (OLA): An Act that represents the status and use of the official languages of Canada (i.e., English and French). Official languages are used in government administration, judiciary proceedings and legislation.

Office of the Commissioner of Official Languages (OCOL): an agent of Parliament promotes Canadian linguistic duality and manage the implementation of the Official Languages Act.

Policy on Official Languages: a policy to assist compliance with and ensure successful implementation of the OLA.

Public Service Official Languages Appointment Regulation (PSOLAR): a regulation that details the language requirement of positions within the public service.

Utilization Argument: validation of intended test use by gathering evidence consisting of claims from the validity data; warrants, backing, and rebuttals are collected in terms of relevance of the score, utility to the TLU, intended consequences of the assessment and the decision. Sufficient evidence is required to support the decision made by the assessment (Bachman, 2005)

Validity: the extent to which the inferences drawn from a test lead to meaningful interpretations that, in an ideal world, result in fair decisions (Kane, 2010; Kunnan, 2004; Messick, 1987, 1994, 1996; Moss, 1998, 2013; McNamara & Ryan, 2011).

Washback: the narrow influence of a test at the classroom level on teaching and learning (Alderson & Wall, 1993).

Chapter One: Introduction

French and English are recognized in the Canadian constitution as the two official languages of Canada¹; however, promoting and maintaining French-English bilingualism among the Canadian population continues to be met with challenges. Drawing on statistics from 2007, Gentil, Bigras and O'Connor (2010) reported that country-wide, 17.4% of Canadians had bilingual English-French conversational knowledge; 57% of Canadians were English-dominant; 22.1% were French-dominant. Statistics reported ten years later (Statistics Canada, 2017) indicate that, “the rate of English–French bilingualism in Canada was 17.9% in 2016, the highest proportion ever” (p.1). However, over the decade, the statistics do not reflect a dramatic increase in bilingualism, raising concerns about the effectiveness of bilingual initiatives in Canada.

Another issue of concern is the unequal distribution of English-speaking to French-speaking Canadians. Regional differences show the highest number of French-speaking Canadians are concentrated in Quebec, New Brunswick, and the National Capital Region (NCR) of Ottawa. These differences likewise extend to the retention of bilingualism. When analyzing different groups of children aged 5-17 in Quebec and the rest of Canada, Turcotte (2019) reports that one of the groups located outside of Quebec were no longer considered bilingual 10 years later; in comparison, children growing up in Quebec tend to retain their bilingualism and therefore contribute to the higher rate of bilingualism in the province (Turcotte, 2019). The challenge then for the Canadian public service is to ensure equal representation of both official languages within the workforce.

¹ Since the adoption of the constitution, Inuktitut has been recognized as an official language in Nunavut.

However, according to government publications (e.g., Statistics Canada, 2017), the numbers related to bilingualism have not changed much over the last ten years.

French-English bilingualism in the Canadian public service is an effort strengthened by the Official Languages Act (OLA), a policy originally implemented in 1969. Since its enactment, the OLA has subsequently seen several rounds of strengthening in 1988, 2005, with its last amendment in 2017. The act stipulates that English and French have equal status in federal institutions and functions within the public service must reflect this standard. Secondly, it supports the development of linguistic minority communities; thirdly, it prescribes that certain duties within the federal public service must accord to the official languages of Canada by offering equal opportunity for employment and advancement; in addition, Canada-wide regardless of first language, Canadians have the right to access services in the official language of their choice. One outcome of the act is that it has resulted in institutionalized rather than individualized bilingualism. In order to foster a more bilingual workforce, the public service either hires bilingual employees, or invests in language training for those who are not (Gentil et al., 2010; Department of Justice [DJ], 2005; Treasury Board Secretariat [TBS], 2017). This has had significant effects on Canadian public service employees. For example, English-French bilingualism leads to better employment and higher salaries (Turcotte, 2019), months are dedicated to language training, and the stress of achieving the language requirement is a lingering concern (Bessette, 2005; Gentil et al., 2010). Furthermore, the government invests a considerable amount of money into language training, and other official languages initiatives (see for example Canadian Heritage [CH], 2018).

Given the hierarchical nature of the federal government, there are a number of policies that trickle down to ensure that the OLA is implemented. Subsidiary to the OLA is the Policy on Official Languages, which oversees Parts IV (Communications with and Services to the Public), V (Language of Work), VI (Participation of English-speaking and French-speaking Canadians), and VII (Advancement of English and French). Canadian Heritage (CH) and Treasury Board Secretariat (TBS) are the two main departments responsible for managing compliance from other departments and agencies with the OLA and overseeing implementation (Bessette, 2005; TBS, 2017). Furthermore, deputy heads are responsible for ensuring compliance with the official languages regulations within each institution that subscribes to the relevant parts of the OLA. The Office of the Commissioner of Official Languages (OCOL) is responsible for carrying out investigations to ensure the departments are following the directives as set out in the OLA. Finally, the Public Service Commission (PSC) plays an integral role regarding implementation by facilitating second language (L2) testing of government employees in both official languages, also known as the Second Language Evaluation (SLE). The effects of the SLE is the focus of this study.

Despite the many policies in place that recognize the equal status of English and French, as well as the provisions for a bilingual workforce within the public service, there have been issues with implementation, as a number of reports have presented concerning numbers regarding the development and progress of Canadian bilingualism. For example, between 2000 and 2017, no significant changes had been reported with respect to part V (i.e., language of work [CH, 2018; OCOL, 2019]). To address these issues, the government has recently published its most recent overhaul of the OLA which will

comprise of roughly \$2.5 billion in investments over five years, from 2018 to 2023 (CH, 2018). According to the new action plan, three key initiatives will be emphasized overall: strengthening communities, strengthening access to services, and promoting a bilingual Canada. Of that investment, \$500 million will be spent specifically on L2 development initiatives to improve French-English duality. Given that achieving a bilingual workforce has a major impact on both the government and the employees, it should be the case that the tests are yielding accurate data, as well as supporting the implementation of bilingual initiatives. It is these concerns that provide the rationale for this study.

As a language teacher who has taught in the context of preparing employees for their SLE, I have experienced the washback effects from the tests. *Washback* is essentially the influence of a test on teaching and learning (Alderson & Wall, 1993). Previous research on the SLE has examined in-house language training programs and documented the prevalence of English-dominant work environments. However, the effects of the SLE do not end in the language classroom; the *impact* of the SLE extends much farther. Thus, the rationale for this study is threefold. First, an exploration into the Test of Written Expression (TWE) begins a conversation about validity in language testing within the context of the SLE, an area that would benefit from more research. Secondly, since the SLE and language training are closely connected, investigating the ways in which the test is affecting language preparation classes further adds to the discussion of the SLE. Finally, raising awareness of the consequences of testing that extend beyond the individual and into broader Canadian society may help inform future policy-makers, should the SLE be revisited in the future.

The overarching aim of this study was to explore the construct of the SLE's TWE, if there was anything undermining the validity, as well as the consequences of the test in the narrower sense (i.e., washback) and the broader one (i.e., impact). This study was guided by the following research questions:

1. What is the Test of Written Expression (TWE) assessing? Are there any factors undermining the validity of this test? If so, what are they?
2. Do the test tasks on the TWE correspond to those in the TLU Domain?
3. What, if any, are the washback effects of the test?
4. What is the impact of the SLE?

To address the research questions, this thesis is divided into five chapters. Chapter one has described the background, rationale, aims and research questions central to this study. Chapter two provides additional information on the SLE and a review of previous research concerning the concepts of *validity*, *washback* and *impact* in language testing. Chapter three describes the research methodology and provides information on the setting, participants, instruments and procedures as well as methods for analysis. Chapter four details the findings and discussion in relation to the research questions. Chapter five summarizes the findings of the study, addresses the limitations, and suggests areas for future research.

Chapter Two: Literature Review

There are four main themes on which the following literature review is based: (1) validity, (2) washback, (3) impact, and (4) threats to validity. These themes were guided by the research questions and are each addressed in the literature review that follows.

2.1 Validity

Validity is a central concept in this study, and also one that has been the focus of much attention and debate in the language testing literature. This section will focus on the concept of validity as it relates to language testing within the context of the current study. As stated in the previous section, SLE testing is mandated by the PSC in an effort for departments to meet the requirements outlined by the OLA (Bessette, 2005; Gentil et al., 2010). Kane (2013) explains how a test developed by one person and used for their own personal reasons is not likely to be challenged for its validity. However, for a test such as those in the SLE – that is designed, developed, and implemented by the public service *for* the public service – the validity of the inferences of such a test is of public concern (Kane, 2013). This section will first briefly discuss the contentious nature surrounding the term *validity*; following this discussion, general qualities within the field of validity research will be considered. Lastly, several researchers and their approaches to validity and validation research will be shared.

It is important to note from the outset that the term *validity* is troublesome; each researcher holds differing beliefs as to what *validity* refers to, what evidence to consider when creating a validity argument, as well as to what extent consequences should be considered in validity and validation research. Newton and Shaw (2016) contrast three main schools of thought on the term: *traditionalists*, *new liberals*, and *conservatives*.

Traditionalists are concerned with the predictive power of the *validity coefficient* and numerical measurements; this would be considered a narrow definition of validity which looks at aspects related to test design (McNamara & Ryan, 2011). New liberals take a broader perspective on validity and include intended and unintended consequences that result from test-based decisions (Newton & Shaw, 2016). The conservative standpoint, much like the traditionalists, returns to the narrower perspective of the instrument and the intended measurement of said instrument. Conservatives view broader issues of consequences of test use as a separate matter altogether (Newton & Shaw, 2016; Cizek, 2012). Taking these definitions by Newton and Shaw (2016) into consideration, this paper takes a new liberalist perspective with respect to validity, as it is exploring the context of the SLE which is rooted in Canadian policy with consequences that stem beyond the individual level of the test-taker. This paper defines *validity* as the extent to which the inferences drawn from test scores lead to meaningful interpretations that, in an ideal world, result in fair decisions (Kane, 2010; Kunnan, 2004; McNamara & Ryan, 2011; Messick, 1987, 1994, 1996; Moss, 1998, 2013).

Validity is not so much a feature of a test, but rather developed through argumentation after gathering enough evidence to support claims (Bachman, 2005; Cizek, 2012; Kane, 2013; Messick, 1994, 1996). Research in this area is often divided into two categories: (1) qualities related to test design and (2) consequences that result from test use. The latter has received considerably more attention since the 1990s, when studies of *washback* and *impact* grew in interest, spear-headed by Samuel Messick and Lee J. Cronbach, two prominent figures in promoting the awareness of *consequences* of test use (Cheng, 2014; Moss, 1998). Conversations of fairness, ethics and justice in language

testing have grown in popularity as an attempt to rationalize uses of tests (Bachman, 2005; Kane, 2010; Kunnan, 2004; McNamara & Ryan, 2011). While validity as a concept has evolved over time and the debate is still ongoing in the literature (cf. Bachman, 2005; Cizek, 2012; Kunnan, 2004; McNamara, 2007; McNamara & Roever, 2006; Messick, 1987, 1994; Shohamy, 2005), there are several agreed upon features of contemporary validity theories and the validation process (see, for example, Chapelle, 2021; the Standards of Educational and Psychological Testing, 2014).

First, it is widely agreed that test scores have meanings and interpretations, and it is these interpretations that are validated, not a test itself (Kane, 2013; McNamara & Roever, 2006; Messick, 1987). The scores from a test ultimately reflect the *construct*, which is an unobservable characteristic (Cronbach & Meehl, 1955), hence the literature often refers to *inferences* drawn from interpretations. The second agreed upon feature of validity is that it is a varying quality of a test; tests can be more or less valid depending on the inferences drawn from them (Bachman & Palmer, 1996; Cizek, 2012; Chapelle, 1999; Cronbach & Meehl, 1955; Kane, 2013; McNamara & Roever, 2006; Messick, 1987, 1994). Thirdly, *validity* is different from *validation*, the latter of which is an ongoing process that combines evidence and justification for the appropriateness of test use (Bachman, 2005; Cizek, 2012; Kane, 2013). As new evidence and theories emerge, the validity judgement has the potential to change, which necessitates the need for ongoing validation (Bachman, 2005; Cizek, 2012; Kane, 2013).

Researchers place the emphasis on *construct validity* to gather evidence and evaluate the validity judgement (Bachman & Palmer, 1996; Cronbach & Meehl, 1955; Messick, 1987). Construct validity is based on three qualities: the theories that define the

construct, the justification the construct is measured through test design, and the value of this construct as represented in the test (Bachman & Palmer, 1996; Cronbach & Meehl, 1955; Messick, 1987); it is the closest conceptualization of a unified theory of validity. Since current views emphasize construct validity, this then begs the question of how to define the construct (Bachman, 2007; Bachman & Palmer, 1996; Chapelle, 1991).

Cronbach and Meehl (1955) define it as “some postulated attribute of people, assumed to be reflected in test performance” (p. 283). Messick (1987) states constructs are “inferences about underlying processes or structures” (p. 5). Bachman and Palmer (1996) define it as evidence of the specific language ability the test intends to measure. Bachman (2007) highlights how defining the construct is not so simple by analyzing the history of theoretical approaches to construct definition throughout the years. For example, one of the older approaches to defining and testing the construct is the *skills and elements* approach, which attempts to measure characteristics of language skills; test tasks may consist of discrete points of grammar. On the contrary, a performance assessment values how a language speaker would perform in a real-life situation, and as such, the test tasks would be similar to real-life tasks (see Bachman, 2007 for an overview of approaches from 1960 to the present). Therefore, the way the construct is defined will have an influence on test design, as well as how the construct validity argument is supported.

Another strategy to validation is the argument-based approach popular among researchers such as Bachman (2005) and Kane (2013). This approach does not rely as much on theoretical background as construct validity, but rather focuses on claims and gathering evidence to support the claims.

Several models have been proposed: Cronbach and Meehl (1955) were some of the first to discuss features of construct validity, Messick (1987) offers a widely accepted framework for his unified theory of validity, and Kane (2013) suggests his chain of inferences model. Displeased with Messick's model, Cizek (2012) puts forth a framework that highlights the two separate processes of validation of test design followed by the validation of test use. Bachman and Palmer (1996) suggest their own validity framework by discussing test usefulness. Finally, Bachman (2005) attempts to draw a clear link between test validity and test use through the utilization argument. What follows is a discussion of these noteworthy validity researchers to offer a comparison of perspectives.

Cronbach and Meehl (1955) discuss thoughts on validity popular at the time by identifying four types of validity: predictive, concurrent, content, and construct. Predictive and concurrent validity are *criterion-oriented* measures, which aim to assess underlying abilities, such as grammar knowledge on a language test, or an ideal performance with which to compare a workplace performance assessment. Concurrent validity is a measure of the quality at the time of testing. For example, in the case that there are two different tests (e.g., a multiple-choice test and a dictation test) that measure the same purported feature of grammar, high validity would be a consistent measurement across both tests. Predictive validity is the extent to which the test performance can predict how the individual will behave in the real-world context. Content validity refers to the test items and how they relate to the "universe in which the investigator is interested" (p. 282), also known as the target language use domain (TLU). In the case of the SLE, the TLU would be language for completing work-related tasks.

Cronbach and Meehl (1955) were the first to define construct validity as a characteristic that a test is *interpreted* to measure; it is a quality which is not “operationally defined” (p. 282). The construct is “some postulated attribute of people, assumed to be reflected in test performance” (p. 283). Cronbach and Meehl (1955) add that the construct is an indirect measure which depends on the position of the test creator and how they define the construct as claimed to be reflected through their test design. On construct validation, they add, “We seek to specify how one is to defend a proposed interpretation of a test; we are not recommending any one type of interpretation” (pp. 283-284). Here, Cronbach and Meehl (1955) are asserting that in order to create an argument for construct validation, it is not enough to merely look at the content of the test, but to gather a body of evidence, and to examine the assumptions of the test along with the context to which the test applies. Additionally, justifying the construct validity of the test is not a yes or no option; constructs have to be specified in order to be properly investigated. This specification is supported by the theoretical background that roots the construct, as decided by the creators of the tests. As stated earlier, construct validity is based on three qualities: the theories that define the construct, the justification that this construct is measured through test design, and the value of this construct as represented in the test. Cronbach and Meehl (1955) refer to this as a “nomological network” (p. 290), and state that this is a set of laws that guide construct validity. In the case of negative evidence, three interpretations can be made: (1) the test did not measure the construct, (2) the theoretical framework is incorrect, or (3) the design was not such that it facilitated an accurate measure of the construct. Should any of those three options arise, they state that the test ought to be redesigned and gather new evidence for validation. Tests themselves

cannot be validated; inferences and interpretations based on evidence, however, can. This is a belief in the field that is still widely held today and one that is greatly reflected in the literature.

Messick

Another noteworthy researcher on the subject of validity is Samuel Messick who has influenced many works on the subject of validity, test fairness, and consequences of test use (see Cizek, 2012; Kunnan, 2004; Kane, 2013; Shohamy, 2005, to name a few). One of the criticisms of validity theory at the time was that it was often presented as a *reliability coefficient* which emphasizes the predictive ability of tests, and not enough on the consequences that stem from test-based decisions, thus Messick (1987) presented his progressive matrix (see Table 1) as a unified theory of validity which reflects his thinking.

Table 1: Facets of validity (Messick, 1987, p. 17)

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/Utility
Consequential Basis	Value Implications	Social Consequences

The matrix is divided into two columns — *test interpretation* and *test use*— intersecting with two rows —*evidential basis* and *consequential basis*. Going from left to right, the evidential basis of test interpretation and test use refers to the evidence required to support the validation of construct validity as well as its relevance to the targeted domain; such evidence can be gathered from a comparison of the tasks and abilities needed in the real world to the test items which should reflect the construct. Additionally, regardless of

how well supported the construct may be, to Messick there must be a direct link of the relevance of the test to the targeted domain.

The consequential basis of test interpretation and test use include value implications and social consequences. The value implications section accounts for the theoretical backing required for validating construct validity in addition to the ideologies reflected in such theories. The consequential basis for test use includes social consequences which would require an examination of the potential or actual, intended or unintended, consequences of test use (Messick, 1987; 1996). In the case of the SLE, a consequence would be a test-based decision allowing or impeding job placement. Messick (1987) explains that his matrix ought to be thought of as progressive to ensure construct validity is present in each cell, adequately representing its prevalent role in validity theory. Furthermore, Messick emphasizes that it is not enough to have a good instrument; it needs to be used appropriately. Since Messick's work, many researchers have studied consequences of testing both on the smaller scale of washback (see for example, Alderson & Hamp-Lyons, 1996; Alderson & Wall, 1993; Hamp-Lyons; 1997; Wall & Alderson, 1993) and the larger scale of impact (Chalhoub-Deville, 2016; McNamara 2007; McNamara & Roever, 2006; McNamara & Ryan, 2011; Shohamy, 2007; Moss, 1998; 2013), the topic of which is elaborated in the *Washback and Impact* section (see section 2.2).

Cizek

In comparison with Messick and Cronbach who would be considered *new liberalists* by Newton and Shaw's (2016) definition, Cizek subscribes to a more conservative approach, which views validity as a matter of test interpretation. Hence

Cizek's (2012) major issue with Messick's framework is it collapses two ideas — validity as a feature and validation as a process — into one conception; to Cizek, these are two independent processes of interpreting score meaning and justification for test use. While Cizek recognizes that each cell of the matrix represents integral considerations for testing practices and building a validity argument, he views the cells as separate concerns, not united ones. Additionally, Cizek has several reservations of specific features of Messick's matrix. For example, he questions the extent to which the consequential basis of test use (i.e., social consequences) should be a source of validity evidence; moreover, he does not view the relevance of social consequences as having weight when it comes to the validity of score interpretations because again, validating test design does not justify test use. However, Messick is firm that both features rely on each other; it would not be possible to make a good argument for appropriate test use without meaningful interpretations from merely test design. Another criticism of Messick has been questioning how construct validity and social consequences are related, stating that issues of ethics and fairness in testing fall under the justification for *use*, *not* validity. Yet, Messick (1994) would say that it is possible to use social consequences as evidence that either supports or refutes construct validity in the case that unfairness or bias results from an invalid test.

To address such issues, Cizek provides his own definition: "Validity is the degree to which scores on an appropriately administered instrument support inferences about variation in the characteristic that the instrument was developed to measure" (p. 35). According to Cizek, tests themselves do not define the construct, nor is the construct a test; he argues that a well-developed test will match what it was intended to measure. As we will see in section 2.3 on threats to validity, regardless of how well-developed a test

is, there will always be some elements of what Messick (1996) termed construct underrepresentation and construct irrelevant variance.

Cizek's model demonstrates his thinking of how the validation process consists of two distinct procedures: (1) validation of score meanings and (2) justification for test use (Figure 1).

Figure 1: Cizek's (2012) model (p. 36)

Cizek emphasizes that first, the test itself and the intended score meanings are validated, then the usefulness of the test to the context it applies should be examined in two distinct processes. Once each step of the process satisfactorily results in relatively little negative evidence, then a validity judgement can be reached. This is similar to Bachman's (2005) discussion of the assessment use argument (AUA), where he argues that validating test use should follow validating the test design. Evidently, there is overlap, and even though researchers, such as Cizek (2012), are certain that the sections of Messick's matrix should be separated, Messick (1987) is firm in the necessity of each category to be considered complementary to one another.

Kane was highly concerned with the claims made by test scores, the meanings they carry, and the reliability of test scores as being crucial for making a valid interpretation. Similar to Cronbach and Meehl (1955), Kane argues that test scores can be interpreted differently depending on who is evaluating the evidence. As has been stated, one of the ways to determine validity is to make inferences based on a set of evidence. Although researchers such as Cronbach and Meehl (1955) and Messick (1987) were advocates of construct validity, Kane (2013) supports the argument-based approach to validation. Whereas construct validation combines evidence with theory to create its case, argument-based approaches do not necessarily require a theoretical framework. Additionally, whereas Cizek makes the claim that *interpretations* and *uses* of the tests ought to be treated separately, Kane and Messick stand firm that those two processes are mutually exclusive. In fact, Kane states that too much emphasis has been placed on *interpretations* of tests, and not enough on the *uses*; thus, he promotes the Interpretation/Use Argument (IUA). The IUA “can be specified as a network of inferences and assumptions leading from the test performances to the conclusions to be drawn and to any decisions based on these conclusions” (Kane, 2013, p. 8).

