Linguistically and Culturally Diverse Students: Their Language Development, Assessment, and Support in the Public Education System

by

Hyunah Kim

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Applied Psychology and Human Development
Ontario Institute for Studies in Education
University of Toronto

© Copyright by Hyunah Kim 2023
Linguistically and Culturally Diverse Students: Their Language Development, Assessment, and Support in the Public Education System

Hyunah Kim
Doctor of Philosophy
Department of Applied Psychology and Human Development
Ontario Institute for Studies in Education
University of Toronto
2023

Abstract

This dissertation centers on the experiences of linguistically and culturally diverse (LCD) students in the K–12 public education system, with a specific emphasis on their language development and assessment. Historically, education systems have been designed to accommodate students from mainstream culture who speak the socially dominant language as their first language. However, recognizing the heterogeneous and complex multicultural backgrounds within the LCD student population, it is crucial to provide them with the necessary attention and support to thrive academically and in their lives. This dissertation comprises three empirical research studies conducted in Ontario, Canada, known for its cultural diversity and substantial immigrant population. By examining the experiences and outcomes of the growing LCD student population in the public school system, the dissertation aimed to inform the development of effective policies and practices to enhance their educational experiences and success.
In Study 1, the focus was on the development of multilingual identities among LCD students over time and its relationship with literacy achievement in school. By analyzing population-level cohort data, the study tracked the home language environments and English literacy performance of students between grades 3 and 10. A comparison was made between students from multilingual homes and those from English-dominant homes. The findings revealed a rapid shift towards English within student homes, yet also highlighted the benefits of maintaining home languages. Multilingual students who maintained their linguistic diversity exhibited superior literacy performance, emphasizing the positive impact of multilingual competence.

Building on the literacy achievement patterns of LCD students established in Study 1, Study 2 investigated the extent to which items on a large-scale reading test provide equitable opportunities for LCD students to demonstrate their learning. The study identified reading test items that assumed a high level of knowledge of mainstream Canadian culture and conducted a multi-group differential item functioning analysis across student groups with varying lengths of residence in Canada. Test items with high cultural sensitivity exhibited differential functioning, disadvantaging LCD students when compared to their Canadian-born English-first-language peers. The findings underscore the need for improvements in test construction to ensure fairness for LCD students.

Study 3 evaluated the support provided for the heritage language development of LCD students. This study was designed as a case study of Korean language classes in a publicly funded heritage language program in Ontario, with a focus on assessment and reporting practices. By capturing the perspectives of teachers and parents through individual interviews,
Study 3 provided valuable insights and practical suggestions to enhance the assessment practices within Korean classes and improve the overall quality of the heritage language program. Collectively, the three studies shed light on equity-related issues and the support received by LCD students within the school system. The findings have important implications for education policymakers seeking to address the needs of this diverse student population. Policymakers are urged to develop a deep understanding of the heterogeneity within the LCD student population and the fluid nature of their characteristics, formulate comprehensive policies that address their holistic needs, and place a stronger emphasis on policy implementation during the policy design phase. By incorporating these recommendations, policymakers can work towards creating an inclusive and supportive educational environment for LCD students in the K–12 education system.
Acknowledgments

The path to this milestone spanned seven years, intertwined by the arrival of my youngest child, navigating a pandemic, and balancing a full-time job. This dissertation would not have been possible without the unwavering support and care from many individuals, to whom I owe immense gratitude.

I have been extraordinarily fortunate to have Dr. Eunice Eunhee Jang as my supervisor, providing exceptional guidance in pursuing the research I am deeply dedicated to. She offered me numerous opportunities for professional growth and was always available for consultation when needed. Eunice, having you as my supervisor and friend is a privilege. As you mentioned, I eagerly look forward to our ongoing friendship and research. I extend my heartfelt thanks to my committee members, Drs. Becky Xi Chen and Esther Geva. I consider myself incredibly lucky to have had their support throughout this journey. Becky, I vividly recall our early collaboration during my early years in the program, and Esther, your initial encouragement when I first met you at your house in my first year was immensely motivating. Your continued support throughout my program is deeply appreciated. I am also grateful to my external examiners, Drs. Margaret Early and Enrica Piccardo, for their thorough readings and insightful feedback that refined my dissertation.

My gratitude extends to more faculty members with whom I have built relationships. I am thankful to Dr. Jeff Bale for including me in the Heritage Languages Program project during my early years and offering significant support for conference presentations and publications. This project deepened my understanding of Ontario’s education system and enriched my dissertation. Drs. Yoonkyung Lee and Jesook Song at the Centre for the Study of Korea, your invaluable feedback on the earliest version of Study 3 was immensely helpful. Dr. Anne Jordan, your words of encouragement over the years have had a profound impact on me, and I am forever grateful for your continuous support. Dr. Patricia Ganea, I deeply appreciate your teaching and our conversations during my first-year proseminar course and continued encouragement. Dr. Kang Lee, I had the privilege of being a teaching assistant for your course for many years and learning from your insights. Thank you.

Starting my doctoral journey at OISE as an international student in an unfamiliar environment with an uncertain path, I would not have survived without my incredible colleagues.
and friends. Drs. Jeanne Sinclair, Clarissa Lau, Megan Vincett, and Jeannie Larson: how I miss our daily interactions, laughter, and shared stories. You all have been the most wonderful companions one could ever ask for. Drs. Maryam Wagner and Maggie Dunlop: although you were part of the lab before my time, I consider myself extremely fortunate to have met and worked closely with you. Bruce Russell, Christie Barron, Samantha McCormick, Melissa Hunte, Liam Hannah, Dr. Lois Maplethorpe, Gina Park, and Meng-Hsun Hunter Lee: I am indebted for your support, care, patience, and understanding. Dr. Jennifer Burton, thank you for introducing yourself to me at the LPP conference and joining the manuscript writing journey alongside Tasneem. Your presence has shaped my graduate life in a significant way. Wonhee Lee: meeting you through the Korean Consulate in Toronto and collaborating on multiple research projects has been a stroke of luck. My dissertation would have been incomplete without you. To Hiwon Hwang and Sunkyu Lee: even though we have been apart for so long, your continuous support and love have bridged the distance, making me feel connected throughout days and nights.

I have gleaned valuable insights from numerous scholars and colleagues I encountered through professional organizations, including the International Language Testing Association (especially the Graduate Student Assembly), the American Educational Research Association, the Canadian Association of Language Assessment, the Canadian Association of Teachers of Korean, and the American Association of Teachers of Korean. I am grateful for the professional relationships and friendships I have forged through these organizations. Additionally, during my final years in the program, my full-time position at the Education Quality and Accountability Office greatly was instrumental in refining my dissertation. I extend my sincere appreciation for the invaluable insights and words of encouragement from my colleagues, as well as the flexibility they provided with my work schedule during the crucial final stages of completing this dissertation.

To the research participants for the studies included in this dissertation, thank you. While your names did not appear in my writing, you know that I am addressing you. As educators, you possess the power to make real changes to the lives of students whom I deeply care for. I gained valuable insights through our conversations, and your willingness to share your experiences and insights made my dissertation something real, transcending mere numbers and statistics in large datasets.
Additionally, I extend my gratitude to the esteemed organizations that supported me throughout this academic endeavor. The financial assistance and resources provided by the University of Toronto (including Connaught International Scholarship for Doctoral Students from the University, Lora Dolores Antonette Pallotta Memorial Scholarship and Doctoral Completion Award from OISE, and CSK Dissertation Writing Award and CSK Publication Support Award from the Centre for the Study of Korea), Social Sciences and Humanities Research Council of Canada, Ontario Graduate Scholarship, Duolingo English Test, the Language Learning journal, Korean Canadian Scholarship Foundation, International Research Foundation for English Language Education, International Language Testing Association, and American Association of Teachers of Korean significantly contributed to the successful completion of this dissertation. Their support not only facilitated my research but also encouraged the dissemination of knowledge, promoting advancements in language education and assessment. I appreciate their belief in the importance of this work and the potential impact it holds in the academic and broader community.

Lastly, I acknowledge that reaching this point was possible only through the support of my family. My parents, parents-in-law, and extended family members: I deeply thank you for your unwavering support, including multiple trips between South Korea and Canada to care for the children. My little ones, Alina, Celine, and Sian: you have always been the inspiration that kept me going, making this achievement possible. Last but not least, my lifelong partner, Dongmin Kim: you had no reservations leaving your successful career behind and boldly relocating our family to Canada, solely to support my pursuit of doctoral studies. I know I would have achieved nothing without your sacrifice and dedication. I love you to the moon and back.
# Table of Contents