Kane (2013) highlights eight features to consider for test validation of the IUA. The first is that test scores are not limited to one interpretation/use, there are multiple possible interpretations/uses which are subsequently validated – not the test itself. Second, validity of an interpretation/use is determined by whether or not the evidence supports the claims being made. Thirdly, larger claims related to interpretations require more evidence than smaller, less ambitious ones. Fourth, ambitious interpretations are

more useful, however, harder to validate. Fifth, as new evidence presents itself, this may affect the interpretation. Sixth, evaluating a test score should include an examination of the consequences of the suggested uses; to Kane, negative consequences can result in an unacceptable score. Seven, rejecting a score use does not necessarily render a prior interpretation invalid. Finally, validating the score's interpretation based on what the score will be used for does not validate the *use* of the score thus exemplifying the distinction between test interpretation and test use.

Kane discusses several inferences to be made from observable qualities that begin with the score a test-taker receives on their test. There has been much discussion on the meanings of test scores and Kane details how the number received on a test has effects that extend to the much larger picture; Kane discusses three main inferences that relate to these qualities: scoring, generalization, and extrapolation to non-test performance (see figure 2).

Figure 2: Kane's Chain of Inferences (Bachman, 2005)

Scoring is the raw score received on the test; it is important that scoring be done accurately, fairly, and where necessary – such as with performance-based tests – reviewed by a panel of experts to determine the scoring was done properly and is free from bias. Generalization is the process whereby the score is interpreted and extended to make assumptions about how one might perform in the *universe of generalization*, which

refers to the hypothetical target domain of the tests intended use. Evidence to support the generalization inference largely comes from sampling test tasks and test conditions to determine that they are representative of the types of tasks to be encountered in the target domain. For example, a test of world geography should include items of fairly equal distribution about different areas of the world. If it focused most of the items on Europe, it is not accurately testing knowledge of *world* geography and therefore the claims made by such a test would be invalid. The third inference, extrapolation to the target domain, connects the score's interpretation from the universe of generalization to the actual target domain, (i.e., the real-world context) and extends to non-test performances and non-test contexts. Similar to the generalization inference, this inference is concerned with test design, how items are created, and how the test is administered. For example, if response modes on the test match those in the target domain, the extrapolation inference is supported. However, if the response mode differs greatly from the target domain, this limits the validity of the interpretation from the inference.

In criticism of previous validity theories, Bachman (2005) states that frameworks such as those discussed by Kane and Messick are “shopping lists” (p. 32) of features with no real plan of how to effectively connect test use, consequences, and validity (a harsh and controversial critique in my view). In response to his critique, Bachman (2005) proposed his Assessment Use Argument (AUA), which is similar to Kane's chain of inferences in that it considers validity and utilization as a framework for validating inferences; however, Bachman establishes cyclical procedures for justifying test use. His framework involves a cycle of claims, warrants/backing — evidence in favor of the argument being made — and rebuttals — evidence against. According to Bachman

(2005), following this kind of argument structure can more clearly guide test designers and validity researchers to support the tests' interpretations and uses.

To summarize, this section provided an overview of selected validity theory and concepts that have developed in the field over time. Validity researchers are concerned with test interpretation, and test use, as high-stakes tests are used to make important decisions in people's lives. Validity is a quality of the inferences drawn from test scores which consists of overlapping features to decide if a test is more or less valid. Whereas validity is a property of the inferences drawn from tests, validation is the process of gathering assorted forms of evidence for justifying claims set out by the test creators and challenging their assumptions. While Messick (1987), Kane (2013), Bachman (2005) and Cizek (2012) have vouched for their own frameworks, construct validity tends to be at the heart of many, if not all, of these arguments. The construct is an ability that cannot be measured directly, resulting in inferences that stem from test scores. Researchers can analyze different features that surround the test-taking process, such as design, context, and consequences, to determine construct validity. There are a number of different inferences that can be made depending on which element of the testing process the researcher chooses to examine. Finally, the consequential aspect of validity in terms of test interpretation and justification for test use is controversial and is addressed further in the *Washback and Impact* section below (see section 2.2).

It is important to acknowledge the seminal contributions of the *Standards for Educational and Psychological Testing* (American Education Research Association, American Psychological Association, & the National Council on Measurement in Education, 2014), which within this context of debate and discussion, has nonetheless

developed a working consensus on validity and validation (Chapelle, 2021; Fox & Artemeva, forthcoming).

As stated previously, construct validity is a central notion to current validity theories. When examining the construct of a test, researchers must rely on the definition provided by the test designers. As this study is concerned with the construct of the TWE, the next section details the PSC's definition of the construct of writing within the public service.

2.1.1 The Construct of Writing in the Public Service Commission

Section 91 of the OLA stipulates that certain positions within the public service require employees to have a pre-determined level of proficiency in both official languages with respect to reading, writing and speaking; therefore, each position is attached to a *language profile*. Public service employees likewise hold a language profile and their bilingual proficiency standing controls whether or not they are eligible for a promotion or can continue to occupy their current position. For the employees, the process of determining their linguistic profile is regulated through SLE testing. Employees are appointed either because they met the requirement upon hiring, or they are given a period of two years to attend language training to achieve the level, as stipulated by the Public Service Official Languages Appointment Regulation (PSOLAR) (Department of Justice [DJ], 2005). The regulation additionally states that should the employee not fulfill the requirement to become bilingual within the allotted time period, they may be moved to a position that matches their bilingual proficiency. Treasury Board Secretariat (TBS) regulates the qualification standards which detail the abilities needed in order to perform the duties of the position held. The TBS website states, "Official

language qualifications are identified objectively and are relevant to the duties and responsibilities of the position as it relates to communications with and services to the public and language of work” (TBS, 2016, sec. 1).

The PSC administers three tests to determine an employees’ proficiency in their L2: Written Comprehension (reading), the Test of Written Expression (writing) and Oral Interaction (speaking). The levels of proficiency range from X-A-B-C-E, with X at the lowest (i.e., the test-taker did not even qualify) and Exemption (E) at the highest. E indicates that the proficiency level is of level C or higher, and such that it will be maintained indefinitely signifying they will never again have to retake the test. The reading and writing tests are multiple-choice and aim to test ability by having the test-taker fill-in-the-blank and identify errors, or answer comprehension questions. The oral proficiency test is a 30-minute face-to-face interview.

Table 2: Table 2: Qualification Standards – Official Languages (TBS, 2016, sec. 4

Level	A	B	C
Can accomplish	Tasks at level A	Tasks at level A	Tasks at level A
		Tasks at level B	Tasks at level B
			Tasks at level C

Table 2 shows the cumulative nature of each proficiency level; if someone achieves a level C, it is assumed that they have met the qualifications of both level A and B.

The following definitions are with respect to writing and were taken from the Treasury Board website regarding the qualification standards (TBS, 2016).

Level A

Positions that require a level A in writing will involve writing smaller components of information. TBS (2016) states, “A person writing at this level can write isolated words, phrases, simple statements or questions on very familiar topics using words of time, place or person” (sec. 8). Mistakes with grammar, vocabulary and spelling are accepted, so long as the communication is understood. Some examples of writing ability are simple phrases, lists, or brief notes. Functions for this level include filling out a form, completing a table of contents, or requesting and providing simple information. To receive an A, the test-taker must score between 20-30 points (PSC, 2019).

Level B

Positions that require level B involve writing short descriptions or fact-based texts in the L2. “A person writing at this level can deal with explicit information on work-related topics since they have sufficient mastery of grammar and vocabulary” (sec. 9). The texts are expected to require some corrections in grammar, vocabulary, and revision for style. In addition to the level A tasks, a person writing at this level can write short work-related messages, tests, descriptions, or correspondence. Functions for this level include adapting templates, requesting or providing information, explanations, or instructions, explain or request an action to be taken, or summarize a text or meeting in point or note form. To receive a B, the test-taker must score between 31-42 points (PSC, 2019).

Level C

Positions that require level C in writing will have the employee write informal and formal work-related explanations and descriptions. The description states “A person writing at this level can write texts where ideas are developed and presented in a coherent manner” (sec. 10) This section of the website additionally states, “A person at this level can also modify or correct texts to improve meaning, tone, clarity and conciseness.” In addition to the tasks at level A and B, a person writing at this level can write work-related correspondence, briefing notes, reports, research papers, comprehensive summaries, and detailed presentations. Functions include providing or requesting detailed facts and reasons and providing or commenting on contentious issues. To score a C, the test-taker must score within 43-51 points (PSC, 2019). Bearing this information in mind, it would be the definitions provided by TBS to consider when contemplating the construct of the TWE.

The TWE would be considered an indirect assessment method which involves the test-taker making a judgement regarding the appropriateness of a piece of writing (Stiggins, 1982). The test consists of 65 multiple-choice questions, involving the test-taker to perform two tasks: either fill in the blank with the correct response or identify the error in a text out of a choice of three highlighted sections. In comparison, direct writing assessments would require the test-taker to produce a written sample. Testing writing using indirect methods has trade-offs when compared with direct assessment methods. Advantages of indirect methods include high score reliability, (i.e., it is scored as either right or wrong), low administration costs, low scoring costs (the test can be marked by a machine), and the test developer has control over the skills tested. Disadvantages involve

a lack of fidelity in terms of real-world writing tasks, the test-taker uses their reading skills as opposed to their writing skills, and lower *face validity*, that is the test-taker may doubt the effectiveness of the design. Comparatively, direct assessments are associated with lower test development costs, provide more information about ones' genuine writing ability, which increases the reliability between the stimulus and response as the test items relate more directly to the real-world situations. This in turn increases the face validity of the writing samples. However, disadvantages may include high scoring costs, as a human rater is required for grading, and variation in the raters' judgements (Stiggins, 1982). Raters also must undergo training to ensure they are rating the test in similar ways.

2.2 Washback and Impact

The previous section on *validity* noted that more and more researchers are concerned with the consequences that may stem from a test, and whether or not they should be integrated as part of the validity judgement. Hamp-Lyons (1997) argues that there needs to be: (1) a different way of looking at test influence that includes a narrower one of washback and the wider term of impact, and (2) an ethical framework that encompasses the broader influences of testing on the societal level. Issues related to these concerns will be addressed in the discussion that follows, beginning with washback, then moving on to impact, and finally connecting these issues to validity theory.

2.2.1 Washback

The area of *washback* has gained more popularity since the 1990s, which saw an increase in washback studies (Cheng, 2014; McNamara, 2006). The term *washback* is generally used to describe the micro-level influence of assessments on teaching and learning (Alderson & Wall, 1993). A well-recognized definition of washback comes from

Alderson and Wall (1993) who conceived the *Washback Hypothesis*. They state “The Washback Hypothesis seems to assume that teachers and learners do things *they would not necessarily otherwise do* because of the test [emphasis in original]” (p. 117).

Alderson and Wall (1993, pp. 120-121) additionally posit 15 sub-hypotheses:

5. A test will influence teaching.
6. A test will influence learning.
7. A test will influence what teachers teach.
8. A test will influence how teachers teach.
9. A test will influence what learners learn.
10. A test will influence how learners learn.
11. A test will influence the rate and sequence of teaching.
12. A test will influence the rate and sequence of learning.
13. A test will influence the degree and depth of teaching.
14. A test will influence the degree and depth of learning.
15. A test will influence attitudes to content, method, etc. of teaching/learning.
16. Tests that have important consequences will have washback.
17. Tests that do not have important consequences will have no washback.
18. Tests will have washback on all learners and teachers.
19. Tests will have washback effects for some teachers and some learners, but not for others.

It is important to note that Alderson and Wall do not claim these to be definitive, but rather offer them as a guideline. Many studies that followed have referenced these hypotheses (e.g., Bailey, 1996; Green, 2007; Hamp-Lyons, 1997; McCullough, 2018;

Wall & Alderson, 1993). Furthermore, thoughts towards washback have changed as the field continues to develop over the years. Washback used to be thought of as negative (Hamp-Lyons; 1997), however now it is widely documented that washback can be positive, negative, or neutral, depending on the context (Alderson & Wall, 1993; Bailey, 1996; Cheng, 2014; Wall & Alderson, 1993). Originally, it was thought that good tests result in positive washback, and bad tests yield negative washback, however it is now recognized that seemingly good and valid tests can result in both outcomes. Alternatively, bad tests have the ability to create positive influence, depending on how one defines positive (Alderson & Wall, 1993; Bailey, 1996; Cheng, 2014). Therefore, it is important to consider different perspectives before deciding whether or not the washback is positive, negative, or neutral (Cheng, 2008, 2014).

As we can see from the 15 sub-hypotheses, they are concerned mostly with the influence of a test on teaching and learning. For teaching, this affects matters such as content of the study materials (i.e., *what* is taught), and methodology (i.e., *how* it is taught). Materials are created to support the development of skills needed for the test, which subsequently have the potential to be taken over by teachers; these materials may or may not influence what and how they teach (Alderson & Hamp-Lyons, 1997; Green, 2007; Shohamy et al., 1996; Wall & Alderson, 1993). For example, in terms of content, Wall and Alderson (1993) examined washback in the classrooms of Sri-Lanka to see if and how the O-Level exam (a high-stakes exam that either grants entrance to a university or a promising career) influenced the classroom. They noted that in the beginning most of the teachers were teaching using the new communicative textbook; however, towards the end of the four terms the new textbook was dropped by teachers almost entirely; they

switched to more explicit test preparation materials. Alderson and Hamp-Lyons' (1996) study observed two teachers who taught both a TOEFL preparation course, in addition to a regular ESL course; the preparation course was guided by a textbook that mirrored the structure of the tests' tasks; teachers additionally reported spending less time on lesson planning, and homework. Green's (2007) study investigated washback to learning outcomes with the International English Language Testing System (IELTS) test. He compared entrance and exit scores of three ESL programs on the improvement of the participants' writing. The programs were an IELTS preparation course (Type 1), a pre-sessional EAP (Type 2), and a combination of both (Type 3). Interestingly, Green (2007) did not find significant differences between the three courses; all three types resulted in an improvement in the students' writing.

The same studies reported differences in the methodology used by teachers when the focus shifted to test preparation. Alderson and Wall (1993) recorded less student talk time, and an increasing influence of the test on what was being taught. Alderson and Hamp-Lyons (1997) noted how in the TOEFL preparation courses, the teacher talked more, there was more use of metalanguage, less pair-work, as well as less instances of shared laughter. Green (2007) concludes that the test preparation courses, thought to speed up the process exponentially, were no better or worse than a regular sessional ESL for academic purposes course, and the combination course worked just as well. This finding from Green suggests that perhaps test preparation courses are the wrong approach; other course structures were equally effective.

While teachers have the ability to adapt their methodology, the content is typically decided by a curriculum. Ideological values of education are present within an

assessment, which then influences the curriculum (McNamara, 2006, 2007; Shohamy, 2005). For example, section 2.1.1. of this study highlighted the written language skills that are valued within the context of the public service; the SLE tests attempt to reflect those constructs through test design. Educational governing bodies may decide which materials the teachers should use; comparatively, experienced teachers may supplement their lessons with their own test preparation materials (Alderson & Hamp-Lyons, 1996; Alderson & Wall, 1993; Cheng, 2014; Shohamy et al., 1996; Wall & Alderson, 1993). In the context of the current study, CSPA has prepared a number of resources to help employees and teachers alike with language training. Private academies that have a standing offer with the government additionally have access to CSPA materials.

Generally, negative washback on teaching and consequently learning are a result of *curricular alignment* or a *narrowing* of the curriculum, which occurs when a teacher begins to center their course around teaching to the test, rather than developing language (Alderson & Wall, 1993; Bailey, 1996). Such examples of negative washback on teaching would be abandoning a text to use their own materials, using test preparation texts to supplement the curriculum (Bailey, 1996), ignoring the methodology suggested (e.g., communicative teaching) and using older versions of the test as in-class assessments and activities (Shohamy et al., 1996; Wall & Alderson, 1993). On the contrary, positive washback on teaching would be minimal narrowing of the curriculum and, in the case that there is a course text as part of the curriculum, teaching the course text as it should be taught without overly emphasizing an external test (Alderson & Wall, 1993). In a general L2 class, further examples of positive washback on teaching would be balancing all four language skills rather than focusing on reading and writing, while

additionally having in-class activities that supplement the content from the course text, rather than designing activities based on previous versions of tests. Thus, washback on teaching through the content (i.e. materials) used in language classes plays a prominent role.

Much like tests affect teaching, tests likewise affect learning because what is taught in the classroom is consequently what is learned (e.g., Alderson & Wall, 1993; Green, 2007; Shohamy et al., 1996). In the classroom, for example, tests that motivate students to work harder in class can be regarded as positive washback on the student. Contrarily, tests tend to induce anxiety and stress, which are seemingly negative factors and are therefore considered negative washback on the student (Alderson & Wall, 1993). Bailey (1996) discusses that the goals of the students may be in conflict; a student may have a personal goal of passing the test and achieving a certain level, but likewise have a general goal of improving his/her language skills. Achieving goals is positive washback on the learner, but could possibly conflict with sound pedagogy thus negatively influencing teaching. Other factors influencing the learner are referred to by Bailey (1996) as *washback to the programme*. As discussed at length in this paper, tests are used to make decisions. Educational institutions use tests to place students into programs using placement tests; diagnostic tests have been used by teachers to inform classroom decisions. The sheer existence of a high-stakes test may influence learners to sign up for test preparation courses which within themselves have goals the program wishes to achieve.

In the case of the SLE, it is not exactly clear which came first, the test or the washback since the PSC is responsible for the testing portion, and CSPS is in charge of

language training. Historically, language training was centralized within the PSC; this changed when the CSPA was created in 2004 (OCOL, 2013). In the context of the Canadian public service, these two aspects of language training and testing are closely intertwined yet separate. Both language training and testing can be considered programs that have goals to help employees meet the language requirement of their positions and consequently lead to successful implementation of the OLA (OCOL, 2013). As stated earlier, the Public Service Official Languages Appointment Regulation prescribes that employees who do not meet the bilingual requirements of their intended position may still be placed so long as they meet this condition within two years (JD, 2005). The federal government offers language training through the CSPA; alternatively, there are a number of standing offers between the government and private language schools (PSPC, 2020); Bessette (2005) and Gentil et al. (2010) have documented experiences with the in-house language training. The intersection between the classroom, the program and the policy where the conversation shifts to discussions of impact.

2.2.2 *Impact*

Impact is a term used to document effects of testing on the larger societal level, such as in educational institutions and employment (Bachman & Palmer, 1996; Cheng, 2014; McNamara, 2006). As the previous discussion of washback has demonstrated, the presence of a test can have an effect on what transpires in the classroom. Additionally, the results of an assessment do not end with a test; rather, they extend to broader society. As a result, researchers (e.g., Wall & Alderson, 1993; Alderson & Hamp-Lyons, 1996; Hamp-Lyons, 1997) express the need to consider such effects of testing.

Bailey (1996) and Hamp-Lyons (1997) discuss *washback to the program*; considering Figure 3, the program resides in the educational level and is concerned with curricular decisions that ultimately feedback into the classroom and can control what is taught (Markee, 1993).

Figure 3: Levels involved in hierarchical decisions (Markee, 1993, p. 234)

Curricular decisions are often implemented by even broader functions, as is the case in Canada with the SLE. We can think of the cultural level as relating to the discussion of bilingualism; the OLA has been implemented and exercises its control on the political level. The SLE can be considered an administrative role that facilitates implementation of the policy; language training and language schools fall at the bottom of the hierarchy in Figure 3. Each level operates within their own set of beliefs and constraints (Markee, 1993), thus changing elements of the SLE is faced with challenges as it is controlled by the OLA. Positive washback to the program can be measured in terms of the program meeting its goals; conversely, negative washback would be the opposite (Bailey, 1996; Messick, 1996). We can therefore consider achieving a bilingual institution the goal of the OLA and consequently the SLE.

On discussing the impact of tests, Shohamy et al. (1996) identify five factors that influence the level of potential impact of a test: low-stakes versus high-stakes, language status, the purpose of the test, the format, and the skills tested. They state, “The power and authority of tests enable policy-makers to use them as effective tools for controlling educational systems and prescribing the behavior of those who are affected by their results” (p. 299). Hamp-Lyons (1997) adds “the impact of assessment and the expectations laid upon it spread all the way out to the society as a whole; it is not only test developers whose work has ‘impact’; it is also testing agencies who make policy and economic decisions about the kinds of testing to support and the kinds that will not be supported [...]” (p. 298). McNamara (2007) further emphasizes the social context of testing. He states:

...the certification we gain as a result of our subjection to processes of examination confers an identity — we become socially visible as possessing the qualities the examination seeks to establish. We thus become socially visible in the social roles — the subject positions — available within the discourse” (p. 134).

These quotations describe how impact refers to the consequences of testing on the broader educational/societal level, the discussion of which can be traced back to Messick’s (1987) consequential basis of test use. Studies in this field document larger issues that deal with fairness and ethics in testing contexts (Bachman, 2005; Bachman & Palmer, 1996; Hamp-Lyons, 1997; Kunnan, 2004; McNamara & Ryan, 2011; Messick, 1994, 1996; Moss, 1998; Shohamy, 2005).