Abstract .................................................. ii  
Acknowledgments ......................................... v  
Table of Contents ......................................... viii  
List of Tables ............................................. xi  
List of Figures ............................................ xii  
List of Appendices ........................................ xiii  
List of Abbreviations ...................................... xiv  

**PROLOGUE** .................................................. 1  

**INTRODUCTION** .......................................... 2  
  Theoretical Background .................................. 4  
      Linguistically and Culturally Diverse Students .... 4  
      Language Development of LCD Students .......... 7  
      Assessing LCD Students’ Language and Literacy Skills ... 11  
  Overview of the Dissertation ......................... 16  

**STUDY 1. CHANGE IN HOME LANGUAGE ENVIRONMENT AND ENGLISH LITERACY ACHIEVEMENT OVER TIME: A MULTI-GROUP LATENT GROWTH CURVE MODELING INVESTIGATION** .............. 18  
  Literature Review ....................................... 19  
      Home language environment and language shift ... 19  
      Variation in literacy achievement by home language ... 20  
      Latent growth curve modeling and its applications ... 22  
  The Present Study ....................................... 25  
  Method ..................................................... 26  
      Participants .......................................... 26  
      Measures ............................................. 27  
      Analysis ............................................. 28  
  Results .................................................... 29
RQ1: How do students and their families change their home language use between grades 3 and 6? 29
RQ2: How are changes in home language environment associated with students’ relative literacy achievement patterns between grades 3 and 10? 32
RQ3: Do immigration status and ESL program enrollment in grade 3 predict the relative literacy achievement patterns of students with different home language environments? 38

Discussion 40
Acknowledgments 45

STUDY 2. DIFFERENTIAL ITEM FUNCTIONING DUE TO CULTURAL FAMILIARITY ON A READING ACHIEVEMENT TEST: DOES THE LENGTH OF RESIDENCE MATTER? 46

Literature Review 47
Cultural Knowledge and Familiarity as a Source of Item Bias 47
Previous Research on DIF for Language Learners in Large-scale Assessments 49
DIF Analysis Approaches 52
The Present Study 54

Method 54
Participants 54
Measure 56
Analysis Procedure 58
Item Content Review 58
DIF Detection Methods 59
DIF Analysis Procedure 60

Results 61
Identification of Items to Be Tested for DIF 61
DIF Analysis 62
Discussion 68
Acknowledgments 73

STUDY 3. ASSESSING HERITAGE LANGUAGE LEARNERS IN ONTARIO’S INTERNATIONAL LANGUAGES ELEMENTARY PROGRAM: A CASE STUDY OF KOREAN CLASSES 74
List of Tables

Table 1.1 Sample Sizes, Means, and Standard Deviations of Literacy Assessment Performance by Group 33

Table 1.2 Results from the Fitted Multi-group LGCM without Covariates (Left) and with Covariates (Right) (n = 18,000) 35

Table 1.3 Pairwise Wald Test Statistics of Invariance of Latent Intercept Means ($\pi_0$, Upper Diagonal) and Slope Means ($\pi_1$, Lower Diagonal) Across Groups (n = 18,000) 37

Table 2.1 Sample Size by DIF Group 55

Table 2.2 Reading Passages and Corresponding Items 56

Table 2.3 Descriptive Statistics of the Reading Achievement Test Across Multiple Groups (N = 82,922) 57

Table 2.4 Item Difficulty and Discrimination (N = 82,922) 58

Table 2.5 Goodness of Fit Indices of Two Confirmatory Factor Analysis Models 60

Table 2.6 Results of Pairwise Comparison Analysis (Lord) 64

Table 3.1 Teacher Participant Profiles 82

Table 3.2 Parent Participant Profiles 83
**List of Figures**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1 Sample Path Diagram of An Unconditional Linear Latent Growth Curve Model</td>
<td>24</td>
</tr>
<tr>
<td>Figure 1.2 Final Hypothesized Model as a Linear Function with Two Covariates</td>
<td>29</td>
</tr>
<tr>
<td>Figure 1.3 Home Language Use of Students and Their Families at G3 and G6 (N = 89,609)</td>
<td>30</td>
</tr>
<tr>
<td>Figure 1.4 Most Prominent Patterns of Changes in Home Language Environment of Multilingual Children from G3 to G6 (n = 34,996)</td>
<td>32</td>
</tr>
<tr>
<td>Figure 1.5 A Path Diagram for the Fitted Multi-group LGCM Without Covariates, with Associated Fitted Parameter Estimates for Means ($\pi_0, \pi_1$), Variances ($\zeta_0, \zeta_1$), and Covariance ($\psi$) for Each Group (n = 18,000)</td>
<td>36</td>
</tr>
<tr>
<td>Figure 1.6 Estimated Changes in Comparative Literacy Performance by Group (n = 18,000)</td>
<td>37</td>
</tr>
<tr>
<td>Figure 2.1 Means and Standard Deviations of Expert Ratings of Test Items</td>
<td>62</td>
</tr>
<tr>
<td>Figure 2.2 Results of Simultaneous Multi-group Comparison Analysis (Generalized Lord)</td>
<td>63</td>
</tr>
<tr>
<td>Figure 2.3 Changes in the DIF Magnitude by Item as a Function of Length of Residence</td>
<td>65</td>
</tr>
<tr>
<td>Figure 2.4 Item Characteristics Curves for Three DIF-flagged Items</td>
<td>67</td>
</tr>
<tr>
<td>Figure 3.1 Logic Model of Ontario’s International Languages Elementary Program</td>
<td>81</td>
</tr>
<tr>
<td>Figure 3.2 Grounded Theory Paradigm Model</td>
<td>87</td>
</tr>
</tbody>
</table>
List of Appendices

Appendix A. Contents of the Five Items Selected to Be Tested for DIF (Study 2) 142
Appendix B. Interview Protocols (Study 3) 149
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AERA</td>
<td>American Educational Research Association</td>
</tr>
<tr>
<td>APA</td>
<td>American Psychological Association</td>
</tr>
<tr>
<td>CLD</td>
<td>culturally and linguistically diverse</td>
</tr>
<tr>
<td>DIF</td>
<td>differential item functioning</td>
</tr>
<tr>
<td>EL1</td>
<td>English as a first language</td>
</tr>
<tr>
<td>ELL</td>
<td>English language learner(s)</td>
</tr>
<tr>
<td>ESL</td>
<td>English as a second language</td>
</tr>
<tr>
<td>EQAO</td>
<td>Education Quality and Accountability Office</td>
</tr>
<tr>
<td>G3</td>
<td>grade 3</td>
</tr>
<tr>
<td>G6</td>
<td>grade 6</td>
</tr>
<tr>
<td>G10</td>
<td>grade 10</td>
</tr>
<tr>
<td>ICC</td>
<td>intra-class correlation</td>
</tr>
<tr>
<td>ILE</td>
<td>International Languages Elementary</td>
</tr>
<tr>
<td>IRT</td>
<td>item response theory</td>
</tr>
<tr>
<td>L1</td>
<td>first language</td>
</tr>
<tr>
<td>L2</td>
<td>second language</td>
</tr>
<tr>
<td>LCD</td>
<td>linguistically and culturally diverse</td>
</tr>
<tr>
<td>LGCM</td>
<td>latent growth curve modeling</td>
</tr>
<tr>
<td>MOE</td>
<td>Ministry of Education</td>
</tr>
<tr>
<td>NCME</td>
<td>National Council on Measurement in Education</td>
</tr>
<tr>
<td>OECD</td>
<td>Organisation for Economic Co-operation and Development</td>
</tr>
<tr>
<td>RTI</td>
<td>Response to instruction</td>
</tr>
</tbody>
</table>
Prologue

I received my master’s degree in Korean Language Education from Seoul National University. During my time there, my research interests spanned a wide range of topics within language education, ranging from children’s second language development to language learning curriculum and assessment. My wide interest in language education narrowed to that of school-aged language minority students after an impactful experience teaching adult immigrants in South Korea. Among my students were parents deeply concerned about their children struggling to comprehend lessons due to language barrier, resulting in less likely academic success in school. Within a few years, these children gradually improved their proficiency in Korean to keep up in class, but it often came at the cost of gradually losing their home language, creating communication challenges for their parents. Witnessing these challenges first-hand led me to focus my interest on language minority children, and this was what motivated me to pursue my doctoral studies in Canada. As an immigrant and a mother of language minority children in Canada, this research is personal to me. I am greatly invested in the issues of linguistically and culturally diverse children and their education journey, and it is my personal investment that truly drives my work. I sincerely hope that my research contributes to promoting multilingual and multicultural integration in the public education system and creating more enriching experiences for linguistically and culturally diverse students not only in Canada but also globally.
**Introduction**

Modern societies are becoming increasingly diverse, with multiple cultures and languages co-existing as a result of voluntary migration for better life opportunities and involuntary migration to escape persecution or violence. According to the Organisation for Economic Co-operation and Development [OECD] (2022), one in every five individuals aged 25–64 living in OECD member countries was born in a different country than their current country of residence. This increase in diversity is also reflected in the school-aged student population. Canada, one of the largest immigrant-receiving countries, reported that as of 2016, 38% of children under 15 years old were either first-generation immigrants or second-generation immigrants with at least one foreign-born parent (Statistics Canada, 2017). Moreover, recent statistics demonstrate that not only has the number of immigrants increased, but so has their diversity in terms of country of origin, cultural and linguistic traditions, and socio-economic status (OECD, 2021).