Drawing from the studies cited in the washback section to exemplify this idea, each one was situated within a broader context which has impact. For example, Alderson and Hamp-Lyons (1996) demonstrated washback effects on the classroom level with TOEFL preparation courses, however the impact of TOEFL extends beyond the classroom to the institutional curriculum. The USA and other countries around the world were offering students the option of taking TOEFL preparation courses as an additional class. As stated earlier from Alderson and Wall (1993), the O-Level exam is a high-stakes exam that either grants students in Sri Lanka entrance to a university or a promising career, therefore making it the “the most important [test] in his or her academic career” (p. 42); subsequently resulting in a curriculum to facilitate success on the exam. Green (2007) discusses the importance of IELTS, a high-stakes English proficiency test that is widely used to assess the readiness of international students to engage in academic study in their L2, the scores of which are widely accepted as a condition of admission to a post-secondary institution. Depending on their score, the student may have to take additional English classes which will cost them time and money. In reference to the current study, the results of the SLE determine if someone is linguistically qualified to perform their job or eligible for a promotion.

Policy-based decision that come from the top-down, such as with the SLE, can cause a lack of clarity (Fox, 2009). Fox, Haggerty, and Artemeva (2016) provide a positive example of the power of tests when an institution works together rather than separately in a Canadian university. They used diagnostic testing to establish how likely new undergraduate students were to succeed in their first-year engineering program. The test identified students at risk which then informed the academic support unit; ultimately,

an academic success center was created specifically for the engineering students who were identified as being at risk. Interestingly enough, some of the students identified were not entirely the ESL international students. A series of studies documents this process (see Fox & Artemeva, 2017; Fox, von Randow, & Volkov, 2016). This is an example of the power test data can have when used appropriately, the result being positive impact on an entire academic unit. Furthermore, it is a reminder to question the efficacy of tests; even though one test determined admission to the program, another test identified risk.

To summarize, effects and influence from tests appear to trickle down into classrooms while likewise trickle up into society. Washback has been evidenced to have effects on teaching and learning at the classroom level; in some cases, washback is either positive, negative, or neutral, as perceived by the stakeholder. It has been highlighted how tests have consequences, which adds to the importance for the inferences of a test to be justified because of its potential impact (Bachman & Palmer, 1996; Messick, 1994, 1996; Shohamy, 2005). As stated at the beginning of this section, an ethical framework that encompasses the broader influences of testing on the societal level is needed (Hamp-Lyons, 1997). How these concerns of washback and impact relate to validity will be discussed in the following section.

2.3.3 *Washback, Impact and Validity*

Researchers in this field use terms such as *ethics*, *fairness*, and *justice* to frame their discussions of impact in language testing. In discussing such topics, McNamara and Ryan (2011) contrast *fairness* and *justice* stating that the former is concerned with what they reference as the *technical* quality of tests (i.e., test design); they reserve the latter for

issues of test use. If the purpose of the SLE is to regulate bilingualism within the public service, then the justice of the use of the test is supported. Messick (1987) arguably paved the way for researchers to include impact as part of validity by dedicating a part of his matrix to consequences of test use. In later publications, he refined his case to include consequences as a part of construct validity, where intended and unintended consequences of tests should be evaluated as part of validation. Some researchers (e.g. Bachman, 2005; Kunnan, 2004; Messick, 1994) argue that washback should be considered a feature of validity, that if a test has negative influences on teachers – and consequently students – this could be a case of invalidity.

However, Messick (1996) explains that because washback is such a complex concept consisting of different dynamics, it is simply too wide to claim that *any* negative influence should result in the invalidity of a test. For Messick, washback should be considered a facet of validity to the extent that properties of test design lead to negative washback. He says, “...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 of other forces operative on the educational scene” (p. 242). As discussed earlier concerning Messick’s view on validity, two main threats are construct underrepresentation and construct irrelevant variance. These characteristics evidently are related to the technical qualities of test items, design, and administration (McNamara & Ryan, 2011). Likewise, Messick (1996) argues that every test, no matter how well or poorly designed, will have some elements of construct underrepresentation and construct irrelevant variance. The task then for designers is to create a test that is administered in such a way that minimizes these

effects. This, along with relatively authentic tasks, will ideally promote more feelings of fairness and accordingly positive washback to the test-taker.

There are two issues with a narrow view which limits considerations of validity to aspects of test design and the technical quality of a test: (1) it ignores the fact that test interpretations and meanings are interpreted against societal values (McNamara, 2007), and (2) tests have the power to control and serve political agendas (Shohamy et al., 1996; Shohamy, 2005). Therefore, regarding the impact of a test, the narrow perspective stems from the test items and design; broadly, it extends to larger issues of test use, or justice (Kunnan, 2004; McNamara & Ryan, 2011). Moss (1998) argues that those who develop, mandate, and implement a testing program have a responsibility to monitor the context in which it is implemented. She says “the study of consequences becomes an essential aspect of validity even for those who choose to limit the scope of validity to a test-based interpretation” (p. 7). Moreover, she states:

If the goal is to make decisions about how to improve teaching and learning or to make choices among alternative courses of action or policies, evidence of student outcomes alone is insufficient; one must consider information about the conceptual and material resources, the teaching processes and practices, and the organizational routines and cultures that shape or influence those outcomes (Moss, 2016, p. 93).

She discusses the issue that researchers and test designers gather data, but policy makers may not necessarily know how to use the data effectively (Moss, 2016). Moss suggests that the conception of validity should shift depending on the purpose, emphasizing that broader examinations may be needed in order to support an individual; thus it is critical for test creators, policy makers, as well as test users to work together when considering

consequences of testing in both broad and narrow definitions to facilitate effective decision making.

Many contemporary theories of validity include features of test use; however, an issue that still remains is how to adequately link test validation to these broader concerns. For example, Kane's decision and policy inferences demonstrate his preference for integrating the consequential aspect into the validity argument, yet Kane does not provide a methodology to do so (McNamara, 2006). While McNamara (2007) and Shohamy (2005) address important issues related to ethics, fairness, and the social realm of testing, there have been criticisms against these types of discussions from researchers (e.g., Cizek, 2012) for attempting to combine different aspects of the testing context with respect to validity of the interpretations on the one hand, and theories of test use on the other. Two resources we can reference as a framework for test use are Bachman's (2005) two-part assessment/use argument (AUA), and Kunnan's (2004) test fairness framework.

Bachman (2005) states "These discussions, though clearly enlarging our view of the social and educational milieux in which language testing is situated, do not, in my view, bring us any closer to a coherent theory of test use" (p. 4). According to Bachman, the utilization argument consists of similar features as the validity argument, namely claims, warrants, backing and rebuttals, however with some differences. Claims in the utilization argument draw on the validity data as the starting point to begin evaluating decisions to be based on these claims. Warrants, backing, and rebuttals are collected in terms of relevance of the score, utility to the TLU, intended consequences of the assessment and the decision, as well as the sufficiency of the evidence to support the decision made by the assessment. Bachman vouches for a two-part system in which

validity of test interpretation, and validation of intended use, is determined separately. Although Bachman states that one could begin with either step, he recommends beginning with the utilization argument in order to ensure the test will be designed with a useful purpose. While Bachman attempted to create a framework for test utilization, McNamara (2006) states that Bachman deals with issues of use too narrowly, ignoring the broader social context and arguably freeing test creators of being held accountable for their instruments.

Kunnan (2004) has been widely cited in the literature for his fairness framework for test evaluation, which was guided by two ethical principles: (1) the Principle of Justice, which states that a test should be fair for all test-takers, and (2) The Principle of Benefice which states that a test should have a positive, rather than harmful, effect on society. Furthermore, his framework includes five qualities: validity, absence of bias, access, administration, and social consequences. Absence of bias refers to the content of test items that may or may not be offensive, discriminating, or otherwise lead to construct-irrelevant factors that can affect test performance. Access addresses issues such as educational access, financial access, geographical, personal, and technical requirements; test-takers should not be penalized due to an inability to access one aspect or another of the test. Administration refers to qualities of the testing conditions; if the test room is noisy, too cold, etc., or if the test administrators treat the test-takers differently in a performance assessment, this can also lead to irrelevant-variance which would corrupt the score. The final feature refers to social consequences which includes washback and remedies; test-takers should have the opportunity to be re-evaluated in the face of adverse testing consequences.

While researchers have attempted to fill in the gaps, it seems there is nevertheless a lack of a recognized methodological framework for how exactly to connect test validation and test use (Bachman, 2005; Cheng, 2014; McNamara, 2006). The discussions to date have raised important issues when considering the impact of testing, however they remain just that — issues to consider. Cheng (2014) suggests researchers collect evidence from multiple stakeholder perspectives and establish a methodology that can appropriately link test validation and test use, as this is still required within the language assessment literature.

2.3 Threats to Validity

Since this study is concerned with possible threats to validity, what follows is a discussion of factors that can threaten the validity of the inferences of a test. Adopting McNamara and Ryan's (2011) distinction between fairness (i.e., technical quality) and justice (i.e., test use), threats to validity are associated with design, content, testing environment, and negative consequences that occur as a result of the uses of the test. One other outside factor to some extent is the influence of coaching (Messick, 1996; Kane 2013). Following these considerations, previous SLE studies will share the ways in which evidence of these threats have been found.

Firstly, considering test design and content, *construct underrepresentation* is the lack of clarity and explicitness of the construct (Messick, 1994; 1996). McNamara and Ryan (2011) give the example of listening and speaking tasks mediated through a computer to make judgements about ones' interactive speaking ability. This would threaten validity of the inferences gathered because not all abilities needed for adequate generalization to the real-world context are facilitated through the tests' design. Messick

(1994) adds that if the construct being tested is not clear, the test-taker may not know how to adequately perform; test-takers should know what is being assessed and have a baseline idea of what constitutes a good or bad performance (Messick, 1994). Cronbach and Meehl (1955) and Messick (1994) call for rationale of test specifications that tie together the purpose of the test to the domain, combined with construct theories (such as in Bachman, 2004). This would support qualities of transparency and meaningfulness central to authenticity, and for many researchers, validity (Bachman, 2005; Bachman & Palmer, 1996; Cronbach & Meehl, 1955; Kane, 2013; Messick, 1994, 1996).

Content validity is one aspect of construct validity; threats of this nature involve the correspondence between the test tasks and the TLU domain. Bachman and Palmer (1996) address the importance of making a strong connection between the content and the TLU. Additionally, Messick (1996) adds:

A key issue for the content aspect of construct validity is the specification of the boundaries of the construct domain to be assessed –that is, determining the knowledge, skills and other attributes to be revealed by the assessment tasks (p. 249).

Since language tests such as the IELTS, TOEFL-iBT (McNamara and Ryan, 2011), and the SLE are used to assess one's ability to operate within a certain context, it is necessary that the test items reflect the qualities needed to perform in the real-world setting; if they do not, then the ability to generalize from the test score will be invalid, and the construct validity argument will fall apart (Kane, 2013). Kane further extends this to *irrelevant method variance*, meaning if the response mode varies greatly from what would be encountered in the real-world setting, this could threaten the interpretation made.

Construct irrelevant variance refers to adverse factors that can negatively impact the results of a test (Cizek, 2012; McNamara & Ryan, 2011; Messick, 1994, 1996). For example, the researchers in DeLuca et al. (2013) took the TOEFL-iBT test in order to document the entire test-taking process. They commented on adverse effects that distracted them while taking the test such as a cold room, not being allowed to wear shoes, and people speaking when reaching different parts of the test. Similar issues with construct irrelevant variance were noted in McCullough's (2018) study which looked into both IELTS and TOEFL. If the test score is used to determine academic English ability, these adverse effects could result in an inaccurate score for the test-taker, and ultimately an unfair decision (Kunnan, 2004).

Moving the discussion to broader considerations, issues of fairness could threaten validity if a test results in negative consequences for the stakeholder(s) (Bachman, 2005; Kane, 2010; Kunnan, 2004; McNamara and Ryan, 2011; Moss 1998, 2016). For example, according to Kane (2010, p. 181):

An assessment that is unfair, in the sense that it systematically misrepresents the standing of some individuals or some groups of individuals on the construct being measured or that tends to make inappropriate decisions for individuals or groups is, to that extent not valid for that interpretation or use. Similarly, an assessment that is not valid in the sense that it tends to generate misleading conclusions or inappropriate decisions for some individuals or groups will also be unfair.

This quote from Kane touches on two aspects: (1) test quality and (2) inappropriate decisions made from tests. Extending beyond test design, if tests have negative

consequences from inappropriate decisions, this could invalidate the inferences drawn from the test. In justifying his AUA, Bachman (2005, p. 32) adds:

We need to be able to justify the interpretations and uses we make of any assessment so that we can be accountable to the stakeholders. Therefore, the primary reason for articulating an assessment use argument, and for collecting evidence in support of this, is so that we can be accountable to stakeholders.

Finally, McNamara and Ryan (2011) emphasize that tests could be of sound design, yet still have negative impact. Hence it is necessary to consider broader issues of negative impact as a facet of validity as Messick (1987) does in his matrix.

Coaching has been identified as another threat to validity, as coaching and test preparation courses could lead to a misrepresentation of the abilities of the test-taker (Cronbach & Meehl, 1955; Kane, 2013; Messick, 1994). Cronbach and Meehl (1955) state

When the coaching is of a sort that improves the pupil's intellectual functioning in school, the test which is affected by the coaching has validity as a measure of intellectual functioning; if the coaching improves test taking but not school performance, the test which responds to the coaching has poor validity as a measure of this construct (pp. 288-289).

We can apply this idea to language testing; if coaching results in a better score, but no significant improvement in language ability, the test would not be an accurate representation of the ability tested. Kane (2013) also claims that if a test is coachable, the validation of the IUA may decline. Messick (1994) gives the example of coaching for a

proficiency test, stating “coaching emphasizing test-wiseness strategies that might increase test scores without correspondingly improving the skills measured by the test” (p. 246). Thus, coaching and test preparation may result in an inaccurate depiction of the individual’s true language ability. Furthermore, as reported in the washback section above, narrowing of the curriculum is considered negative washback to both teachers and learners (Bailey, 1996; Hamp-Lyons, 1997), and is a negative consequence of the test.

2.3.1 Validity in Context: the SLE

After considering these threats to validity of inferences, the focus now shifts to previous SLE studies. The current study is interested in exploring the SLE-TWE, understanding more about what it is assessing, and if there is any aspect of the test-taking process that undermines the validity of the inferences drawn from test in question. Apart from government publications, Bessette (2005), Gentil et al. (2010) and McNamara and Roever (2006) are three of the few publicly available studies that have discussed the SLE to date. Bessette (2005) and Gentil et al. (2010) focus on the language training process rather than the tests; however, it is difficult for the tests to be disregarded entirely, as one aspect of government language training is test preparation. McNamara and Roever (2006) elaborate on how the creation of these tests is rooted in policy. Since the tests are controlled in this way, it makes it more difficult to change; changes to any of the tests would have to navigate through the various levels of government responsible for the implementation of the OLA outlined in the introduction.

Bessette’s (2005) study looked specifically at opinions of higher-level employees taking French training in preparation for the SLE. Bessette (2005) aimed to investigate the learner’s impressions of the language training process through analysis of the one-on-

one interviews. She found that there were many comments that pertained to the oral test. For example, several of the participants commented that the pressure of passing the test was a constant and lingering thought. Additionally on the topic of assessment, the employees commented on how the language training program seemed to lack explicitly stated learning outcomes, which Bessette refers to as negatively impacting curricular validity. This can be extended to an issue with alignment, (i.e., where what is taught is not necessarily what is tested, and vice-versa, Cheng & Fox, 2017). McNamara and Roever (2006) agree that there exists a “lack of explicitness of the construct” (p, 188) where the SLE is concerned. Bessette (2005) suggests making information about the skills the test intends to measure explicitly clear and more widely available in order to improve this aspect of the training program, ultimately resulting in less construct underrepresentation.

Gentil et al. (2010) examined French language training and attrition in the workplace. They found evidence of a disconnect between the tasks the employees do in preparation for the test (i.e., in language training) and their actual workplace (Bessette, 2005; Gentil et al., 2010), which could indicate a threat to the representativeness of the test to the target domain (Messick, 1987). Gentil et al. (2010) report that the government’s in-house training program takes a more general approach to teaching work-related tasks and suggests that if the language training were more tailored to the specific needs of the employees, this may foster positive attitudes to their second language. Additionally, the apparent dominance of English in the workplace and limited opportunities to effectively use French for work-related functions results in employees losing their French knowledge, a process known as *language attrition*. Gentil et al.

(2010) reported how at times, conversations will start in French, but when it becomes time to talk about actual duties and assignments, everyone switched to English.

Other comments from stakeholders in Bessette (2005) were related to how the oral test was a one-shot summative test, 30 minutes in length. Students felt that a bad day or a bad 30 minutes could have immense consequences; moreover, they commented on how there is only one rater for the test which made them feel as though a different rater on a different day could yield different results. Comments such as these raise issues that relate to the validity of the test. For one, the lack of clarity and explicitness of the construct signifies construct underrepresentation (Messick, 1994, 1996). Secondly, the participants commented on how they felt many factors could have impacted their results, such as mood of the evaluator and the time of day. These are issues of construct irrelevant variance; the fact that the stakeholders feel their results highly depend on how they or their rater is feeling that day is another threat to valid inferences drawn from the test because elements outside of their language are affecting the results (McNamara & Ryan, 2011). Furthermore, McNamara and Roever (2006) comment on how frustration from the testing experience then leads to frustration with the policy; tests that are perceived as unfair or arbitrary have low face validity (Bailey, 1996) and make people reject the policy. The policy itself is well-intentioned — to promote bilingualism in Canada — however the way it is being implemented is perhaps not fostering positive attitudes, as evidenced in McNamara and Roever (2006), Bessette (2005), and Gentil et al. (2010).

Considering further aspects that contribute to the consequential aspect of validity, there is a major impact of the SLE tests on the individual, as the tests can decide whether

the employee keeps their current position or is able to receive a promotion. This has been shown to produce anxiety around taking the test because in a sense, their future is determined by 55-65 multiple choice questions, or a 30-minute interview (Bessette, 2005; Gentil, 2010). Multiple-choice as a method includes distractors which can provide misleading information. Bessette's (2005) study documented the experiences of employees preparing to take the test who can occupy and have been occupying their positions; for them to risk being changed to a lower position because of the results of their tests is perhaps unfair.

In addition to the effect on the individual, there is a much larger impact on Canadian society as a whole. For example, according to a study by Villaincourt et al. (2012), the federal government spent roughly \$1 billion on bilingualism efforts, of that nearly \$340 million was spent on language training for public servants in 2006. Values and goals of society are reflected within assessments; if we follow the thinking of Messick and Kane who integrate impact as part of their validity argument, it would be worthwhile to ensure that the system is working effectively and that this money is being put to good use. This then emphasizes the need to ensure the SLE tests are effective, reliable, and most importantly, yield valid inferences.

In summary, main threats to validity consist largely of the technical qualities of test design and content, such as construct underrepresentation, construct irrelevant variance, and a strong connection to the TLU domain (Bachman & Palmer, 1996; Kane, 2013; McNamara & Ryan, 2011; Messick, 1987, 1994, 1996). Coaching has been considered another threat, as it potentially alters or distorts the measurement of the ability of the test-taker (Cronbach & Meehl, 1955; Kane, 2013; Messick, 1994). To some extent,

negative effects of washback and impact should be treated as a threat to validity, however opinions on this vary (cf. Bachman, 2005; Cizek 2012; Kane, 2013; Kunnan, 2004; Messick, 1994). Previous research on the SLE indicate that there may be some issues with the current testing system as well as with implementation of the OLA (Bessette, 2005; Gentil et al, 2010; McNamara & Roever, 2006); if this is the case, such issues are of public concern (Kane, 2013).

The literature review highlighted relevant themes in relation to validity, washback and impact. The next chapter details the methods that were involved in answering the research questions.

Chapter Three: Methods

Having introduced the focus of the present study on the SLE-TWE and discussed relevant literature, this chapter details the research methodology employed to answer the research questions.

3.1 Research Design

The current study employs qualitative research methods by collecting stakeholder perspectives in a case-study approach to examine the phenomena of the SLE and the construct of the TWE. Five participants consisting of two types of stakeholders are included in this study: three government employees and two language teachers. In a case study approach, the researchers are able to deeply analyze the individual experiences of the participant and relate it to broader conditions (Polio & Friedman, 2017). It has been suggested by Fox and Cheng (2015) that future research on testing and validity include evidence from multiple stakeholder perspectives, and employ various methods in order to effectively connect the micro- and macro-level consequences of a testing program (Cheng, 2014).

In this study, I am interested in the accounts of test-takers and language teachers, their experiences, and what these accounts suggest about the construct that is being measured, anything that may be underlying the validity of the TWE, how well the test tasks match with tasks in the TLU, and if there are any washback effects. To investigate these issues, three types of qualitative data were collected for analysis: (1) a concurrent verbal protocol using a shortened version of a practice test (Appendix B), (2) one-on-one interviews with the test-takers (Appendix C), and (3) one-on-one interviews with experienced language teachers (Appendix D). Having three sets of data allows for

methodological triangulation, which when employed in case study research, improves the credibility of the claims being made (Polio & Friedman, 2016). A concurrent verbal report was used as a way to understand the cognitive processes involved while the employees were taking the test to better understand the extent of their L2 knowledge (Bowles, 2010).

3.2 Setting

The setting for the data collection phase was focused in the National Capital Region (NCR) of Canada due to the concentration of government headquarters and linguistic make-up of the region. There are several reasons an employee would take their SLE. The first is that language profiles expire every five years for government employees and as such, must be renewed for the employee to continue to meet the requirements of their positions. Secondly, an employee receiving a promotion or transferring to a new position which requires a higher level in any of the language functions would have to take the test of that particular function. For example, someone could occupy a position that requires BBB, but they have an ECB profile. If they transfer to a new position that requires CBC, they would then need to meet the requirement for their oral expression and would only need to take the oral exam. Depending on the position, the government would pay for the employee to attend language training to help them improve their language ability. The participants intended for this study were those who were involved in such processes.