For students from linguistically and culturally diverse families, including immigrants, refugees, and Indigenous children, the development of bi/multilingualism is crucial. Proficiency in the societal language is necessary for academic success, social integration, and future career opportunities. Equally important is the maintenance and development of their heritage language, which connects them to their culture, identity, and family. The development of two or more languages and cultures needs to be recognized as a right for these students (Ruiz, 1984; Skutnabb-Kangas, 2015) as stipulated in the United Nations Convention on the Rights of the Child (Baker & Wright, 2021; United Nations International Children’s Emergency Fund, 1989):

> In those States in which ethnic, religious or linguistic minorities or persons of indigenous origin exist, a child belonging to such a minority or who is indigenous shall not be denied the right, in community with other members of his or her group, to enjoy his or her own culture, to profess and practise his or her own religion, or to use his or her own language. (Article 30, emphasis added)

As such, it is essential to promote and celebrate their linguistic and cultural diversity in society. In school, they should not face discrimination or stereotypes based on their diverse backgrounds.
They deserve equal access to quality education, fair assessments, and support for learning both the socially dominant language and their heritage language.

However, historically, many education systems have catered to students from mainstream culture who speak the socially dominant language as their first language (L1). In some instances, minority languages were even banned in schools, such as Gikuyu in colonial Kenya by the British (Thiong’o, 2005), Korean in colonial Korea by the Japanese (Rhee, 1992), and many Indigenous languages in residential schools in Canada and the United States by European settlers (Child, 2015; Haque & Patrick, 2014). Unfortunately, bilingualism continues to be perceived as a problem in many parts of the world, with students being discouraged and even punished for speaking their own languages in the classroom, as is the case in the United States, one of the most diverse countries (Wright, 2005). With the language-as-a-problem orientation (Ruiz, 1984), schools tend to focus only on identifying students with limited proficiency in the language of instruction as the targets for remedial action, thus linking bilingualism to a deficit framework (Shapiro, 2014).

Linguistically and culturally diverse students require considerable attention and support to succeed in school and in life. Adequate education policies are necessary to promote their linguistic and cultural development and protect them from discrimination in assessments or deficit-oriented discourse. Despite growing interest in this student population in both scholarship and practice, existing research points out that current policies are insufficient and often used merely for political rhetoric (Kim et al., 2020). Even with the right policies in place, implementation may be inadequate due to the generic nature of policies or lack of proper monitoring (Lara & Volante, 2019). In many educational jurisdictions, linguistically and culturally diverse students underperform, or are perceived to underperform in school (OECD, 2015), experience discrimination when assessed (Abedi et al., 1997; Ercikan et al., 2014; Fox & Cheng, 2007), and receive little support for heritage language learning (Carreira & Kagan, 2018).

This dissertation revolves around the experiences of linguistically and culturally diverse students, placing them at the centre. The primary goal was to explore their experiences related to language development and assessment within the public education system through three empirical research studies. Study 1 investigated the longitudinal changes in their home language environment and how these changes are associated with their literacy achievement in school.
Study 2 focused on the extent to which items on a large-scale reading test provide equitable opportunities for linguistically and culturally diverse students to demonstrate their learning. Study 3 evaluated the support provided for the development of their heritage language. Through these studies, I aimed to gain a deeper understanding of how this ever-growing subgroup of students is faring within the school system in terms of academic achievement, equitable access to fair assessments, and support for heritage language development. This dissertation work holds the potential to provide insights that can shape policies and practices, ultimately enhancing the overall educational journey of these students.