3.3 Participants

This study is concerned with test validity and stakeholder perspectives; therefore, two types of participants were recruited. The first type is government employees who

have experience with the SLE process. The population targeted were those who have a government linguistic profile of B/B/B or higher, as this would ensure that they have enough communicative ability and understanding in their second language to comfortably participate in the verbal protocol, and provide a good insight into the inner-workings of their L2 processes. A language profile was chosen as the target for the study rather than job classification due to the fact that the SLE is designed in such a way that it generalizes to all government employees who take the tests. Of course, there is a link between job classification and language profile — different classifications require different profiles — however the SLE itself does not necessarily distinguish between classifications, and neither did this study. Three English-dominant employees were selected to participate.

Table 3: Characteristics of employee participants

No.	Pseudonym	Years Worked	Current Profile	Goal
1	Holly	10	E/B/B	E/C/C
2	Rodney	18	E/B/C	*
3	Abigail	32	C/C/B	Oral C

As we can see in Table 3, the years of service ranged from 10-32 years. Each employee has gone through the rounds of testing several times throughout their career. At the time of the study, all three employees held different profiles, and had different goals.

In addition to the data from government employees, data were collected from Bruno and Willow, two French teachers who have experience training employees for the test. They work at a private language school in the NCR. The school creates its own materials for general language development, however due to its location, its customer base consists largely of public service employees and has access to the CSPS language

training materials. Their interviews and perspectives supplement the conversation regarding the research aims of this study.

3.4 Researcher Positionality

It is important to note my role as a social other within the context of this study. I myself am an anglophone who has worked for nearly six years as an English language teacher, with two years of experience preparing French government employees for their SLE tests in English. I have heard and experienced the frustrations from many employees, as well as experienced the difficulties associated with training employees for these tests. It is this work that inspired this study and therefore influences my interpretation of the data.

3.5 Instruments

Three instruments were used to collect data for the study: a shortened practice test, and a semi-structured interview with the test-takers, as well as French teachers.

3.5.1 Practice Test

For the purposes of this study, I created a shortened sample of a standard practice test, which can be freely obtained from Canada School of Public Service (CSPS, 2014). Normal practice tests are 65 questions in length and begin with easier questions and progress through to more difficult ones. To maintain a level of difficulty reminiscent of the original, the practice test was divided into quarters, and a random number generator was used to select 7-8 questions from each section. The final instrument to be used consisted of 30 questions. I considered this to be a more effective way to gather the data for this section as a normal practice test would take 1.5 hours to complete. I was looking

for something that could be completed within 30 minutes to ease the cognitive burden on the participants.

3.5.2. Interview - Employees

Interviewing is a commonly used qualitative research method as it allows for the participants to share their experiences, feelings, and attitudes towards the topic under discussion (Saldaña, 2011). The interview was tailored in a way that would allow me to address the research questions guiding the study and delve deeper into the experiences and opinions regarding the TWE, the SLE, language training and their use of French in the workplace. The interview followed a semi-structured format with the purpose of facilitating more of a conversation and address any topics that could arise (Saldaña, 2011). The questions in the interview covered the following topics: (1) what they thought the test was assessing, (2) if/how they use their second language at work, (3) their experiences with language training (see Appendix C for the questions).

3.5.3. Interview - Instructor.

The purpose of the interview with the instructors was to gain a complementary perspective about the test and what it is assessing, as well to gain insights to any washback effects from the teaching perspective. Similar to the interview with the employees, this interview followed a semi-structure format (see Appendix D for the questions).

3.6 Procedures

Upon receiving Ethics approval (Appendix A), the participants were recruited. Data collection commenced on April 3, 2020, and was completed on April 26, 2020. An

initial interview was arranged to provide more information to the potential participants and allow them the chance to ask any questions. Once they agreed to participate and confirmed they understood what would be required of them, a date was scheduled to conduct the practice test, with the semi-structured interview immediately following. Due to the pandemic at the time of data collection, the study procedures were administered online using Zoom, an online web-conferencing platform.

3.6.1. *Verbal Protocol*

Using a verbal protocol does not mimic a real test-taking scenario exactly, however for the purpose of this research, I was not interested in this aspect. I was more interested in having the test-takers demonstrate either their implicit or explicit L2 knowledge when dealing with test items they would encounter while taking the test. Based on my experience teaching French-dominant employees, I anticipated four types of responses: responses based on grammar and vocabulary, debating between two choices, answers that could not be explained, and some guessing.

While the participants completed the practice test, they were asked to perform a metalinguistic concurrent verbal report in an attempt to elicit both implicit and explicit language knowledge (Bowles, 2010). A concurrent verbal protocol, or *think aloud*, is a process whereby the participant verbalizes their thought processes at the time of the task, thereby providing insight into the cognitive processes employed when completing the task. A retrospective verbal report is a process in which the participant is prompted by the researcher to share their thought processes *after* completing the task. A concurrent report was chosen over a retrospective one, as there is a risk with *veridicality*, which is the notion that the participant may not fully remember what they were thinking at the time

and could result in their report being inaccurate. Since the test is a multiple-choice test, they were instructed to do their best to provide reasons for why one answer was better over another one, thus tapping into their metalinguistic knowledge. Following Bowles' (2010) recommendations, instructions were provided as a guide for how to structure their responses. They were also advised that if they felt they knew the answer but could not quite explain, that they could simply state this reason, which would allow the procedure to tap into their implicit language knowledge. Additionally, a warm-up task was included to offer a short practice, thereby familiarizing the participants with the process, as well as to allow the opportunity for questions before the operational study began. The warm-up consisted of two questions, each with the two types of formats the employees were to encounter (i.e., fill-in-the-blank or identify the error). Once they felt comfortable and did not have any further questions, the test began. For each participant, the test took about 30 minutes to complete.

3.6.2. Interviews

The interviews with the test-takers were conducted immediately following the test. The interviews with the teachers were scheduled on separate days. All interviews lasted about 30 minutes.

Once all the data were collected, the recordings were transcribed. Pseudonyms were assigned to each participant, and any identifying data such as departments, places of work, or language schools attended were removed to ensure anonymity.

3.7 Analysis

The verbal protocol and the interviews were transcribed into a Word document and analyzed using qualitative coding methods to draw out recurring and relevant themes

related to the SLE in order to answer the research questions. Deductive and inductive reasoning was used, as well as a number of coding processes highlighted in Saldaña (2011) were applied throughout the process. According to Saldaña (2011), “codes function as a way of patterning, classifying, and later reorganizing each datum into emergent categories” (p. 95). Several rounds of coding were done as a way of analyzing the data from different perspectives.

To ensure coding reliability, a sample transcript was given to two experienced coders. The second coder is an expert within the field of language assessment and thus applied their own bottom-up codes. After some discussion, agreement reached 95%. The third coder, who is equally experienced in qualitative coding with a slightly different area of expertise, was given the coding scheme in Appendix G and instructed to code top-down. Agreement was 96% (see Appendix H and I for samples of primary researcher and second coder coding, respectively).

This chapter described the research methodology used to investigate the research questions. The next chapter will share the results of the qualitative coding, followed by a discussion of the findings.

Chapter Four: Findings and Discussion

The previous chapter shared the research methods used in this study. The following section will share the results of the qualitative coding that was used for analysis of the data in order to answer the research questions. This section is organized as follows: results of the coding, followed by a short profile of each employee participant, a comparison of the verbal protocol (VP) responses, and finally a summary of the findings with respect to each research question.

4.1 Results

4.1.1. Verbal Protocol

The first data to be coded were the VP. There are many different methods to coding verbal reports depending on the types of tasks, as well as the research questions guiding the study (Bowles, 2010). As the verbal report for the practice test was to inform the first research question, *What is the test assessing?*, coding related to the reasoning provided for their responses were prioritized over strategies such as *re-reading question* and *translating*. Given the multiple-choice design of the test, strategies such as *eliminating* and *debating* were prioritized in the coding, and is reflected in the coding scheme provided in Appendix J. There were four main categories that emerged from the analysis of the VPA data:

1. *Grammar*
2. *Response types*
3. *Certainty*
4. *Reasoning*

4.1.2. Interview Data

Interview data were coded using a number of coding methods as outlined in Saldaña (2011) were applied throughout the process to help make sense of the large amount of interaction and interplay evident in the data gathered (Saldaña, 2011). The first round of analysis applied the same codes from the VPA top-down to see if certain categories repeated. Once those categories were identified, I analyzed the data from the bottom-up using inductive reasoning to create new categories. Further analysis revealed that there were many tensions evident in the participants' reported experiences. As such, a *versus coding* method was applied to highlight the conflict with which the stakeholders were dealing (Saldaña, 2011). The versus coding was then categorized as *conflicts*, and contained the following sub-categories:

1. in theory vs. in reality
2. English vs. French
3. Employee vs. SLE

I analyzed the interview transcripts in relation to the research questions, as well as themes identified in the literature review. After the coding scheme was refined, eight major categories from the interviews were organized as such:

1. SLE,
2. Conflicts,
3. Perceptions,
4. TLU,

5. Language training,
6. Washback,
7. Impact, and
8. Threats to Validity

I reflected on these categories, and as I reflected on the seventh one, *impact*, I searched for broader themes to help understand how everything relates to one another. As I looked at these issues from the top-down in relation to the overarching topic of the study — the SLE — four themes became apparent:

1. SLE is a business
2. SLE is an investment
3. SLE is a gatekeeper.
4. SLE is a game.

The full breakdown of codes, categories, and themes can be found in Appendix G.

4.2. Employee Participant Profile

4.2.1. Holly

The first participant to be interviewed, Holly, was a government employee of 10 years who was the only one out of the other two in language training at the time of the study. She reported having gone through the rounds of testing two or three times in her career. Her profile was E/B/B. She stated that her goal was to get E/C/C, as she was in several pools for a promotion, which would require an Oral C. In terms of response style, she preferred to read the questions out loud, and exhibited the most metalinguistic ability

out of the other three; each reason she gave was accompanied by an explicit language rule. For example:

Holly: Question 9: “..it’s gonna be three or four because it’s singular, it’s [la note], [regler] is the correct vocab in this sentence, not [payer], and it’s feminine and it’s also after [être] so it has to be number 2.”
(correct answers, correct reasons).

4.2.2. Rodney

Rodney had worked in the government for about 18 years. After some fluctuation in his levels, he held an E/B/C at the time of the study. He reported being in and out of language training over the previous year in order to renew his Oral C. Six years ago, he did 10 months of full-time group training. This past year, he completed roughly 100 hours of part-time training in both group and individual settings. Over the course of the year previous to the study, he had attempted the oral test four times to finally reinstate his C level.

His response style varied; he began responding in French but switched to English as the test progressed. He scored 19 out of a possible 30. Out of the three participants for this section, he seemed less confident in his answers; he admittedly guessed three of the questions, in which two of those three were guessed correctly. When he was debating between two options, he chose the correct answer more often than the incorrect one. Rodney generally gave an explicit explanation for why the answer he was choosing was correct; he did have to be prompted at times to give more explanation for the other options in the multiple-choice distractor set.

4.2.3. Abigail

Abigail has worked in the government for 32 years and has gone through the testing process several times. Her current profile is C/C/B. Abigail has had a great deal of experience with studying French. She studied French in university and received a BBB upon graduation when she first took her government tests. She had taken training through college programs, as well as through her agency. Her most recent training was September-December of 2019 to achieve the Oral C, which she has been unable to get, although she feels as though she may be close. At the time of the study, Abigail was on a break from language training.

Although she sounded confident in her answers, she scored the lowest of the three, with 16 points out of 30. Abigail seemed to rely on her implicit language knowledge more-so than the others, with more references to *looks right* or *sounds right* in her reasoning. For example, question 24 gave everyone some trouble:

Abigail: ...going between number one and number four ... number one looks better than number four does.

For this test at least, her implicit knowledge did not yield correct answers. With five of her responses using *looks right*, *sounds right*, signaling implicit knowledge, or guessing, only two were correct. What follows is a comparison of correct and incorrect answers, along with the reasons provided.

4.3 Comparison of VPA Results

Table 4 shows the scores from each participant. According to the test scores from the TWE practice test, it would appear that Holly is the best writer, followed by Rodney, and finally Abigail. The participants often switched between English and French, as such French words and phrases in their transcriptions are marked by italics.

Table 4: Participants' scores out of 30

Participant	Score
Holly	25
Rodney	19
Abigail	16

4.3.1. Correct Answers

For 12 of the questions, all participants selected the correct answer. For nine of those correct answers, the same reasons were provided.

- Ils _____ s'assurer de ne pas contrevenir aux lois sur la concurrence.
- Question 12
1. auraient dû
 2. auraient du
 3. auraient due
 4. auraient dut

Figure 4: Question 12

For example, question 12 (Figure 4), each participant identified the correct form of the past participle of the French verb *devoir* as being spelled *dû*.

Holly: That's the right spelling for the *participe passé* for *devoir*.

Rodney: *dû* needs the little hat whereas number 2 doesn't have the little hat...

Abigail: dû, needs an accent circonflex...

Question 15
<p>Quel mot ou groupe de mots complète le mieux le texte?</p> <p>_____, j'ai téléphoné à Zabulon.</p> <ol style="list-style-type: none">1. En retourner dans mon bureau2. De retour à mon bureau3. Revenu à mon office4. Au retour dans mon bureau

Figure 5. Question 15

Question 15 (Figure 5), in their elimination of the distractors, each participant identified option three as an anglicism and proceeded to select the correct expression which in this case was option two.

Question 21
<p>Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».</p> <p>Nous aimerions savoir si M. Bouchard a l'intention d'accepté (1) le poste de vérificateur principal à Toronto à la suite du concours qui a eu lieu (2) dans cette région le mois dernier. La date d'échéance pour donner sa réponse est fixée (3) au 20 juin.</p> <ol style="list-style-type: none">1. d'accepté2. a eu lieu3. fixée4. aucune correction

Figure 6. Question 21

In question 21 (Figure 6), they all identified the correct error and provided the correct rule:

Holly: ...it's number one because you have *l'intention de* plus the infinitive so it's *accepter* rather than *accepté*.

Rodney: ...the first one there is wrong 'cause it would be *accepter* avec *e-r*, the infinitive in this case.

Abigail: number one is incorrect because it needs to be infinitive after the preposition *de* ...

In some instances, everyone chose the correct answer, yet different reasons were given.

For example, in question 13 (Figure 7):

Question 13

Question 23
Quel mot ou groupe de mots complète le mieux le texte?
Ils _____ s'assurer de ne pas contrevenir aux lois sur la concurrence.
1. auraient dû

Figure 7. Question 13

Holly: I think it's number one because it's unusual to have two R's in a French word so I'm gonna go with number one.

Rodney: I just don't know if it's one R or two R's so it's kind of a trick. I would go with one, I guess...

Abigail: ...this would definitely be one that I would be guessing at ...

and in this case I would pick number one, just because it looks right.

Here we can see that Holly knew a rule she had learned in language training, Rodney was unsure but was on the right track, and Abigail was admittedly guessing; three different reasons that resulted in a correct answer.

4.3.2. Incorrect Answers

There were only two instances where all three participants were wrong in their choice: questions 20 (Figure 8) and 22 (Figure 9). Both were an *identify the error* type of question:

Question 20
<p>Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».</p> <p>Vu les circonstances exceptionnels (1), l'agence accordera une semaine de congé de plus à tout le personnel. Vous bénéficierez (2) d'une période d'une année financière pour vous prévaloir de cette semaine de congé. Veillez (3) à respecter cette durée.</p> <ol style="list-style-type: none">1. exceptionnels2. bénéficierez3. Veillez4. aucune correction

Figure 8. Question 20

Interestingly, all three incorrectly chose the same option —number three — for what they thought the error was. While the correct answer was option 1, which required the adjective to agree in gender and number with the noun to which it refers — *circonstances* — the participants thought there was something wrong with the third option, *Veillez*.

Rodney was close; he was debating between option one and three and reported that it could be an issue with agreement with the adjective in option one. Regardless, he chose option three. The second instance where each chose the wrong answer was question 22 (Figure 9):

Question 22
Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».
Le représentant ministériel en Europe doit <u>prémouvoir (1)</u> et administrer <u>conjointement (2)</u> avec le représentant de la CE (Communauté européenne) et le directeur des Affaires ministérielles et internationales, un programme d'assistance technique pour les pays <u>émergents (3)</u> .
1. prémouvoir
2. conjointement
3. émergents
4. aucune correction

Figure 9. Question 22

Both Holly and Abigail thought there were no errors thus choosing option four; Rodney thought there was a spelling mistake with option three, *émergents*. The answer to this question was related to the verb *prémouvoir*. While it looks like a perfectly fine French word, it is in fact not a verb. The correct verb for this phrase would be *promouvoir*.

For the other questions, two or more people chose the wrong response. For example, question 1 (Figure 10), Holly and Rodney chose the wrong answer whereas Abigail chose correct:

Question 1
<p>Quel mot ou groupe de mots complète le mieux le texte?</p> <p>Il a refusé _____ nos offres.</p> <ol style="list-style-type: none">1. tous2. tout3. toute4. toutes

Figure 10. Question 1

Holly: I think that *offres* is masculine so I'm going to go with number one for masculine plural.

Rodney: *Je pense que c'est le premier, parce que c'est pluriel masculin.* [I think it's the first one, because it's plural, masculine.]

Abigail: I believe *nos offres* is feminine, so that's why I would pick 4.

The correct answer was with respect to the gender of the noun *offre*. Holly and Rodney both identified the plural noun, however they mistook the gender for masculine.

Question 5 (Figure 11) was another instance where each participant chose a different response:

Question 5

Quel mot ou groupe de mots complète le mieux le texte?

Ils nous ont promis _____ à haute vitesse.

1. une branche
2. une branchement
3. une connexion
4. une connection

Figure 11. Question 5

Holly: connexion is the correct one, connection with C-T is like an anglicism and *branche* and *branchement* are just not the correct words.

Rodney: Je pense que c'est une connection ... I think three is just a spelling mistake. It could be *branche*, *une branchement* ... I don't think the expression would be *une connexion*. Could be an anglicism, could be a trap but, I'll go with four.

Abigail: I believe it's number one, just because I believe *connection* is an anglicism and *branchement* would be negative so it's definitely not *une* so I would say that its *une branche*, number 1.

When we compare Holly and Abigail, both were correct in stating that *connection* was an anglicism. However, Holly correctly identified that *branche* and *branchement* were not the correct vocabulary for this answer. Rodney debated all options and ultimately chose the fourth one, which was incorrect.

4.3.3. Summary of VPA

The purpose of the verbal report was to provide an interesting look into the cognitive processes of these three individuals, along with the knowledge and skills that were measured by this test (Bowles, 2010). Since multiple-choice tests contain distractor sets, the verbal protocol was employed as a way to verify the extent of their L2 knowledge and see if the test-takers were accurately identifying the correct answers, and to what extent guessing contributed to their score. The test-takers were not forced to give a reason if they did not have one, allowing them to rely on their implicit language knowledge. Previous studies that have employed such methods have not found significant links between proficiency level and amount of metalinguistic ability (see for example Renou, 2000; Hu, 2002; Green & Hecht, 1992; Alderson, Clapham & Steel, 1997; Roehr, 2006). Therefore, even though response styles differed regarding amount of explanations provided, this should not be used to make a judgement of their L2 knowledge.

The results from the VPA will be used in the next section to contribute to answering the first research question in this study. In the next section, the interview data along with the VPA data will be combined to inform the remaining research questions.

**4.4. Research Question 1: What is the Test of Written Expression (TWE) assessing?
Are there any factors undermining the validity of this test? If so, what are they?**

In response to the first research question, this section provides the reported perceptions from the interview participants, as well as the data collected through the VPA. The second part of this question: *Are there any factors undermining the validity of this test? If so, what are they?* will be answered in the conclusion once all of the data have been considered. This section will present each participants' comments in comparison with one another.

From the VPA results alone, the TWE appears to be assessing grammar and vocabulary, as each answer from the participants to the 30 questions on the practice test included references to grammar or vocabulary. In addition to the verbal protocol, each participant was asked what they thought the test was assessing. Their comments are as follows.

Holly: ...It's an interesting mix, like I used to think it was just a grammar test, and I still call it the grammar test, but it is vocab as well, which is I guess why they call it *Written Expression*. Really, it's grammar. It's about all the rules and tricks that you have to know.

Rodney: That's the million-dollar question. I mean it should be testing your ability to write in your second official language. It's definitely mostly focused on grammar.

Abigail: I mean, it's supposed to test your ability to write effectively... Some of the questions are definitely intended to test

your vocabulary, but other questions are intended to test your grammar, which both of those are important parts of writing.

As we can see in their comments, all of the employees would agree that this is more a test of grammar and vocabulary rather than writing skills in perhaps the more common conception of writing (i.e., productive use). Given that the test is called the *Test of Written Expression*, the employees assume that it ought to test writing, however, they express doubts and uncertainty about whether or not the test is assessing their L2 writing skills. These doubts and uncertainties are consistent with what Stiggins (1982) reports as low face validity, a disadvantage of writing tests of this kind. While Abigail sounds doubtful, she does express a more understanding attitude when it comes to the test. Like the others, she references grammar and vocabulary, however in her interview Abigail conveys the necessity of these skills as a part of writing while adding a limitation to the tests' design.

Abigail: Somebody can be good at grammar and not necessarily a good writer. And by the same token, somebody can be good at vocabulary but not necessarily a good writer.

Abigail's comment suggests that perhaps the test is not assessing what it is intending to assess, signalling construct underrepresentation. Comparatively, the French teachers, Willow and Bruno, share the same perspectives as the employees regarding the construct of the test.

Willow: I will say grammar and syntax. Especially those two.

Interviewer: Anything else?

Willow: Sometimes also vocabulary, [and] for some students, expressions.

Bruno expressed a positive perspective towards the writing test. Similar to Abigail, Bruno expresses how one needs to know grammar and vocabulary in order to write effectively.

Interviewer: How well do you think this test assesses their writing ability?

Bruno: I think it evaluates a lot of aspects of writing skills. I think it's a good test.