The geographic context for the three studies is Ontario, Canada, renowned for its multicultural policies and diverse population. Canada officially adopted multiculturalism as a policy in 1971, and subsequently, Ontario, being one of the most diverse provinces with the largest share of immigrants, has developed numerous policies to support linguistically and culturally diverse students in public schools (Lara & Volante, 2019). However, previous research indicates that despite these seemingly favourable circumstances, students’ linguistic and cultural diversity may not receive sufficient support in Ontario schools. These students are often discouraged from maintaining and developing languages other than Canada’s two official languages, namely English and French (Kim et al., 2020; Kubota & Bale, 2020). Moreover, they are not consistently provided with equal opportunities to demonstrate their learning without being penalized (Buono & Jang, 2021; Cheng et al., 2007; Fox & Cheng, 2007). The findings of these studies conducted within the context of Ontario can offer valuable insights and recommendations for policymakers and educators, not only in Ontario but also in other educational jurisdictions facing with similar demographic challenges.

The remainder of the Introduction chapter comprises the theoretical background on linguistically and culturally diverse students, their language development, and assessment of their language and literacy skills. It concludes with a brief overview of the dissertation.

**Theoretical Background**

**Linguistically and Culturally Diverse Students**

The term “linguistically and culturally diverse (LCD)” began to gain prominence among educators and policymakers in the 1980s. It was originally used in the United States to replace
“limited English proficient (LEP),” a term widely used to refer to students who require additional English language support in the school system (National Association for the Education of Young Children, 1995; Wiese & García, 2001). Due to the negative connotation of LEP that perpetuates deficit thinking about those students who are still developing their English skills, LCD became a preferred term to refer to this particular group of students, along with other popular terms such as language-minority students and English language learners (García & Cuéllar, 2006). The term LCD shifts the focus away from language proficiency and appreciates a range of linguistic and cultural assets that students can bring to the learning environment. It also recognizes the inseparable relationship between language and culture, implying that a student’s cultural background may impact their language learning and academic success. Overall, it reflects a more inclusive and positive approach to understanding and supporting the needs of students from diverse linguistic and cultural backgrounds.

Over time, the definition and use of this term have evolved in response to changing educational contexts and needs, resulting in varying use of the term across different geographical regions and contexts. In many instances, the term LCD or “culturally and linguistically diverse (CLD)” is now being used as a more general term. Geva and Wiener (2015) define it as:

an umbrella term referring to children and adolescents whose families belong to cultural or linguistic groups that are distinct from the culture and/or language of the majority culture. CLD refers to immigrants, refugees, aboriginal people, and marginalized groups who do not speak the language or share cultural values of the dominant group in the society. (p. 4)

This broad definition of LCD does not limit the level of language proficiency of the LCD students; rather, they may have varying levels of proficiency in the dominant language of the education setting. More broadly, the above definition highlights heterogeneity within the LCD student population with respect to culture, ethnicity, language, previous education, and legal or immigration status in the society. Considering the wide range of students on which the research studies in this dissertation focus, Geva and Wiener’s (2015) definition of LCD students above is best positioned to situate this dissertation research.
Many LCD students face common challenges in school settings. The most considerable challenges are related to lower academic achievement and language barriers (Early et al., 2017; OECD, 2015, 2021). The results from the Programme for International Student Assessment administered to 15-year-old students suggest that in most OECD countries, first-generation immigrant students perform worse in reading, mathematics, and science than second-generation immigrant students, who, in turn, underperform their peers without an immigrant background, after controlling for mastery of the language of instruction and socio-economic status (Bilgili et al., 2018; Lara & Volante, 2019; OECD, 2021). LCD students who are still developing their proficiency in the dominant language in the society, particularly for those who migrated at a later age, tend to struggle to understand instruction, participate in class activities, or complete assignments, resulting in larger achievement gap between LCD students with limited language proficiency and non-LCD students (OECD, 2015). Furthermore, some countries that have accepted refugees at high rates are seeing a growing number of LCD students with interrupted schooling (OECD, 2021), which may have limited their opportunity to learn at grade level in any language.