Interviewer: Can you explain more?

Bruno: I think there [are] a lot of aspects of the test which help the students to improve their skills of writing and communicat[ion] during administrative correspondence.

When discussing other features of test design, both Holly and Rodney reference ways in which they perceive the test deceives the test-taker.

Holly: I hate the ones with the three underlined 'cause I feel like they're designed to trick.

She elaborates on the use of anglicisms as a form of deception when talking about having enough time to complete the written test. She reports:

Holly: ...Written Expression, I had enough time to do it but, partly that's because I didn't know some of them so I just kind of ... skipped past them...and I feel like it's designed to throw you off and to trip up anglophones, so there are lots of little anglicisms and stuff like that.

[...] So often, one of my techniques is to look for what seems right in English and then eliminate that question.

This comment from Holly also speaks to some of her test-taking strategies and how she decides to eliminate one answer. Such strategies have been identified as threatening the judgements and interpretations of test scores (Kane, 2013; Messick, 1996).

Complementary to Holly's comment, Rodney reports that he felt the test involved deception:

Rodney: I don't know if it's just a perception, but they do seem to be trying to trick you, and they don't make it easy.

Interviewer: What are some ways they try to trick you?

Rodney: Well they're just throwing in random spelling mistakes, so it sounds right, but it's spelled wrong, overly complicating the structure or the verb tense there to make it more difficult. Yeah, I don't know, it's a bit of a game.

These comments from Holly and Rodney that reference deception can be interpreted to be issues of fairness that stem from the technical quality of the tests' design as discussed by McNamara & Ryan (2011). If the test-taker feels the test involves too much deception or trickery, then this may result in feelings of unfairness towards the test. If the test is in fact unfair, this would then have ramifications for how this test is being used (Bachman, 2005; Kunnan, 2004; McNamara & Ryan, 2011).

After gathering their perceptions of the construct, the participants were asked how well they thought the test measured their French writing skills they must use at work. As

already mentioned, our French teacher Bruno felt it was a good test that measured important aspects of writing. The employees offer their perspectives.

Holly: So not at all because I don't write in reality. I don't know how many anglophones would actually write in French.

Rodney: Well, no, it doesn't really test your ability to write at all, [it] tests your grammar and how well you can give [corrections to] errors. 'Cause everything's already written for you, it's just multiple-choice, right.

Abigail: I don't believe that the writing test is that great of a tool. It would be a better measure to assess actually writing something and having that assessed. However, I do understand why they do it this way because it's easy, it's fast, it definitely assesses the principles that you need to be a good writer, but it doesn't assess the whole package.

These comments touch on three different issues in testing: representativeness to the TLU domain (Messick, 1987), the technical quality of the test items, and representativeness of the construct (McNamara & Ryan, 2011; Messick, 1987). Holly's remark that she does not write in reality would threaten the representativeness of the test score to the target domain (Messick, 1987). Rodney comments on the multiple-choice design of the test which in and of itself does not allow for productive writing, but rather, tests receptive skills related to identifying and understanding rules of grammar. This speaks to another disadvantage of testing writing indirectly, as it relies more on reading (Stiggins, 1982). Abigail's comment that the test does not assess "the whole package" is most concerning

as this is evidence of construct underrepresentation and would be a threat to the validity of the inferences drawn from the test.

4.4.1 *Summary of RQ1*

The information from the VPA, the interviews with the employees and language teachers suggest that the TWE is assessing grammar and vocabulary; however, as pointed out by Abigail and Bruno these components are necessary functions of writing. If we consider different constructs over time as outlined in Bachman (2007), tests that measure discrete elements of the language, such as vocabulary and grammar, fall under the category of *skills and elements*, and was a common method in the 1960s. This test design is in opposition with direct methods, such as performance or task-based assessments (Bachman, 2007; Stiggins, 1982). Had their comments made references to testing skills such as *communicating with colleagues* or *writing an email*, this would be considered to test communicative ability, which focuses more on language used in certain situations or to perform functions. Bailey (1996) would argue that since this test lacks a communicative focus, the design is outdated.

Several problems are evident after considering the comments with respect to the perceived construct of the test. Firstly, through the comments from the employee-participants, it is clear that this test has low face validity regarding the technical quality of the test (McNamara & Ryan, 2011; Stiggins, 1982). The participants made reference to the multiple-choice design, feeling as though it did not necessarily represent the “whole picture” of their L2 writing ability, indicating construct underrepresentation as put forth by Messick (1987, 1996). Additionally, as evidenced in the VPA, some of the points were received through guessing; if one or two points separates a B level from a C level, this

would then not be a fair and accurate representation of one's language ability. As Stiggins (1982) clarifies, there are trade-offs to testing writing in such a way; disadvantages include low face validity and relying on reading skills, whereas the advantages include clear and irrefutable scoring and low test costs. In comparison, the oral test is designed for productive use of speaking. Why the PSC has opted for a direct method of assessment for speaking, and not writing, is unclear. Moreover, Holly and Rodney referenced ways in which they felt the test was trying to trick them; feelings of deception could lead to feelings of unfairness. This would then violate Kunnan's (2004) fairness framework which states that tests should be fair to all test-takers.

Finally, all three of the employees do not believe that this test assesses their ability to write in their L2 at work. This topic will be discussed in relation to the second research question in the next section.

4.5. Research question 2: do the test tasks on the TWE correspond to those in the TLU domain?

This section provides data to answer the second research question. In the context of this study, language of the workplace corresponds to Part V of the OLA. A test that is assessing workplace bilingual proficiency should have test tasks that mimic the workplace tasks, as this would result in more accurate inferences regarding ability tested because of its representativeness to the TLU (Bachman & Palmer, 1996; Kane, 2013; McNamara & Ryan, 2011; Messick, 1987). To answer this question, the participants were asked about the TWE; additionally, since TLU refers to the real-world situation in which French would be used, they were asked in what other ways they use French at work to

offer a comparison with respect to reading, writing, and speaking. The conversations consisted of more emphasis on speaking in the office concerning meetings, presentations, and interaction with various colleagues (see Appendix C and D for a list of the interview questions; for additional details, see Chapter Three, Section 3.5). References to writing tasks revolved mostly around sending emails. In the section which follows below, each participant's account will be presented on a case-by-case basis, beginning with Holly.

4.5.1. Holly

Holly reported having very limited opportunities to use French in her workplace. She explained that she could read the official documents in French if she wanted to, however since those documents tend to be sent out in English and French, she opts to read them in English. When asked if she has any opportunities to write in French, she reports:

Holly: Zero, there would be none.

Interviewer: Zero, eh?

*Holly: Yeah, I mean, if I want to send a little email to somebody to say you know, *Can you help me with this, please?* Or just a short email, I could, but I don't have to. And I certainly wouldn't if it needed to be something important, like something professional that needed to be done properly.*

Based on this answer then, as Holly does not write in her job, the test tasks do not correspond to the TLU domain. Going beyond the scope of writing in the workplace, Holly references some of the ways she may have to speak French at work, however she reiterates the level of necessity in which she has to speak French.

Holly: If I was delivering a presentation... So, there are times I deliver a training program a few times a year at work, and it should be bilingual. So, if I could give some of the presentation in any French that would be good. It's more likely that I would do some of the introduction, which is what I have done in the past and then try to recruit a francophone employee, 'cause there's two of us that do it together.

Here we can see how English and French interact in her office on a day-to-day basis; there are tasks that *should* be done in French, however in reality, it is not always the case that English-dominant employees will be required to perform French tasks on their own; they have resources in the form of French colleagues to help them. Holly shares other ways she uses French at work.

Holly: I work in policy, I don't do service delivery so, I'm not gonna have a member of the public come up to me and ask, *How do I apply for a passport?*, that's not the kind of work I do. If I was supervising staff and then they have the right to be supervised in the language of their choice but then again yeah, the question is the quality [laughter] of the discussion.

She shares more of her frustrations:

Holly: One of the frustrating things in my office is that I've got a bunch of colleagues who have level C and they do not open their mouth. I've never heard them speak. So, I don't actually know what

level C sounds like because [...] the only people who I hear speaking French are francophones.

In terms of interaction between English and French in her workplace, Holly offers some insights in relation to how she interacts with various colleagues in her L2. She compares her interactions with the administrative staff and her boss.

Holly: I might talk with admin staff... I had francophone bosses who said to speak French with them and I thought *I'm not even comfortable speaking English with this particular guy* 'cause he was super shy, there's no way I was gonna sound like an idiot talking to my boss in a professional context.

In this example, Holly's issue is less to do with language and more to do with features of social interaction in general. She says the administrative staff are friendly and polite, and she feels more comfortable speaking French with them, however with an authority figure such as her boss, she feels uncomfortable in any situation, regardless of language. She further compares speaking with friends in unofficial contexts:

Holly: ...with friends too, I find it really hard to speak French with them because you have an equal relationship when you can speak one language well, and their English is fine, and to switch languages it then becomes a much more awkward interaction.

Holly's relative comfort with using French seems to be directly related to her relationship with her interlocutor (i.e., the person with which she is interacting).

In summary, Holly's workplace does not offer much incentive to use French. It would require a great deal of effort and motivation on Holly's part to use French

effectively. In terms of writing, she reports having no requirements in her current position. Should she get accepted by one of the job pools for a promotion, she speculates that she may have to use French more in the future.

4.5.2. Rodney

When asked about the ways Rodney writes in French in his position, he reports:

Rodney: A lot of our stuff goes out bilingual so, we have to do some quality control but, most often, if it's more than a couple paragraphs, we'll send it out to external translators and have to review it when it comes back.

To this extent, reviewing and checking for grammatical errors would match the tasks on the TWE. One caveat would be the TWE gives a hint as to where the errors could be through the multiple-choice design and distractor set; in the real world, Rodney may have to be able to identify the errors himself. In this respect, the test tasks do not match the TLU domain perfectly.

Similar to Holly, Rodney comments on how English and French interact in reality in his workplace.

Rodney: I'm in [department] so we do a lot of interviews. If we are interviewing somebody in Quebec, sometimes we'll have to do the interview in French.

Interviewer: And do you yourself do the interview in French?

Rodney: I've done a couple. I think the big thing for my job, why it's CBC is 'cause it's supervisory so, if I have an employee who wants to

do a performance review in French, then yeah I have to be able to do that.

Interviewer: And have you done any of those in French?

Rodney: I will if I have to but most of the time it's like we start in French and then switch to English at some point.

Interviewer: Why do you think that is?

Rodney: Well, people know that I am more comfortable in English than I am in French.

This is similar to Holly's comments about interaction with her boss. In this case, Rodney is superior to the colleague to whom he would be giving the performance review. This could explain why the conversation turns to English. Rodney goes on to discuss meetings:

Rodney: ... for bilingual meetings, I think the intention's good, and we have a lot of francophone people, so they tend to start in French and then if [there's] one unilingual person that doesn't understand French at all, then somebody's got to translate for them for real time and [it] just becomes easier for everybody to switch to English.

These types of comments show the disparity between French and English in the workplace which raises concerns with how section V of the OLA is being implemented in these offices; unfortunately, a solution is not so simple. Section V of the OLA states that employees have the right to work in their preferred official language. On one hand, the French-speaking employees should be able to participate in the meeting in French; on the other hand, if there is someone who is not bilingual, they too have the right to participate

in the language of their choice. As stated in the beginning of this study, French-speaking Canadians tend to be more bilingual, therefore they switch to English. The government of Canada spends a massive amount of money for English employees to go for language training to promote bilingualism, yet the offices reported in this study, are not operating bilingually. Similar findings are reported in Bessette (2005) and Gentil et al. (2010).

4.5.3. Abigail

Out of the three employees, Abigail reported feeling the most comfortable with French at work. On writing in the workplace, she states:

Abigail: I do from time to time, and I've had to craft some pretty high-level emails that have gone out across our agency so, I'm comfortable and competent doing that.

She goes on to discuss how she uses tools and resources available to help her when writing in the real world.

Abigail: I'm not going to say that I can write an email without having to check my work, and I do have to use some translation tools just to make sure that I'm using precise vocabulary, because, you know, when you are writing at that level it's important to make sure that you're not using more general terminology. It has to be the right verb so, working for [department] there's a lot of vocabulary that's very specific to our agency and in that case what I generally do is I'll use our intranet, I'll do a search for the term in English and then I'll flip the page over to French to see what the actual correct translation for that term is.

This comment demonstrates another way in which the writing test does not quite mimic real-world scenarios. When writing in an L2, people have access to resources such as dictionaries, translation devices, or internet services in case they need to search a meaning of a word or double-check the grammar. Language tests that do not allow access to such resources violate this very real-world function of writing in an L2.

4.5.4. The teacher's perspective

While the employees can provide an account of their TLU tasks, it is not uncommon for language teachers to be aware of how the employees use French in their work. Depending on the pedagogic approach a teacher takes, this could inform what they teach. The teachers who participated in the study shared their perceptions of what their students do, and if they had any knowledge of their students' job duties. Again, their comments, although somewhat limited and situated within the context of language training, contribute to the picture of the SLE.

Bruno: Some of them told me that they have no opportunity to communicate in French, because most of [their] colleagues speak in English and in order to be understandable, they communicate almost every day in English.

Willow was unsure of what exactly her students' duties are – she deals with different students in different positions on a daily basis – but she gives the example of students who have to write legal documents for their jobs.

Willow: Law, *des politiques, des lois* so, they need to write [those]. I'm not sure it's enough [SLE test preparation]. Emails between colleagues ... it might help them to communicate with other

colleagues. But for policies, they need more time and more contact with the language.

Taking Willow and Bruno's comments into consideration, we again see a lack of opportunity to use French at work. Additionally, the test itself does not necessarily have any tasks on it that require writing an email or writing laws, which were referenced in the interviews.

4.5.5. Summary of Research Question 2

This section reported details about the TLU domain, as told by the participants in the study. Speaking contexts included meetings, giving presentations, training sessions, performance reviews, and day-to-day interactions among colleagues. Reported writing tasks included emailing, reviewing documents that have been returned from translation services, or writing laws.

Abigail's case is perhaps the most interesting and likewise concerning as she is the only one of the employee-participants who reported actively writing in French and feeling comfortable crafting important documents. In her real-world context, she makes use of tools and resources such as translation tools, the Intranet site, or a French colleague for verification. Discussed at length in the literature review was how inferences are drawn from test scores; test scores are used to make judgements. Referring back to Table 3 above, looking solely at the test scores, it would seem that Holly is the best writer, followed by Rodney, then by Abigail. However, it could arguably be the opposite. Based on Abigail's comments in the interview, the inferences drawn from her test may be invalid because even though she scored the lowest, she reported feeling the most comfortable out of the three with writing in French, provided she has the right tools to

help her. This finding speaks to the limitation of the TWE's design and how it has the potential to inaccurately depict people's writing ability. Surely, if there were a productive writing element, and Abigail had access to resources, the scores would be more reflective of her true ability.

The denial of resources to complete the TWE is an additional limitation of the test's design whereby it is lacking in representativeness to the target domain thus signifying construct underrepresentation (Messick, 1987). When writing in the real world, we are able to use dictionaries, translators, or ask bilingual friends/colleagues for advice. Messick (1994; 1996) argues that tests should be designed to facilitate the best possible performance; Bailey (1996) stresses the value for allowing dictionaries. In questions on the practice test where they got the answer wrong because of gender agreement, a simple search in the dictionary would have solved this and improved the test-taker's score. In cases where they did not know a vocabulary item, they could search it to see if it matched the context; a spell-checker on a computer would identify spelling mistakes; in these cases, the scores would yield more valid inferences about writing in the real world.

After considering the comments from Holly, Rodney, and Abigail, it is clear that there are discrepancies between the way English and French ought to interact in the workplace and the ways in which they interact in reality. Moreover, there seem to be mismatched expectations of English-speaking employees in terms of how much French they themselves have to use. Rodney and Abigail hold management positions which require them to use more French than Holly. They are aware of the responsibility to interact with their employees in the language the employee prefers, as stipulated by the

OLA and the Policy on Official Languages. Nevertheless, the offices are not as bilingual as they should be which signals issues with implementation. These issues have been documented many times over the years through government reports (CH, 2018; OCOL, 2019); it is these types of issues that have led to the most recent attempt to revitalize the OLA. However, over the last 20 years, similar comments return regarding language of work in the public service (see Bessette, 2005; Gentil et al., 2010; OCOL, 2019). If nothing changes, these efforts will be in vain and the \$500 million the government plans to invest in the next ten years risks going to waste.

4.6. Research question 3: what, if any, are the washback effects of the test?

To reiterate, washback is the idea that the presence of a test will influence what is taught and consequently, learned (Alderson & Wall, 1993; Cheng, 2014). The Public Service Official Languages Appointment Regulation (DJ, 2005) states that employees may occupy positions regardless of language profile at the time of hiring provided they reach the language requirements within two years; to support these initiatives, employees are able to attend language training paid for by their department. The employees were asked about their experiences with such training, as well as how confident they feel to accomplish their SLE goals after training has completed. As discussed in the qualitative coding section (section 3.7), codes were focused on negative and positive washback on the content, teachers, and learners.

4.6.1. Holly

Overall, language training for Holly has greatly improved her knowledge of French and has helped her feel more confident to try to use it in her office. When she was asked how comfortable she feels using French at work, she reports positive outcomes.

Holly: Before I went on language training, really not comfortable at all, 'cause I feel like I'd make a fool of myself. If I got my C and then I went back to work, then I would probably try to use it more.

She discusses how if she reached her goal of Oral C, she would attempt to use French more. As we learned in the previous section however, using French in her office may be challenging, as other English-speaking employees do not seem to use French much.

In terms of the materials being used in her language training, Holly explains how at the moment, although she does not have to redo her writing test, she finds it helps her to learn and review grammar rules to improve her speaking.

Holly: The reason why I'm redoing my grammar exam now is to help my speaking. I don't think it helps you with writing, but it does help you with speaking because you need to know those rules and you need to know the vocab to be able to speak well.

Whether or not this is positive or negative washback could be debated. On one hand, studying past tests with no communicative focus would be negative washback on the content in terms of study materials (Bailey, 1996; Hamp-Lyons, 1997;). On the other hand, if this way of studying is helping her reach her goal, this could be considered positive washback on her because she feels it is improving her speaking as she can incorporate the grammar and vocabulary. It is interesting to note that several times

throughout her interview, Holly refers to the writing test as the *grammar exam*, which is something that I am not sure she was aware of half the time; this reinforces her perception of the construct of the test, as reported in section 4.4.1.

Classroom context seems to have an effect on learning and Holly's perceived progress towards her goal. She has tried group lessons as well as individual lessons. At the time of the study, she was in full-time individual training, which she felt was working better for her.

Holly: Now, I'm in full-time individual, I'm learning stacks more because I've got the time and it's every day. I do homework at night and then the next day, you kind of reinforce what you learned the day before.

As Holly stated earlier, she was studying the TWE as a way to improve her grammar and vocabulary; she describes her frustrations with her progress.

Holly: I found the grammar stuff frustrating. A couple of weeks ago I kept feeling like I was doing practice tests and I was still making loads of mistakes. Or I could tell you between two, which were the right ones, but I would often make the wrong choice. So, that was frustrating, 'cause it's like even with all this study and all this practice, I'm still like making all these mistakes, 'cause they're just such fiddly little rules that are really irritating at some point.

Her frustrations are evident, and likely due to the nature of the TWE, where it is exactly as she says—many different rules of French combined into one test. Holly reported feelings of *frustration* and as though she was not making much progress is evidence of

negative washback to the learner (Bailey, 1996). Regarding classroom study material, studying through past tests promotes negative washback on the content, as using a past test as a study material is perhaps not the best way to promote natural productive use of the language and potentially signifies a narrowing of the curriculum (Bailey, 1996; Wall & Hamp-Lyons, 1997). This consequently has a negative effect on the student in terms of long-term language development.

Considering her personal goals, Holly's goals for her writing test were beyond the requirements of her position. It is interesting to note how the tests are often referred to as *pass/fail*. The SLE is not designed as a *pass/fail* test; it is the employee who applies this term based on their own personal goals, and depending on the score they receive (i.e., A, B, or C), they decide for themselves if they have *passed* or *failed*. Holly reports she is not concerned about the written test, it is the speaking test that causes her uncertainty.

Holly: So I think I would get a B if I pass, I would like to get a C just because I like to do well [...] I only need a B because for EX you gotta be CBC so, for Written Expression, I'm not at all worried, I think that I'll get a B. It's my own goals that are what would get in the way. Speaking, another matter altogether [laughter], I don't know whether I'd get there or not.

Even though she has experienced frustration during the language training, the fact that she feels confident to achieve at least one of her goals is evidence of positive washback. Moreover, if she returns to work and is able to use more French as she hopes to, this would be further evidence of positive washback. However, there would be challenges

facing her in this endeavor, as she has previously commented on the disparities of French being used in her office.

Considering Holly's verbal account, it would appear that washback on her has been both positive and negative. Positive in the sense that through language training, she has improved her confidence in using the language, as full-time training has allowed her to feel optimistic about achieving her goals, at least with reading and writing. There is evidence of negative washback on her as the student because the writing test is arguably a grammar test with many different rules that lack a communicative focus. Studying through practice tests entails learning various rules from scattered contexts.

4.6.2. Rodney

Similar to Holly's comments, Rodney reports that language training has helped him to achieve his goals. The methods used in his training however would exhibit negative washback on teaching and consequently learning, as he reported studying through past tests.

Rodney: It's been good enough to get me the levels that I need, but it's more teaching the past exam than teaching to learn the language.

Again, this is evidence of negative washback as the curriculum has narrowed to the extent that it is the main source of input in this case. Furthermore, this type of comment is concerning because the array of literature on L2 development would argue that this is not a good strategy for promoting L2 ability. Current language teaching methodology being taught in graduate L2 teaching programs highlights communicative and task-based language teaching as this has been shown to improve use of the L2. For example, Nation's (2007) Four Strands approach vouches for meaning-focused learning at each

stage of a language program. Studying in the way that Rodney mentioned above is not thought of as beneficial for long-term language gains (Bailey, 1996; Burns & Richards, 2012; Ellis, 2000, 2015, 2016; Nation, 2007; Nation & Newton, 2008; Willis & Willis, 2007). This would then raise issues with the implementation of the OLA because once the employees pass the test, they return to work and may not use French anymore. One reason for this could be a result of what and how they studied in their language training.

Moving on to classroom context, Rodney took language training at a private language school. He shared some of his perceptions on the operations of the school in terms of the teachers he was given.

Rodney: When I did ten months full-time, we had a lot of different teachers. I think most language teachers are [paid] minimum wage and [it's] not really great work conditions, [there] tends to be a lot of turnover which makes it hard on people trying to learn. [It's] pretty much similar teachers and similar style no matter where you go.

While this comment from Rodney shares insights into operations within private-language institutions, this topic is beyond the current study. What is of relevance is his account of the negative effects of learning as a result of the school environment. Future research could examine this issue further.

Similar to Holly, there is a shared perception of the difficulty of the oral test. However, Rodney reports how he owes much of his success with his SLE to language training.

Rodney: The oral C is tougher, [...]. The fact that I got the E in comprehension is great and I do feel, compared to where I was before

I did full-time training, my comprehension is way better now than it used to be. [...] There's no way I could've gotten CBC or close to it without full-time training.

The positive washback to Rodney is clear in his comments, especially when we consider the previous section where it was reported that his position requires him to use French. Evidently, language training has played an important role in his success with the SLE.

4.6.3. Abigail

Much like the other two employees, Abigail owes success to the language training programs she has undertaken. With respect to the written test, she commented:

Abigail: The training courses do a really good job of preparing you for the tests. For me, the, the training course that I did was great, it was perfect, it gave me all the tools that I needed to succeed in the written test, which I got my two level C's in [reading and writing].

At the time of the study, Abigail had recently taken a break from language training in which she was aiming to reach her oral C level. While she has had success in the other two skills, the oral test still evades her. Abigail shared her experience with her current teacher who was helping her reach the goal. She explained:

Abigail: What was great about her was that she basically gave strategies to pass the test. I still didn't pass, but I feel like my last time was definitely much closer.

While on the surface, it is great that these employees are finding success through language training, teaching strategies and specifically focusing on test preparation are both signs of negative washback on the content, the learner, and potentially threaten the

validity of the inferences from test scores; as Rodney had mentioned, these methods do not focus on learning the language. What is notable is Abigail's uncertainty with how close she *feels* she is to achieving her level in oral communication. Not knowing what exactly she needs to do in order to reach the C level would further be considered a threat to validity in terms of construct underrepresentation as identified by Messick (1994).

This section reported the experiences of the employees; what is clear from the interviews are two things: all employees owe their success on the SLE to the language training they have received. For the TWE, practice tests are used as a main study tool in their training. The section that follows details the interviews from the French teachers to understand to what extent the test influences what and how they teach.

4.6.4. Bruno

Bruno had experienced teaching group lessons, as well as individual lessons to government employees. Earlier, Bruno's perceptions of the skills measured by the test, as well as the effectiveness of the test as a useful measure of the employees' writing skills were documented. While he remarked that the test measures useful aspects of writing, he adds a condition.

Bruno: If they learn grammars, vocabularies, and if they pass their test, if they don't practice in communicating with the same language, they will lose [their] communication [in] writing.

Bruno's comment is reflective of test preparation scenarios: once the student passes their test, their goals change. For example, Table 2 shared the current levels and goals of the employees; Rodney did not have a goal anymore after reinstating his oral C. Moreover, Bruno's comment suggests that after testing, the employees may not necessarily maintain

the knowledge that they gained during language training; Rodney additionally shared that he was concerned about maintaining his level, as the workplace is not necessarily conducive to language maintenance.

Bruno discussed the classroom context when doing test preparation and he shared some of the pressures and tensions he experienced when he has students in his class for test-taking purposes.

Bruno: The problem is some of them want to resolve exercises without preparing the basic grammar. So, sometimes they want to arrive on the goal, but they don't build the process of arriving there.

When asked about why he thought this was the case, he stipulated that the employees do not have enough time to thoroughly develop the language knowledge while also continuing to work at their jobs. This raises issues of what exactly the goals are with language training and the SLE that perhaps the government should consider. If the goal is to have a bilingual workforce, then something may have to change.

In terms of classroom context, Bruno shared some of the ways group lessons with mixed levels of students affect his teaching. He said:

Bruno: The student experience [i.e., knowledge of French] affects my class because I must begin by the level of the students, with [what] they need, and [their] difficulties. As soon as they have different levels, I don't teach in the same way. It depends on the level, the motivation, and the experience of the students with whom I am working.

Bruno shared some of the positive and negative effects of individual classroom context. On one hand, he was aware that the student was also a customer, and paying a large amount of money to be there; on the other hand, due to the one-on-one nature of the class, it was easier to help that individual.

Bruno: The advantage is there because [the] teacher must accommodate the payer... [the teacher] must adapt his teaching strategies to the person and to the test which is [being] prepared.

Through these comments, it would appear that the classroom context plays a role in the language training experience, a perception that was equally shared by our three employees. When he was asked if the TWE influences what he teaches, Bruno reported:

Bruno: It influence[s] my teaching because I must be sure that there is some basic grammar. If there is no basic grammar, I must help the students improve the basic grammar before continuing other strategies or aspects [of writing]. The grammar must be there in order to prepare good applications of writing skills.

When asked what he would do differently if there were no test involved, he said:

Bruno: I will ask them in everyday life, what they need in professional communication. And after identifying the needs in a professional context, I will plan some activities in order to help them reach there.

What Bruno is referring to here is common in current pedagogical approaches called a *needs analysis*. Generally, in these types of language for specific purposes contexts, it

would be this needs analysis that would determine in which situations the learner would need to use their L2; the language training activities would then be centered around those needs in order to achieve communicative ability (Burns & Richards, 2012; Ellis, 2000, 2015, 2016; Nation, 2007; Nation & Newton, 2008; Willis & Willis, 2007). However, because the test needs to be applicable to the entire public service of Canada, the test is very general in nature. It does not test language functions specific to positions, nor departments which Gentil et al. (2010) postulate is the reason that language gains are not maintained once employees return to work.

4.6.5. Willow

Willow reported the ways in which the test influenced her experience with language training. Regarding her methodology, she stated:

Willow: If they don't have to [take] the test, it's easier to go by sounds, just correct them, and you don't have to give them rules. But, if they are [taking] the test, they need to know the rules to be able to identify the mistakes. This is why we teach them more rules when they are [taking] the tests.

Her comment exemplifies the way in which the TWE is feeding back into her classroom teaching and subsequently how her students are learning. Since the test is mostly grammar rules and vocabulary, this is then the focus of such test preparation classes. This way of teaching lacks communicative focus and is not conducive to long-term language development. Most teachers can recognize that this is not the best way to promote L2 usage, as was evidenced in Bruno's comment above, however the influence of the test and the expectations of the student can be difficult for the teacher to manage. Willow

expressed the importance of teaching skills of grammar in order to help her students succeed.

Willow: So they need to know the rules, they need to understand them, to understand what's going on, so they can actually identify the mistakes.

Similar to Bruno, Willow expressed her perception that students who are preparing for their SLE level do not seem interested in learning about the language insomuch as they would rather get enough knowledge to pass the test.

Willow: Some of them like easy things so, they don't want to go deeper into the language. They don't want to stop and then to have a huge lesson about all of the exceptions, how you use it, and how we do it. It's different when they want to learn the language.

While this comment is not representative of every public service employee who attends language training, it does speak to Alderson and Wall's (1993) Washback Hypothesis where it says a test will influence the depth and rate of teaching and learning. It would appear that some students— at least ones that Willow has worked with— do not see the value in learning more about the language and this may cause tension for the teacher. Consequently, the students are not learning as much about the language as they could be. This could be due to the fact that they will likely return to monolingual work environments thus limiting the necessity of learning more about the language.

At private language academies that have a standing offer with the government (PSPC, 2020), it is not uncommon to see employees attend language training for varying

time periods. Some employees attend for two weeks full-time, others for two or three months full-time. Others do part-time, for example six hours a week. Teachers have to be aware of how much time the employee has to prepare for their level.

Willow: Sometimes I think that it's not enough [laughter]. Because to learn a language you need more time than that. ...It depends really on the positions.

Once the contract is finished, the employees return to work. Willow elaborates on the time needed to realistically develop lasting language skills:

Willow: But they need to practice more sometimes, and it's not enough to stay for two or three months, to have the language, and to be able to write in such a sustained, supported language.

It may be worthwhile for the government to consider what they hope to gain from sending employees to language training. If the purpose is to have a workforce of bilingual employees, then it seems the current system is not as successful as it could be.

Even though test preparation scenarios may not be the most pedagogically sound, it is nice when the teachers see their students succeed. These feelings of accomplishment are not only felt by the employees, but also the teachers when they receive good news about a student being successful.

Willow: It's good to see them progressing, I recently had a student who had a C instead of a B.

These feelings of accomplishment can be considered positive washback on the teacher, as well as the student.

4.6.6. Summary of research question 3

Several issues related to washback are evident in the interviews with the stakeholders. As stated in Alderson and Wall's (1993) Washback Hypothesis, a test will influence what is taught and consequently what is learned; evidence from the current study supports this conception. Positive washback on the learner was found in relation to the employee's achieving their goals (Bailey, 1996; Messick, 1996). Additionally, Rodney and Holly reported feeling more confident and able to use French when they return to work after completing language training. There was evidence of negative washback to the employee-participants when they discussed feeling stress, pressure, or frustration with language training and the SLE process. Additionally, the employees reported using past tests as study materials which indicates a narrowing of the curriculum (Bailey, 1996; Hamp-Lyons, 1997).

Both French teachers reported the test influences their teaching by focusing their lessons on teaching grammar, vocabulary, and how to identify errors, further evidence of narrowing the curriculum and thus negative washback to the program, as discussed by Bailey (1996) and Hamp-Lyons (1997). The teachers commented that if there were no tests, their teaching style would change. For example, Bruno would conduct a needs analysis and teach communicative functions; this method is popular with current pedagogy. Willow said she would focus on oral explanations and practice, an additionally effective method. Both methods subscribe to communicative language teaching (CLT) and are approaches popularized in the field of L2 pedagogy today (Bailey, 1996; Hamp-Lyons, 1997).

Washback is one consequence of testing that affects the classroom level. Messick (1996) argues that negative washback stems from threats to validity; in order to promote positive washback, these threats should be reduced as much as possible. Section 4.4 above suggested that there were elements of construct underrepresentation within the TWE and as a result, the employees do not feel as though it tests their full ability to write in their L2. The design of the test is evidently feeding back into the language classrooms and as such, writing is not necessarily being taught in a way that results in long-term learning. Bailey (1996) contrasts the notions of linguistic knowledge (i.e., rules of grammar) with linguistic competence (i.e., knowing how to use these rules productively); ideally, promoting the latter should be the focus of language training. Outdated testing methods, such as the design of the TWE, prevent such teaching, thus affecting learning, resulting in studying rules of grammar to identify errors to pass the test. Messick (1996) suggests direct assessments that contain tasks that mirror real-world tasks could solve this problem. Gentil et al. (2010) further support this notion by suggesting that if the test content was tailored to their position, they could develop vocabulary and language functions specific to their duties. If the employees were trained to use their L2 communicatively, this could promote confidence in using the language (Burns & Richards, 2012; Ellis, 2000, 2015, 2016; Nation, 2007; Nation & Newton, 2008; Willis & Willis, 2007).

Another concerning case that continues to present itself in the literature is that the employees report returning to fairly unilingual work environments after completing language training. Similar findings have been documented in Bessette (2005) and Gentil et al. (2010) and presents issues for language maintenance. Taking these cases into

consideration, in addition to the 2017-2018 annual report from the Commissioner on Official Languages, not much has changed in the last ten years. This is evidence of negative washback to the program (Bailey, 1996), that is to say that if the purpose of language training and the SLE is to ensure implementation of the OLA, the program is failing in this regard. Such broader consequences of testing, known as impact, are rooted in and extend to broader societal institutions (Alderson & Wall, 1993; Bailey, 1997; Hamp-Lyons, 1997; Shohamy et al., 1996; Wall & Alderson, 1993). The following section deals with this issue.

4.7 Research question 4: what is the impact of the SLE?

The previous section discussed the washback effects from the SLE which is the influence of a test on teaching and learning (Alderson & Wall, 1993). The impact section of the literature review discussed how testing in a situation, such as within the PSC, has effects that extend beyond the immediate testing context. For this section, I will use the four themes identified in the qualitative coding section to summarize the impact of the SLE. These themes are as follows: (1) the SLE is a business (2) the SLE is an investment, (3) the SLE is a gatekeeper, and (4) the SLE is a game.

Theme 1: SLE is a Business

The employees in this study reported owing their success on the SLE to the language training they had received. Between Holly, Rodney and Abigail, they have attended hundreds of hours of language training over the course of their careers. Rodney commented on how the SLE fuels a major business. While many departments and agencies have their own in-house program (Gentil et al., 2010), there are a number of

private language academies that receive a sizeable amount of business from government employees specifically for training employees for the SLE (PSPC, 2020).

Rodney: It's a huge business, especially in Ottawa. So many people need their levels and there're so many language schools out there trying to help people get their levels, but I don't think anybody has a really good understanding of the magic formula to get people there.

In the previous section, Bruno even referred to his student as the “payer”; and because of this, teachers need to adapt their teaching methods to help the *payer* achieve their goals. Evidently, if the government is spending millions of dollars on language training that money is going somewhere.

Building off of the SLE being a business, the SLE is a major investment from the employee as well as the PSC.

Theme 2: SLE is an investment

It was clear through the comments in the interviews that the SLE is an investment in two ways: (1) time from the student, and (2) and funds from the government.

Concerning time from the students, all three employees have spent numerous hours in language training. Holly had reported doing part-time training, which she felt was ineffective. Only once she had committed to full-time training did she feel as though she were making progress. Abigail additionally had been on full-time training preparing for her oral C, but at the time of the study she was taking a break because “It was just too much.” Rodney shares his insights on how to succeed the SLE:

Rodney: It just takes time, basically. You do have to try to immerse yourself in the whole second language, all these sorts of tips and tricks to do that. You basically just gotta commit yourself to doing it.

When talking about his most recent stint in language training, he says:

Rodney: It was pretty intense. In December, I was doing 5 hours a day, almost.

In comparison, while the employees feel as though they invest substantial time into their language training, the teachers commented on how they feel two or three months is *not enough* time to really develop one's language. Consequently, there exists tensions between employee, teacher, and the policy.

When sharing her frustrations about colleagues in her office, Holly says:

Holly: ...these staff that don't ever speak, like the government has invested thousands of dollars in your language training, and you're not using it. Why not?

Rodney also shared how there is pressure from the department leads.

Rodney: They're investing in your full-time or part-time training and you can't get your C back. After 2, 3, 4 years, eventually, you're not meeting the requirements of your position.

Again, this is an issue with the implementation of the OLA, a goal which is just (McNamara & Ryan, 2011), however it faced with many challenges. The department leads have a responsibility to ensure the act is being implemented, or they may face repercussions.

*Rodney: The first time I went full-time, definitely the employer really tries to instill like *Ok, [we're] paying for you to go [on] full-time language training. It's a big investment and it's up to you to maintain that.* There's no guarantee that if it expires, that they're gonna send you on training again to get it back.*

Perhaps this idea of “no guarantee” is one of the ways the government attempts to exert control, as Shohamy would argue (Shohamy, 2005; Shohamy et al., 1996).

Theme 3: SLE is a gatekeeper

Holly and Rodney had commented on ways in which the SLE affects career advancement. For Holly's current position, she meets the bilingual requirements, however, she reported being in a few pools for higher positions. This means if she were accepted into one of those positions, she would need an oral C. Rodney had the most to say on this matter:

Rodney: [It's] tons of stress too. If I want a promotion, I need it. It's definitely career limiting. I know a lot of people [who are] unilingual English and you kind of hit a ceiling where you can't progress any further in your career without being bilingual.

Rodney's comment does not speak for himself, rather he references a larger population of the public service of Canada. Evidently, if one wants to progress in their career, then being bilingual is necessary; the opportunity for career advancement can serve as motivation. Given the pressure and necessity of bilingualism the government is required to implement, it is all the more important to have tests that yield valid interpretations. Moreover, to support the investments that are being made, workplaces that operate

bilingually are crucial. Rodney shares one of the ways he was penalized as a result of losing his oral C status:

Rodney: They actually took away my bilingual bonus. Even though I was still BBB, my position was CBC so, after the second B, they're like Okay well, you're not meeting the requirements of your position, so you lose your bilingual bonus.

These types of comments are always interesting because here, he feels as though he is being punished for not meeting the language requirements of his position. Ignoring the policy aspect for a moment, evidently five years prior, Rodney had met the requirements of his position. It was only for a short period when he was unable to renew his C, that he did not. He was able to continue to occupy his position suggesting he continued to be more or less capable of performing his job. Yet from the policy perspective, he risked losing his job because he did not meet the language requirement.

Rodney: A lot of people too where they've had their levels five, ten years ago, and can't get it back. They have to get transferred out of their position 'cause they don't meet the requirements [of their position].

It is clear that he is not the only one who faces such repercussions. This leads us to the final theme.

Theme 4: SLE is a game

In some ways, the SLE is similar to a game that some people have no idea how to beat. For example, if we think of a game like chess that requires strategy to choose the

best move, attempting to predict their opponent's next move, the tests share similarities. For the TWE, several strategies mentioned by the participants were related to how they go about eliminating an answer; when the answer was unknown to them, they make their best guess. The VPA demonstrated how all three employees used this strategy in the practice test portion of the study; it was further mentioned during the follow-up interviews. Rodney said himself, "It's a bit of a game"; Holly references "tips and tricks" that need to be known to pass. Abigail discussed the strategies she has learned through coaching to try to reach the oral C level:

Abigail: You get to the point where you know you're being assessed.
[...] I have to make sure I use the right verb, I have to make sure that
I give three arguments and a counter argument, I have to make sure
that I nuance and blah blah blah and so, I've got all that going through
my brain and at the same time I'm trying to converse.

Through this comment, we can see how it is Abigail versus the oral test, a one-on-one battle. She has no idea if she won until she receives her report. People relying on strategies to succeed the SLE should invalidate the scores, according to Messick (1994), Kane (2013), and Cronbach and Meehl (1955).

Another way in which the SLE can be considered a game is related to, as Rodney discussed, maintenance. During Rodney's interview, he spoke about how his last round of part-time language training was to renew his oral C level. After four unsuccessful attempts, he finally achieved his goal. Abigail is in a similar situation, where she keeps attempting the oral test and continues to receive a B level. It is also not uncommon for people to redo the reading and writing tests when they are a few points off their goal.

Rodney: Now my problem is [I] only go every five years so, it's how do you maintain your level over those five years. Even when I was on full-time last time, there were a lot of people that had their levels five years before, and they're back full-time again.

In a way, there is a strategy to timing as to when to take the SLE, another factor is that it is surrounded with uncertainty. For instance, Holly is trying to get ahead of the game in case she receives her promotion, though it is not yet guaranteed. Employees may question if they should take the test pre-emptively so that they automatically qualify for their next position; in the case they receive the promotion they may wonder when they will need to fulfill the language requirement. Other concerns may be related to not being able to meet the requirement after the department pays for their training, or how to maintain their level over the course of five years until it expires. These types of uncertainties cause a great deal of stress for the employees. What is at stake should be very clear for the stakeholders.

Through discussing the themes, there are several negative consequences that result from the uses of the SLE tests as reported by the participants in this study. The first is that the employee is the primary stakeholder at the mercy of the many systems surrounding them; unfortunately, they reported several negative experiences, as have been shared in this study. Such experiences included stress and pressure to achieve their language requirement, uncertainty regarding the score they will receive, and consequences that may result from not meeting the requirements.

Secondly, there appears to be a strong relationship between investments from the government, language training, and SLE testing, which has created a large business for

language schools in the region. Largely due to the nature of the SLE tests and specifically the TWE, there is evidence of narrowing of the curriculum to the extent that the test is the main focus, which is resulting in positive washback on the learner (i.e., achieving their goals), however more concerningly, negative washback to the program. If employees are exiting language training to return to mostly monolingual workplaces, this is negatively impacting the larger goals of the OLA and resulting in both the financial investments of the government, as well as the personal investments of time from the employee, going to waste; however, language schools are profiting greatly from this mismatch.

4.8 Are there any factors undermining the validity of the test? If so, what are they?

This thesis took a new liberalist perspective and defined *validity* as the extent to which the inferences drawn from a test lead to meaningful interpretations that, in an ideal world, result in fair decisions (Kane, 2010; Kunnan, 2004; McNamara & Ryan, 2011; Messick, 1987, 1994, 1996; Moss, 1998, 2013). After considering the data, factors that may be undermining the validity of the inferences from the TWE and consequently the SLE will be shared with respect to the findings in this study.

Beginning with test interpretation, test scores lead to inferences and decisions. If the inferences drawn from the TWE's test score are to assert that a public service employee can write with a certain level of proficiency, then this claim is questionable for several reasons. Firstly, related to the tests' design, the multiple-choice design as a writing assessment was perceived to have elements of construct underrepresentation, as exhibited by comments such as "it doesn't test writing at all" (Rodney), and "it doesn't assess the full picture" (Abigail). Furthermore, the employee-participants did not report

perceiving the TWE as a writing test; both employee-participants and teacher-participants reported it is more a test of grammar and vocabulary than writing. Such comments further signify low face validity meaning they doubt the efficacy of the tests' design. If the design of the test was improved, this could lead to better attitudes towards L2 language policies the government has in place (McNamara & Roever, 2006).