Other obstacles commonly faced by LCD students, to varying degrees, include socio-economic disadvantage and psychological distress (Lara & Volante, 2019). LCD students from low-income families often have limited access to educational resources such as books, technology, or extracurricular activities. While not all LCD students are economically vulnerable, research suggests that differences in socio-economic status often explain a considerable portion of the academic achievement gap, or almost the entire gap in some countries, between students with and without an immigrant background (OECD, 2021). With respect to psychological distress, LCD students are also more likely to experience discrimination and bullying at school based on their linguistic or cultural background (Lara & Volante, 2019). OECD’s teacher survey data also suggests that the majority of teachers in OECD countries feel unprepared to teach in a multicultural or multilingual setting (OECD, 2015, 2021). The various obstacles mentioned above may have contributed to LCD students’ higher rates of absenteeism and lower rates of secondary education completion in many educational jurisdictions including Canada (Majhanovich, 2006; OECD, 2021).
Despite these challenges, LCD students have demonstrated high potential for academic success. In summarizing its key findings regarding immigrant students, OECD (2021) celebrated their high motivation and goal-setting skills. As well, students with an immigrant background were reported to have higher aspiration for completing post-secondary education and being successful in their career once student socio-economic status and academic achievement were accounted for. With appropriate support, the achievement gap between LCD and non-LCD students has been successfully narrowed in many countries (Klinger et al., 2018; OECD, 2021).

To better support LCD students, policymakers and educators need to be aware of common challenges that LCD students face and adopt culturally responsive practices that meet their needs. This may include providing language support, creating a culturally responsive curriculum, fostering inclusive and welcoming school environments, and addressing bias and discrimination specific to LCD students (Cheng & Yan, 2018; Klinger et al., 2018; Lara & Volante, 2019).

**Language Development of LCD Students**

Many LCD students including those with immigrant, refugee, or Indigenous background, are bi/multilingual. They learn at least two languages, either actively or passively, to effectively communicate in a school setting as well as with immediate or extended family members or members of the local community who share the same ethnic language. Some acquire two or more languages simultaneously from birth (simultaneous bilingualism) while others develop the language primarily used by their caregivers first and then start to learn the societal language after their toddler or preschool years (sequential or consecutive bilingualism). Regardless of the L2 acquisition pathway, they manage to meet, with varying degrees, “the communicative requirements of themselves and their society in normal daily life in two or more languages in their interactions with the speakers of any of these languages” (Mohanty, 2019, p. 17) after a few years of exposure to and learning the second language.

However, not many LCD students are balanced bi/multilingual, that is, equally competent in multiple languages. Most LCD students who mainly speak a different language other than the societally dominant one at home, go through a stage called emergent bilingual (García et al., 2008), during which they are in the process of developing proficiency in the societal language while still being more proficient in their L1. Even if LCD students attended school in the societal
language from the primary grades and appear to be fluent in everyday situations, or in basic interpersonal communication skills (BICS), they may still face difficulties with cognitive academic language proficiency (CALP), which involves a range of more complex vocabulary, grammatical structures, and discourse patterns (Cummins, 1979, 2008).

For individuals who immigrate or are exposed to their L2 during adolescence, acquiring the societal language is often an even greater challenge (Chiswick & Miller, 2008; Pasquarella et al., 2012). As students progress beyond the primary grades in elementary school, they face the need to comprehend more complex language, which sometimes assumes culturally relevant background knowledge, and to absorb new information and concepts in subject matters such as science, math, and social studies. As a result, LCD students who arrive in the host country later in life struggle to develop their L2 proficiency and to learn academic content simultaneously. Existing research that shows the time required to approach grade norms in CALP is five to seven years (Cummins, 2008). This finding suggests that many late arrivals may leave secondary school without having fully developed their academic language skill.

Equally concerning is that the development of bilingualism among LCD students often occurs within a subtractive context (Laundry & Allard, 1992). From the initial stage of bilingual language development, LCD students’ language choice and use constantly vary across different purposes, groups of interlocutors, and contexts (e.g., home, school, church, shopping malls) (Hakuta & Garcia, 1989) or even within the same context as they draw upon their plurilingual competence (Piccardo, 2020) or engage in translanguaging practices (Cummins, 2021). Yet, as their proficiency in the societal language increases and they recognize its more prestigious status in the society, they tend to prefer using it over their L1 across various settings. Furthermore, because classroom instruction is typically delivered solely in the societal language, LCD students have limited opportunities to improve their oral or literacy skills in their home language (Baker & Wright, 2021). As a result, their L1 is at risk of being replaced by the L2, and eventually, they may lose their L1 (Flores, 2015; Hakuta & d’Andrea, 1992; Wong-Fillmore, 1991).