Secondly, the interpretation about the test-taker's ability to write in French may not be a fair and meaningful judgement for the relevant context to which the test applies. For example, Section 2.1.1. described the writing functions for levels A, B and C (TBS, 2016); it claims people writing at the B level can write short descriptive messages. However, if we look at the tasks in Section 2.1.1. and compare them with the TWE tasks (Appendix B), there are gaps between the workplace tasks and the test tasks; this is further supported in the comments from the employee-participants who reported: not writing at all, performing quality control checks, or writing but using resources to aide in the process. This issue is further complicated when we consider the work environment. While it may be bilingual in theory, in reality it does not appear to necessitate the use of French. In the case that an employee writes in French (e.g., with Abigail) they are able to use resources such as translation tools, dictionaries, or French-speaking colleagues. These issues ultimately reflect badly on the fairness of the test and could threaten the validity of the inferences (Bachman, 2005; Kane, 2010; Kunnan, 2004; Messick, 1996; Moss 1998, 2016).

Moving to the overt social context of SLE testing, intended and unintended consequences should be evaluated (Messick, 1996); in the case that negative consequences are found, the validity decision should be questioned (Cronbach & Meehl,

1955). This thesis has found evidence that may be important for the PSC to consider. Promoting bilingualism is a cultural value embedded within Canadian society, thus the use of the test is just (McNamara & Ryan, 2011). Social and cultural values of the public service lead to the definitions of the abilities for level A, B, C, etc. If values are reflected in test constructs (McNamara & Roever, 2006; Messick, 1987), this could entail the TWE values one's ability to identify grammar and vocabulary rather than productively use the language. While grammar and vocabulary are necessary components of language, they do not reflect the full picture of writing; meaningful use of the language should complement such knowledge (Nation, 2007). Unfortunately, these components of the test are washing back into the language classrooms and negatively influencing teaching and learning. This study found evidence of negative washback in terms of narrowing the curriculum by focusing on grammar and vocabulary and teaching using past tests; these are not effective strategies for long-term language gains. Researchers such as Kunnan (2004), Messick (1994), and Bachman (2005) argue that negative washback should threaten the validity judgement of a test.

Furthermore, if language training and testing in the government is a way to achieve bilingualism targets within departments, the program may be failing in this regard, as employees return to monolingual workplaces resulting in language loss. Bilingualism in the public service is institutional, but in some ways operates on a very individual level. These negative consequences warrant a closer examination into the SLE testing process and they could potentially invalidate the inferences drawn from the tests. For the TWE, unintended consequences would be unfair decisions made based on the test results. Further consequences may be the employee revisiting language training resulting

in more money being spent by the government, and more stress and anxiety for the employee as they continue to face language testing.

In the next and final chapter of this study, I will summarize the main findings, discuss their implications and consider the study's limitations.

Chapter 5: Conclusions

The main aim of this study was to explore qualities of validity of the TWE and to consider the impact the SLE has on Canadian society. The research was guided by the overarching concept of *validity* by using a new liberalist definition that includes qualities of test design as well as consequences of test use (Bachman, 2005; Cronbach & Meehl, 1955; Kane, 2010, 2013; Kunnan, 2004; Messick, 1987, 1994, 1996); consequences were considered in terms of the narrow definition of *washback* (Alderson & Wall, 1993; Bailey, 1996; Messick, 1996; Wall & Alderson, 1993), and the broader sense of *impact* (Hamp-Lyons, 1997; Kunnan, 2004; Shohamy, 2005; Shohamy, et al., 1996). Qualitative research methods were used to address the research questions in the form of a verbal report while completing a practice test of the TWE, follow-up semi-structured interviews with three employees, as well as two French teachers working in this field. The next section briefly summarizes the findings of the study with regards to the stated research question. Following this, the implications of the study are considered, followed by its limitations.

5.1 Summary and Implications of Findings

Several salient findings are apparent from the results of this study. First it appears there are issues with the technical quality of the test's design, as the employees reported uncertainty regarding its construct. Given that the test tasks are multiple-choice involving error identification and fill-in-the-blank, there is no productive writing element making it is unfair to assert that the test is assessing workplace writing ability. This signals construct underrepresentation and is a source of invalidity (Messick, 1987). Secondly, when the employees compared the test tasks to their workplace tasks, there was a

disconnect between the two; the only employee-participant in this study who reported feeling competent in writing as part of her work duties is able to make use of tools to help her, a luxury that is not afforded through the administration of the test. Through the interviews with the employees and the teachers, there is evidence that the TWE does not measure the full construct of writing as set out by TBS (2016). If this is the case, then the PSC responsible for administering the tests has a duty to the stakeholders to examine this further (Bachman, 2005; Kane, 2010; Kunnan, 2004).

The second main finding of this study relates to issues of implementation of the OLA on a number of levels, which results in several negative consequences. Firstly, there appears to be evidence of the test negatively influencing language training classes. While there is positive washback on the learner regarding the achievement of their goals on the SLE, there is perhaps more indication of negative washback to the program of PSC testing which directly impacts the public service of Canada in two ways: (1) financial investments and (2) a lack of implementation of the OLA. Technically speaking, since there is a considerable business around language training with the incentive of test preparation, this is an entire industry based around negative washback. Secondly, the Government of Canada invests large amounts of money for government employees' language training. The government is further planning on investing \$2.5 billion dollars into the current initiatives to revitalize the OLA, promote bilingualism, and improve French-English statistics. For years since the OLA was enacted in 1969, reports have consistently come back stating an issue with implementation of the OLA in various government offices. These issues with implementation are especially supported when considering the evidence from the employees in this study and others (Bessette, 2005;

Gentil et al., 2010) that report employees returning to mostly unilingual workplaces and thus losing their language abilities gained through training. As evidenced by the history since the OLA was initially enacted, it is perhaps not enough to plan modernizations of the act. If no significant changes to the entire process transpire, then in 10 years, we risk reporting the same dismal numbers that demonstrate a lack of progress in bilingualism. If the government intends to invest \$2.5 billion, it should be put to good use.

As a language teacher first and foremost, one potential area we can attempt to improve the implementation of the OLA is through the SLE and language training, as this would target a large population of the public service. If the tests were designed in a more communicative fashion that was more representative of the duties and functions of the various departments, this could feed back into the classrooms in a way that would be more conducive to learning language used in real-world contexts rather than studying past versions of tests. The government has to decide if it wants employees to pass a test, or to be functionally bilingual, the latter of which should be prioritized.

5.2 Limitations

This study had several limitations. First, case-study research aims to provide a thick description of one phenomenon bounded by the context, participants and researcher, thus limiting the ability to generalize from the findings. Secondly, this study only included English-dominant employees, and ignored a crucial population in the public service.

5.3 Future Research

There is little public research on the SLE process; the literature had detailed experiences with language training, with mentions of the oral test, however as of now

there has not been a thorough investigation into the constructs and consequently the validity of the SLE tests. This study aimed to contribute to a body of literature that hopefully sees more growth in the coming years. Regarding this study, future research could expand the context to investigate the constructs of TWE Canada-wide.

Additionally, a larger sample size of TWE test-takers combined with a differential item analysis would be better able to discriminate between high and low performers on the test; such data would contribute to the discussion about the validity of the writing test.

Secondly, a thorough exploration into the construct of the Oral Interaction test is warranted as this test seems to cause public service employees and teachers the most stress and uncertainty. Finally, and perhaps most importantly, there is a lack of research that compares the francophone and anglophone perspectives. It would benefit the field to include more francophone experiences and perspectives of language training and testing in the government. A continued investigation into the validity of the inferences drawn from the government tests would serve the literature well.

References

- Alderson, C. J., Clapham, C., & Steel, D. (1997). Metlinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research*, 1(1), 93-121. <https://doi.org/10.1177/136216889700100202>.
- Alderson, J., & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing*, 13(3), 280-297. <https://doi.org/10.1177/026553229601300304>.
- Alderson, J., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115–129. <https://doi.org/10.1093/applin/14.2.115>.
- Anderson, N. (1991). Individual differences in strategy use in second language reading and testing. *Modern Language Journal*, 75(4), 460–472. <https://doi.org/10.1111/j.1540-4781.1991.tb05384.x>.
- Association, A. E., Association, A. P., & Education, t. N. (2014). Standards for educational and psychological testing. Washington, DC: American Education Research Association.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34. https://doi.org/10.1207/s15434311laq0201_1.
- Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, & e. al. (Eds.), *Language Testing Reconsidered* (pp. 41-71). ON: University of Ottawa Press.
- Bachman, L. F., & Palmer, S. A. (1996). Test usefulness: Qualities of language tests. In *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
- Bailey, K. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13(3), 257-279. <https://doi.org/10.1177/026553229601300303>.
- Bessette, J. (2005). *Government French language training programs: Statutory civil servants' experiences*. [Master's Thesis, University of Ottawa]. ProQuest Dissertation Publishing. <http://dx.doi.org/10.20381/ruor-18406>.
- Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York, N.Y: Routeledge.
- Canada School of Public Service. (2014). *Préparation à l'évaluation du français langue seconde: expression écrite 1 et 2*. Retrieved from Government of Canada Publications: http://publications.gc.ca/collections/collection_2016/efpc-csps/SC103-29-4-2015-fra.pdf
- Canada, P. S. (2020, February 5). *Official language training services for the National Capital Region*. Retrieved from <https://www.tpsgc-pwgsc.gc.ca/app-acq/sflo-olts/index-eng.html>

- Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing*, 33(4), 453-472. doi:10.1177/0265532215593312.
- Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272. doi:10.1017/S0267190599190135.
- Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. Sage.
- Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy, & N. H. Hornberger (Eds.), *Encyclopedia of language and education (2nd ed.): Language testing and assessment* (pp. 349-364). New York, NY: Springer.
- Cheng, L. (2014). Consequences, impact, and washback. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1130-1146). Hoboken, NJ: John Wiley & Sons.
- Cheng, L., & Fox, J. (2017). *Assessment in the language classroom*. London, UK: Palgrave/Macmillan.
- Cizek, J. G. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, 17(1), 31-43. doi:10.1037/a0026975.
- Cronbach, J. L., & Meehl, E. P. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4), 281-302. doi:10.1037/h0040957.
- DeLuca, C., Cheng, L., Fox, J., Doe, C., & Li, M. (2013). Putting testing researchers to the test: An exploratory study on the TOEFL iBT. *System*, 41(3), 663-676. doi:10.1016/j.system.2013.07.010.
- Department of Justice. (2005, November 21). *Public service official languages appointment regulations*. Retrieved from <https://laws-lois.justice.gc.ca/eng/regulations/sor-2005-347/page-1.html>
- Department of Justice. (2017, September 21). *Official Languages Act*. Retrieved from <https://laws-lois.justice.gc.ca/eng/acts/o-3.01/>
- Ellis, R. (2000). Task-based research and language pedagogy. *Language Teaching Research*, 4(3), 193-220. doi:10.1191/136216800125069.
- Ellis, R. (2015). *Understanding second language acquisition*. Oxford: Oxford University Press.
- Ellis, R. (2016). Focus on form: A critical review. *Language Teaching Research*, 20(3), 405-428. <https://doi.org/10.1177/1362168816628627>.
- 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*, 5(2), 148-171. <https://doi.org/10.18485/esptoday.2017.5.2.2>.
- Fox, J., & Artemeva, N. (manuscript in final preparation). *Context reconsidered: Social theories in language assessment practices*. Routledge.

- Fox, J., & Cheng, L. (2015). Walk a mile in my shoes: Stakeholder accounts of testing experience with a computer-administered test. *TESL Canada Journal*, 65-86. <https://doi.org/10.18806/tesl.v32i0.1218>.
- Fox, J., Haggarty, J., & Artemeva, N. (2016). Mitigating risk: The impact of a diagnostic assessment procedure on the first-year experience in engineering. In *Post-admission language assessment of university students* (pp. 43-65). Springer, Cham.
- Fox, J., von Randow, J., & Volkov, A. (2016). Identifying students-at-risk through post-entry diagnostic assessment: An Australasian approach takes root in a Canadian university. In *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim*, (pp. 266-285).
- Gentil, G., Bigras, I., & O'Connor, M. (2010). Achieving bilingualism in the Canadian federal public workplace. In *Multilingualism at Work: From Policies to Practices in Public, Medical and Business Settings* (pp. 81-105).
- Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education*, 14(1), 75-97. <https://doi.org/10.1080/09695940701272880>.
- Green, S. P., & Hecht, K. (1992). Implicit and explicit grammar: An empirical study. *Applied Linguistics*, 13(2), 168-84. <https://doi.org/10.1093/applin/13.2.168>.
- Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing*, 14(3), 295-303. <https://doi.org/10.1177/026553229701400306>.
- Heritage, C. (2018). *Investing in our future: 2018-2023 action plan for official languages*. Retrieved from Publication - Official Languages: <https://www.canada.ca/content/dam/pch/documents/services/official-languages-bilingualism/official-languages-action-plan/action-plan.pdf>.
- Hu, G. (2002). Psychological constraints on the utility of metalinguistic knowledge in second language production. *Studies in Second Language Acquisition*, 24(3), 347-86. DOI: 10.101 7.S0272263 1 020030 1 7.
- Kane, M. (2010). Validity and fairness. *Language Testing*, 27(2), 177-182. <https://doi.org/10.1177/0265532209349467>.
- Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. doi:10.1111/jedm.12000.
- Kunnan, J. A. (2004). Test fairness. In M. Milanovic, C. Weir, & S. Bolton (Eds.), *European language testing in a global context: Selected papers from the ALTE conference in Barcelona* (pp. 27-48. <https://doi.org/10.1080/15434303.2014.937486>). Cambridge, England: Cambridge University Press.
- Markee, N. (1992). The diffusion of innovation in language teaching. In *Annual Review of Applied Linguistics* (Vol. 13, pp. 229-243. DOI: <https://doi.org/10.1017/S0267190500002488>). Cambridge University Press.

- McCulloch, C. (2018). Gatekeepers of ambition: Test-taker preparation practices on two high-stakes language proficiency tests. [Master's Thesis, Carleton University]. <https://doi.org/10.22215/etd/2018-13235>.
- McNamara, F. T., & Roever, C. (2006). Validity and the social dimension of language testing. In *Language Testing: The Social Dimension* (pp. 9-42). <https://doi.org/10.1111/j.1473-4192.2006.00117.x>. Oxford, UK: Oxford University Press.
- McNamara, T. (2007). Language testing: A question of context. In *Language Testing Reconsidered* (pp. 131-137). University of Ottawa Press.
- McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly*, 8(2), 161-178. <https://doi.org/10.1080/15434303.2011.565438>.
- Messick, S. (1987). Validity. *ETS Research Report Series*, 1987(2), i-208. <https://doi.org/10.1002/j.2330-8516.1987.tb00244.x>.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23. <https://doi.org/10.3102/0013189X023002013>.
- Messick, S. (1996). Validity and washback in language testing. *ETS Research Report Series*, 13(3), 241-256. <https://doi.org/10.1177/026553229601300302>.
- Moss, P. A. (1998). The role of consequences in validity theory. *Educational measurement: Issues and practice*, 17(2), 6-12. <https://doi.org/10.1111/j.1745-3992.1998.tb00826.x>.
- Moss, P. A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement*, 50(1), 91-98. <https://doi.org/10.1111/jedm.12003>.
- Nation, P. (2007). The four strands. *International Journal of Innovation in Language Learning and Teaching*, 1(1), 2-13. <https://doi.org/10.2167/illt039.0>.
- Nation, P. I., & Newton, J. (2008). *Teaching ESL/EFL listening and speaking*. Routledge.
- Newton, E. P., & Shaw, D. S. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, 23(2), 178-197. <https://doi.org/10.1080/0969594X.2015.1037241>.
- Office of the Commissioner of Official Languages. (2013, September). *Challenges: The new environment for language training in the federal public service*. Retrieved from <https://www.clo-ocol.gc.ca/en/publications/studies/2013/challenges-the-new-environment-for-language-training-in-the-federal-public-service>
- Office of the Commissioner of Official Languages. (2019). *Annual report: 2019-1019*. Retrieved from <https://www.clo-ocol.gc.ca/sites/default/files/annual-report-2018-2019.pdf>

- Polio, C., & Friedman, A. D. (2016). *Understanding, evaluating, and conducting second language writing research*. Taylor & Francis.
- Public Service Commission of Canada. (2018, December 31). *Test of written expression (654)*. Retrieved from <https://www.canada.ca/en/public-service-commission/services/second-language-testing-public-service/second-language-evaluation-writing/the-test.html>.
- Public Service Commission of Canada. (2019, May 7). *Second language evaluation — test of written expression*. Retrieved from <https://www.canada.ca/en/public-service-commission/services/second-language-testing-public-service/managers/sle-written.html>
- Public Services and Procurement Canada. (2020, February 05). *Official language training services for the National Capital Region*. Retrieved from <https://www.tpsgc-pwgsc.gc.ca/app-acq/sflo-olts/index-eng.html>
- Renou, M. J. (2000). Learner accuracy and learner performance: The quest for a link. *Foreign Language Annals*, 33(2), 168-180. <https://doi.org/10.1111/j.1944-9720.2000.tb00909.x>.
- Roehr, K. (2006). Metalinguistic knowledge in L2 task performance: A verbal protocol analysis. *Language Awareness*, 15(3), 180-198. doi: 10.2167/la403.0.
- Saldaña, J. (2011). *Fundamentals of qualitative research*. ProQuest Ebook Central.
- Shohamy, E. (2005). Tests as power tools: Looking back, looking forward. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe, *Language Testing Reconsidered* (pp. 141-152). Ottawa, ON: University of Ottawa Press.
- Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, 13(3), 298-317. <https://doi.org/10.1177/026553229601300305>.
- Statistics Canada. (2017, August 17). *An increasingly diverse linguistic profile: Corrected data from the 2016 census*. Retrieved from <https://www150.statcan.gc.ca/n1/en/daily-quotidien/170817/dq170817a-eng.pdf?st=-5u8z7QB>
- Stiggins, J. R. (1982). A comparison of direct and indirect writing assessment methods. *Research in the Teaching of English*, 16(2), 101-114.
- Treasury Board Secretariat. (2016, July 21). *Qualification standards in relation to official languages*. Retrieved from <https://www.canada.ca/en/treasury-board-secretariat/services/staffing/qualification-standards/relation-official-languages.html>.
- Treasury Board Secretariat. (2017). *Policy on official languages*. Retrieved from <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=26160>
- Turcotte, M. (2019, October 3). *Results from the 2016 census: English-French bilingualism among canadian children and youth*. Retrieved from <https://www150.statcan.gc.ca/n1/pub/75-006-x/2019001/article/00014-eng.pdf>

- Vaillancourt, F., Coche, O., Cadieux, M. A., & Ronson, J. L. (2012, 01). *Official language policies of the canadian provinces costs and benefits in 2006*. Retrieved from <https://www.fraserinstitute.org/sites/default/files/official-language-policies-of-canadian-provinces-rev.pdf>
- Wall, D., & Alderson, C. J. (1993). Examining washback: The Sri Lankan impact study. *Language Testing*, 10(1), 41-69. <https://doi.org/10.1177/026553229301000103>.
- Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford: Oxford University Press.

Appendix A: Ethics Approval Certificate

Office of Research Ethics
4500 ARISE Building | 1125 Colonel By Drive
Ottawa, Ontario K1S 5B6
613-520-2600 Ext: 2517
ethics@carleton.ca

CERTIFICATION OF INSTITUTIONAL ETHICS CLEARANCE

The Carleton University Research Ethics Board-A (CUREB-A) has granted ethics clearance for the research project described below and research may now proceed. 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 # 112386

Project Team Members: Ms. Ellen Nearing (Primary Investigator)

Prof. Janna Fox (Research Supervisor)

Guillaume Gentil (Second Reader)

Project Title: Exploring Validity: The Case of the Government of Canada's SLE Test of Written Expression

Funding Source (If applicable):

Effective: **March 13, 2020**

Expires: **March 31, 2021.**

Please ensure the study clearance number is prominently placed in all recruitment and consent materials: CUREB-A Clearance # 112386.

Restrictions:

This certification is subject to the following conditions:

1. Clearance is granted only for the research and purposes described in the application.
2. Any modification to the approved research must be submitted to CUREB-A via a Change to Protocol Form. All changes must be cleared prior to the continuance of the research.
3. An Annual Status Report for the renewal of ethics clearance must be submitted and cleared by the renewal date listed above. Failure to submit the Annual Status Report will result in the closure of the file. If funding is associated, funds will be frozen.
4. A closure request must be sent to CUREB-A when the research is complete or terminated.

Appendix B: Practice Test

Instructions for the Verbal Protocol:

1. Please read the question number out loud.
 - Ex., “Question number (10).”
 2. When you think you know the answer, state the answer and explain why. Try your best to include reasons for why another answer is wrong. For example:
 - “I think the answer is 2 because”
 - “I don’t think the answer is 1 because... I don’t think it’s 4 because...”
 3. If you feel you can’t explain why an answer is correct over another, that is ok. You can say, “I don’t know why,” or something similar.
-

Instructions pour le protocole verbal :

1. Veuillez lire le numéro de la question à haute voix.
 - Ex., "Question numéro (10)".
2. Quand vous pensez connaître la réponse, indiquez-la et expliquez pourquoi. Veuillez inclure les raisons pour lesquelles les autres réponses ne conviennent pas. Par exemple :
 - "Je pense que la réponse est 2 parce que ..."
 - "Je ne pense pas que la réponse soit 1 parce que... Je ne pense pas que ce soit 4 parce que..."
3. Si vous pensez que vous ne pouvez pas expliquer pourquoi une réponse est correcte plutôt qu'une autre, c'est bon. Vous pouvez dire "je ne sais pas pourquoi" ou quelque chose de similaire.

Warm-up.

EXEMPLE :

Question 1

Quel mot ou groupe de mots complète le mieux le texte?

Plusieurs personnes auront la chance d'obtenir une chaise ergonomique _____ les mois à venir.

1. par
2. sur
3. en
4. dans

Question 2

Laquelle des sections soulignées suivantes comporte une ou plusieurs erreurs? S'il n'y a pas d'erreur, choisissez la réponse 4 « aucune correction ».

Le ministre de l'Environnement a récemment prononcé un discours dans l'assemblée générale annuelle (1) de l'Association canadienne de l'hydroélectricité. Le ministre en a profité pour souligner les initiatives visant à encourager (2) les producteurs d'énergie renouvelable, les particuliers ainsi que les entreprises (3), et à réduire leurs émissions de gaz à effet de serre.

1. dans l'assemblée générale annuelle
2. les initiatives visant à encourager
3. les particuliers ainsi que les entreprises
4. aucune correction

Question 1

Quel mot ou groupe de mots complète le mieux le texte?

Il a refusé _____ nos offres.

1. tous
2. tout
3. toute
4. toutes

Question 2

Question 3

Quel mot ou groupe de mots complète le mieux le texte?

Ce courriel, _____ faut-il l'envoyer?

1. quelle adresse
2. à quelle adresse
3. à quel adresse
4. quel adresse

Question 3

Question 3

Quel mot ou groupe de mots complète le mieux le texte?

Ce courriel, _____ faut-il l'envoyer?

Question 4

Quel mot ou groupe de mots complète le mieux le texte?

_____ à votre question, sachez que je suis tout à fait d'accord avec vous sur ce point.

1. En retour
2. En réponse
3. En répondant
4. En réplique

Question 5

Quel mot ou groupe de mots complète le mieux le texte?

Ils nous ont promis _____ à haute vitesse.

1. une branche
2. une branchement
3. une connexion
4. une connection

Question 6

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Savait-tu (1) qu'Ariane avait été nommée nouvelle directrice adjointe? Je te **dis (2)** que la nouvelle s'est répandue comme une traînée de poudre **partout (3)** dans la division.

1. Savait-tu
2. dis
3. partout
4. aucune correction

Question 7 Les sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Salut Robert. Peux-tu me dire (1) à qui je dois envoyer le chèque en ce qui concerne le contrat de service Internet? Il ne faut pas que j'oublie de l'envoyer (2), car la date d'échéance est demain, le 30 avril (3).

1. me dire
2. l'envoyer
3. le 30 avril
4. aucune correction

Question 8

Question 15

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Salut Robert. Peux-tu me dire (1) à qui je dois envoyer le chèque en ce qui concerne le contrat de service Internet? Il ne faut pas que j'oublie de l'envoyer (2), car la date d'échéance est demain, le 30 avril (3).

Question 9

Quel mot ou groupe de mots complète le mieux le texte?

La note de frais _____ par le service des finances.

1. sera payé
2. sera réglée
3. sera réglé
4. seront réglées

Question 10

Quel mot ou groupe de mots complète le mieux le texte?

Ils _____ de notre mise en demeure hier.

1. ont chargé réception
2. ont accusé réception
3. ont accusé le reçu
4. ont reçu

Question 11

Quel mot ou groupe de mots complète le mieux le texte?

Si nous ne passons pas une commande d'ici peu, je crains que _____ de papier pour les imprimantes de l'étage.

1. nous manquons
2. nous manquerons
3. nous manquions
4. nous manquerions

Question 12

1. auraient dû
2. auraient du
3. auraient due
4. auraient dut

Question 24

Question 13

Question 23

Quel mot ou groupe de mots complète le mieux le texte?

Ils _____ s'assurer de ne pas contrevenir aux lois sur la concurrence.

1. auraient dû

Question 14

Quel mot ou groupe de mots complète le mieux le texte?

On embauchera _____ économiste cette année.

1. un nouveau
2. un nouvel
3. un nouvelle
4. une nouveau

Question 15

Quel mot ou groupe de mots complète le mieux le texte?

_____, j'ai téléphoné à Zabulon.

1. En retourner dans mon bureau
2. De retour à mon bureau
3. Revenu à mon office
4. Au retour dans mon bureau

Question 16

Quel mot ou groupe de mots complète le mieux le texte?

Pensez-vous que ça _____ la peine d'embaucher une personne pour une si courte période de temps?

1. veuille
2. vaille
3. soit
4. veut

Question 17

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Elles ont choisi de s'exprimées (1) devant le groupe pour faire taire les rumeurs voulant (2) qu'elles partent (3) toutes les deux pour des postes dans un autre ministère.

1. s'exprimées
2. voulant
3. partent
4. aucune correction

Question 18

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

On lui a demandé de sauvegarder (1) les données (2) dans une banque compatible à celle (3) de la direction des achats.

1. sauvegarder
2. données
3. à celle
4. aucune correction

Question 19

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Veuillez prendre note que la documentation que vous avez demandée (1) n'est plus disponible en format papier. Par contre (2), vous pouvez depuis (3) vous la procurer en ligne, en format électronique, sur le site Web des Publications du gouvernement du Canada.

1. avez demandée
2. Par contre
3. depuis
4. aucune correction

Question 20

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Vu les circonstances exceptionnels (1), l'agence accordera une semaine de congé de plus à tout le personnel. Vous bénéficierez (2) d'une période d'une année financière pour vous prévaloir de cette semaine de congé. Veillez (3) à respecter cette durée.

1. exceptionnels
2. bénéficierez
3. Veillez
4. aucune correction

Question 21

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Nous aimerions savoir si M. Bouchard a l'intention d'accepté (1) le poste de vérificateur principal à Toronto à la suite du concours qui a eu lieu (2) dans cette région le mois dernier. La date d'échéance pour donner sa réponse est fixée (3) au 20 juin.

1. d'accepté
2. a eu lieu
3. fixée
4. aucune correction

Question 22

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Le représentant ministériel en Europe doit prémouvoir (1) et administrer conjointement (2) avec le représentant de la CE (Communauté européenne) et le directeur des Affaires ministérielles et internationales, un programme d'assistance technique pour les pays émergents (3).

1. prémouvoir
2. conjointement
3. émergents
4. aucune correction

Question 23

Quel mot ou groupe de mots complète le mieux le texte?

C'est à vous de veiller à ce que le travail _____ à temps.

1. termine
2. a terminé
3. est terminé
4. soit terminé

Question 24

Quel mot ou groupe de mots complète le mieux le texte?

Il faudra évaluer les dépenses _____ par cette initiative.

1. encourrues
2. ancourrues
3. en courues
4. encourues

Question 25

Quel mot ou groupe de mots complète le mieux le texte?

_____ que je sache, nous ne sommes pas responsables de ce problème d'impression.

1. Pour
2. Pour autant
3. À ma connaissance
4. Aussi loin

Question 26

Quel mot ou groupe de mots complète le mieux le texte?

La première étape pour prévenir et _____ la violence en milieu de travail est de comprendre les différents types d'événements qui peuvent être jugés violents. Ce n'est pas aussi facile qu'on pourrait le penser.

1. nier
2. rayer
3. étouffer
4. enrayer

Question 27

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Nous désirons vous rapeller (1) que le système est toujours en période de rodage (2) et que des bogues (3) peuvent survenir à l'occasion. N'hésitez pas à nous signaler tout problème que vous pourriez avoir pendant la mise en œuvre du système. Merci.

1. rapeller
2. rodage
3. bogues
4. aucune correction

Question 28

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Puisque (1) les membres du personnel se sont vivement opposés aux nouvelles mesures de contrôle visant à réduire le montant (2) d'absentéisme, il a donc été décidé qu'une rencontre patronale-syndicale aurait lieu sous peu (3).

1. Puisque
2. le montant
3. sous peu
4. aucune correction

Question 29

Question 63

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Question 30

Laquelle des sections soulignées comporte une ou plusieurs erreurs? Si aucune des sections soulignées ne comporte d'erreur, choisissez la réponse 4 « aucune correction ».

Je vous saurais gré (1) de porter (2) ce communiqué à l'attention de (3) votre personnel.

1. saurais gré
2. porter
3. à l'attention de
4. aucune correction

Appendix C: Semi-Structured Interview Questions – Employees

1. What is your experience with taking the TWE?
2. What do you think the test is testing?
3. Do you use English/French at work? If yes, in what ways? If no, why not?
4. How comfortable do you feel using English/French at work?
5. How well do you think the test assess the written language skills that you must use at work?
6. Do you take language preparation courses? If yes, what has been your experience with that?
7. After taking a language preparation course, how confident do you feel to take the test and get your level?

Appendix D: Semi-Structured Interview Questions – Teachers

1. What has been your experience with preparing students for the test?
2. How does test preparation impact what and how you teach?
3. How well does the test assess the writing ability they must demonstrate in their second official language at work?
4. What would you say this test is testing?

Appendix E: Consent Form – Employees

Informed Consent Form

Name and Contact Information of Researchers:

Ellen Nearing, Carleton University, School of Linguistics and Language Studies

Tel.: [REDACTED]

Email: [REDACTED]

Supervisor and Contact Information: [REDACTED]

Project Title

Exploring Validity: The Case of the Government of Canada's SLE Test of Written Expression

Carleton University Project Clearance

Clearance #: 112386 Date of Clearance: March 13, 2020

Invitation

You are invited to take part in a research project because you are a public service employee who is currently in, or has recently completed test preparation courses for your SLE Test of Written Expression. The information in this form is intended to help you understand what we are asking of you so that you can decide whether you agree to participate in this study. Your participation in this study is voluntary, and a decision not to participate will not be used against you in any way. As you read this form, and decide whether to participate, please ask all the questions you might have, take whatever time you need, and consult with others as you wish.

What is the purpose of the study?

Government employees who hold positions that require a level of bilingual proficiency to perform their job must take the Test of Second Language Evaluation (SLE) in order to demonstrate their language proficiency. Little publicly available research has been done to date that looks deeper into what the tests are measuring. As such, the findings from the study will benefit the research field, government employees, and language schools. This will be a qualitative study that will examine what the Government of Canada's Test of Written Expression is measuring and if/how the scores drawn from the test are relevant to the domain (i.e., the workplace) in which it applies.

What will I be asked to do?

If you agree to take part in the study, we will ask you to:

- ☐ Complete a shortened version of a practice test while explaining your thought choices out loud. This is expected to last roughly 30 minutes.
- ☐ After completing the practice test, you will participate in one-on-one interviews to share your perceptions and experiences of the government testing process and how it relates to your workplace. This is expected to last roughly 30 minutes.
- ☐ The test-taking process and interviews will be audio-recorded. The recordings are necessary for analyzing the data and is required.

- ☐ The study will take place at an agreed upon location away from your workplace.

Risks and Inconveniences

We do not anticipate any risks to participating in this study. Some potential risks include feeling stressed from the simulated test-taking atmosphere, or anxiety when reflecting on your testing experiences.

Possible Benefits

Completing an extra practice test and discussing it with a professional will be of benefit to you when you go to take your test. Your participation in this study may allow researchers to better understand what specifically the writing test is testing, and the impact of the testing process on public service employees. Should the testing policy be reviewed by the government in the future, they could consult the study for information.

Compensation/Incentives

You will receive a \$15.00 gift card to Tim Horton's as a small token of appreciation for your participation.

No waiver of your rights

By signing this form, you are not waiving any rights or releasing the researchers from any liability.

Withdrawing from the study

If you withdraw your consent during the course of the study, all information collected from you before your withdrawal will still be deleted. You may request that your data be removed from the study and deleted by notice given to the Principal Investigator (named above) within seven days of data collection, (i.e, the practice test and interview).

Confidentiality

We will remove all identifying information from the study data as soon as possible, which will be within seven days of data collection.

We will treat your personal information as confidential, although absolute privacy cannot be guaranteed. No information that discloses your identity will be released or published. Research records may be accessed by the Carleton University Research Ethics Board in order to ensure continuing ethics compliance.

The results of this study may be published or presented at an academic conference or meeting, but the data will be presented so that it will not be possible to identify any participants.

You will be assigned a pseudonym so that your identity will not be directly associated with the data you have provided. All data, including coded information, will be kept in a password-protected file on a secure computer.

We will encrypt any research data that we store or transfer.

New information during the study

In the event that any changes could affect your decision to continue participating in this study, you will be promptly informed.

Ethics review

This project was reviewed and cleared by the Carleton University Research Ethics Board A. If you have any ethical concerns with the study, please contact Carleton University Research Ethics Board (by phone at 613-520-2600 [ext. 2517 for CUREB A or by email at ethics@carleton.ca).

Statement of consent – print and sign name

I voluntarily agree to participate in this study.

___ Yes ___ No

I agree to be audio recorded.

___ Yes ___ No

Signature of participant

Date

Research team member who interacted with the participant

I have explained the study to the participant and answered any and all of their questions. The participant appeared to understand and agree. I provided a copy of the consent form to the participant for their reference.

Signature of researcher

Date

Appendix F: Consent Form - Teachers

Informed Consent Form

Name and Contact Information of Researchers:

Ellen Nearing, Carleton University, School of Linguistics and Language Studies

Tel.: [REDACTED]

Email: [REDACTED]

Supervisor and Contact Information: [REDACTED]

Project Title

Exploring Validity: The Case of the Government of Canada's SLE Test of Written Expression

Carleton University Project Clearance

Clearance #: 112386 Date of Clearance: **

Invitation

You are invited to take part in a research project because you are language teacher who currently trains students for their SLE Test of Written Expression. The information in this form is intended to help you understand what we are asking of you so that you can decide whether you agree to participate in this study. Your participation in this study is voluntary, and a decision not to participate will not be used against you in any way. As you read this form, and decide whether to participate, please ask all the questions you might have, take whatever time you need, and consult with others as you wish.

What is the purpose of the study?

Government employees who hold positions that require a level of bilingual proficiency to perform their job must take the Test of Second Language Evaluation (SLE) in order to demonstrate their language proficiency. Little publicly available research has been done to date that looks deeper into what the tests are measuring. As such, the findings from the study will benefit the research field, government employees, and language schools. This will be a qualitative study that will examine what the Government of Canada's Test of Written Expression is measuring and if/how the scores drawn from the test are relevant to the domain (i.e., the workplace) in which it applies.

What will I be asked to do?

If you agree to take part in the study, we will ask you to:

- ☐ Participate in one-on-one interviews to share your perceptions and experiences of preparing students for the government testing process, and how it relates to your teaching practice. This is expected to last roughly 30 minutes.
- ☐ The interviews will be audio-recorded. The recordings are necessary for analyzing the data and is required.
- ☐ The study will take place at an agreed upon location away from your workplace.

Risks and Inconveniences

We do not anticipate any risks to participating in this study. Some potential risks include any emotions that may arise as you discuss your teaching practice with your students.

Possible Benefits

The chance to reflect on your teaching practice may benefit you in your future career. Your participation in this study may allow researchers to better understand what specifically the writing test is testing, the impact of the testing process on public service employees, and how test preparation affects what and how teachers teach in the classroom. Should the testing policy be reviewed by the government in the future, they could consult this study for information.

Compensation/Incentives

You will receive a \$15.00 gift card to Tim Horton's as a small token of appreciation for your participation.

No waiver of your rights

By signing this form, you are not waiving any rights or releasing the researchers from any liability.

Withdrawing from the study

If you withdraw your consent during the course of the study, all information collected from you before your withdrawal will still be deleted. You may request that your data be removed from the study and deleted by notice given to the Principal Investigator (named above) within seven days of data collection (i.e. the interview recordings).

Confidentiality

We will remove all identifying information from the study data as soon as possible, which will be within seven days of data collection.

We will treat your personal information as confidential, although absolute privacy cannot be guaranteed. No information that discloses your identity will be released or published. Research records may be accessed by the Carleton University Research Ethics Board in order to ensure continuing ethics compliance.

The results of this study may be published or presented at an academic conference or meeting, but the data will be presented so that it will not be possible to identify any participants.

You will be assigned a pseudonym so that your identity will not be directly associated with the data you have provided. All data, including coded information, will be kept in a password-protected file on a secure computer.

We will encrypt any research data that we store or transfer.

New information during the study

In the event that any changes could affect your decision to continue participating in this study, you will be promptly informed.

Ethics review

This project was reviewed and cleared by the Carleton University Research Ethics Board A. If you have any ethical concerns with the study, please contact Carleton University Research Ethics Board (by phone at 613-520-2600 [ext. 2517 for CUREB A or by email at ethics@carleton.ca).

Statement of consent – print and sign name

I voluntarily agree to participate in this study. ☐ Yes ☐ No

I agree to be audio recorded. ☐ Yes ☐ No

Signature of participant

Date

Research team member who interacted with the participant

I have explained the study to the participant and answered any and all of their questions. The participant appeared to understand and agree. I provided a copy of the consent form to the participant for their reference.

Signature of researcher

Date

Appendix G: Qualitative Coding Scheme

VPA Responses

- Grammar
 - Gender
 - Number
 - Tense
 - Conjugation
 - Agreement
 - Rule
 - Spelling
 - Preposition
 - Vocabulary
 - Phrases / Expressions
- Response type
 - Eliminating
 - Debating
 - Deciding
- Certainty
 - Unsure
 - Guessing
 - Confident
- Reasoning
 - Explicit
 - Implicit

SLE:

- TWE
- Language training
- French
- Reading Test
- Oral Test
- Levels

Conflicts

- Employee vs. SLE

- In theory vs. in reality
- English vs. French

Perceptions - in theory vs. in reality

- TWE
 - skills measured
 - difficulty
 - deception
 - strategy
- Oral
 - skills measured
 - difficulty
 - strategy
- Reading
 - skills measured
 - difficulty

TLU - English vs. French, in theory vs. in reality

- speaking
- writing
- vocabulary
- resources
- interaction
- context
- relationships
- power

Language training - SLE is business/investment

- Classroom context
 - Full-time or part-time
 - Group or 1 on 1
 - Department or private
- Maintenance

Washback

- Positive
 - Content
 - Teacher
 - Student
- Negative
 - Content

- Teacher
- Student

Threats to Validity

Impact

- SLE is an investment
- SLE is a business
- SLE is a gatekeeper
- SLE is a game

Appendix H: Primary Researcher Initial Coding Sample

Ellen: Ok. So now the test is called the test of Written Expression but in your opinion, what do you think this test is testing?

Holly: It's the grammar, all the grammar rules, because none of us actually have to write in French 'cause you can write in your first language, so ... I've sent little emails, in fact my mother-in-law, I text with her in French so it's useful for me for that, but really it feels to me like it's a grammar test, it's about all the rules and tricks that you have to know.

Ellen: Mmhm, and so you mentioned this a little bit but do you use French at work?

Holly: So, not really, I mean like in reality, not a lot. So I have francophone colleagues and sometimes there are conversations as part of team meetings in French. I am not supervising at the moment so I don't need it for that but, I'm in a couple of the EX01 pools so I need to get my oral C to pass that, and the idea is that you can then deal with your francophone staff in French. Whether a level C actually allows anybody to supervise staff in French is another matter all together.

Ellen: What do you mean by that?

Holly: I don't think that, even if you get level C, I don't think you're really good enough to have the kind of managerial conversations that you need to have, like particularly to be you know careful giving performance feedback and stuff like that, those kind of really nuanced conversations, I don't think a level C would get you there. And in fact, one of the frustrating things in my office is that I've got a bunch of colleagues who have level C and they do not open their mouth, I've never heard them speak so I don't know, I don't actually know what level C sounds like because I've never heard [it], the only people who I hear speaking French are francophones.

Ellen: Mmhm

Holly: So, your point of comparison is really skewed because you kind of always comparing yourself to a native speaker, but the people who have C don't use it. Except for in official context, so that's the only time that anglophones I see using

Grammar vs. TWE
TLU = V in theory vs. in reality
threat (CUR)
Real: world
rules-tricks
in theory vs. in reality
TLU
job duties (workplace construct)
oral C - SLE - possible promotion
threat ~ impact
perception - doubting
vs. SLE motivation
perception
TLU in theory vs. in reality
interactions (CUR)
frustrating
SLE ≠ TLU (threat)
(CUR)
in theory vs. in reality
perception "skewed"
(CUR)
"comparing" Emp. vs. Fr TLU
~ NS
context - speaking

Appendix I: Second Coder Initial Coding Sample

Ellen: Ok. So now the test is called the test of Written Expression but in your opinion, what do you think this test is testing?

Holly: It's the grammar, all the grammar rules, because none of us actually have to write in French 'cause you can write in your first language, so ... I've sent little emails, in fact my mother-in-law, I text with her in French so it's useful for me for that but really it feels to me like it's a grammar test, it's about all the rules and tricks that you have to know.

Ellen: Mhmm, and so you mentioned this a little bit but do you use French at work?

Holly: So, not really, I mean like in reality, not a lot. So I have francophone colleagues and sometimes there are conversations as part of team meetings in French. I am not supervising at the moment so I don't need it for that but, I'm in a couple of the EX01 pools so I need to get my oral C to pass that, and the idea is that you can then deal with your francophone staff in French. Whether a level C actually allows anybody to supervise staff in French is another matter all together.

Ellen: What do you mean by that?

Holly: I don't think that, even if you get level C, I don't think you're really good enough to have the kind of managerial conversations that you need to have, like particularly to be you know careful giving performance feedback and stuff like that, those kind of really nuanced conversations, I don't think a level C would get you there. And in fact, one of the frustrating things in my office is that I've got a bunch of colleagues who have level C and they do not open their mouth. I've never heard them speak so I don't know, I don't actually know what level C sounds like because I've never heard [it], the only people who I hear speaking French are francophones.

Ellen: Mhmm

Holly: So, your point of comparison is really skewed because you kind of always comparing yourself to a native speaker, but the people who have C don't use it. Except for in official context, so that's the only time that anglophones I see using

construct [validation] gap

grammar

① + TLU Domain = CUR

② TLU

grammar "rules and tricks"

workplace use - construct

① TLU

② TLU

③ TLU

④ TLU

⑤ motivation

⑥ CUR

CUR

CUR

CUR

TLU

TLU

CUR & Ellen -

CUR

NS

TLU