Research has consistently shown that the loss of their heritage language (also referred to as native, home, or ancestral language; Baker & Wright, 2021) among LCD students can have detrimental consequences (Cho & Krashen, 1998; Hua & Wei, 2016; Wong-Fillmore, 1991; Wright, 2004). One significant impact is the hindrance of effective communication with family
members and relatives who may not speak the dominant language fluently, which may lead to a communication gap between generations. Relatedly, parents may have difficulty passing on their “values, beliefs, understandings, or wisdom about how to cope with their experiences” (Wong-Fillmore, 1991, p. 343), resulting in a loss of cultural heritage and identity, and negative implications for psychological well-being. At the societal level, language loss within individuals can contribute to language decline and even language death (Wright, 2019).

On the other hand, research on the benefits of bilingualism is plentiful (Cummins, 2021). When LCD students have achieved a certain level of competence in their L2 (Cummins, 1977), they often exhibit stronger executive function skills, such as inhibition, working memory, and attention-shifting (Adesope et al., 2010; Bialystok, 2018; Geva & Ryan, 1993) as well as improved metalinguistic awareness in grammar, morphology, and phonology through positive cross-language transfer (Adesope et al., 2010; Bialystok, 1986, 1987; Chen & Schwartz, 2018; Kuo et al., 2017). Bilingualism has also been linked to improved problem-solving abilities, greater cognitive flexibility, and creativity (Cummins, 1977; Kharkhurin, 2009, 2015). Additionally, bilingualism is known to offer social and cultural benefits. Thanks to the heightened level of sensitivity to communicative functions of language, bilingual individuals are able to communicate with a wider range of people and can access more diverse cultural experiences (Ben-Zeev, 1977; Wermelinger et al., 2017). As a result, they may have a greater understanding and appreciation of different cultures, as well as the ability to navigate intercultural interactions more effectively.

Given the extensive research highlighting the importance of maintaining the heritage language for LCD students, as well as the general benefits of bilingualism, some parents opt to enrol their children in dual language bilingual programs (Collier & Thomas, 2004) instead of mainstream programs where only the societal language is used as the language of instruction. Dual language bilingual programs utilize two languages, typically the socially dominant language and a minority language, for instruction. The aim of these programs is to foster students’ bilingual and biliteral skills towards balanced bilingualism, making them a valuable option for LCD students seeking to preserve their heritage language. Recent data indicate that these programs have experienced significant growth in the United States, with nearly 900 dual language programs listed, the majority of which offer English-Spanish bilingual education.
(Center for Applied Linguistics, n.d.). However, it is important to note that many of these programs operate outside of the public education system and may involve fees (Baker & Wright, 2019).

Heritage language education is often marginalized within mainstream education systems, offering limited opportunities for students to develop their heritage languages. Scholars have pointed out the lack of relevant policies in immigrant-receiving countries (Cummins, 2014; Scarino, 2014; Tucker, 2008) as Tucker (2008) put, “[the education system of the United States] have failed to develop or implement educational policies designed to conserve the heritage language resources of our language-minority students” (p. 43). In many cases, heritage language education within the public sphere takes on a weaker form, where the language is taught as a subject in “school-sponsored informal after-school classes that do not carry any credits” (Baker & Wright, 2019, p. 239). An example of such “add-on” programming is the International Languages Elementary in Ontario, where dual language bilingual programs are prohibited (Kim et al., 2020). The limited support from the public school system has led to the emergence of community-based heritage language programs, resulting in significant variations in access, funding, structure, size, instructor qualifications, and program effectiveness (Lee & Wright, 2014).

As noted, the circumstances that most LCD students face in bilingual development, such as family, school, community, and social/political support, play a vital role in their language development (Chen et al., 2012; Geva & Wiener, 2015). For instance, their bilingual development can be significantly influenced by several contextual factors such as: parents’ proficiency in the student’s L1 and L2, interactions with extended family members (family); access to high-quality bilingual instruction, the level of diversity of linguistic groups in the classroom/school (school); presence of same L1 speakers living close by, availability of heritage language programs (community); physical proximity to the homeland, cultural and economic status of the student’s L1 in the society (demographic/social/political factors) (Baker & Wright, 2021; Geva & Wiener, 2015). Therefore, social constructivist theories that focus on these contextual factors are extremely useful in understanding the language development of LCD students (Chong et al., 2022).
For example, complexity theory, introduced in L2 research by Larsen-Freeman and Cameron (2008) provides a comprehensive framework for understanding language development as a complex and non-linear process influenced by multiple interconnected factors. It emphasizes the dynamic nature of language learning, the emergence of patterns and structures, and the self-organizing capacity of learners (Larsen-Freeman, 2019; Larsen-Freeman & Todeva, 2021). Similarly, ecological systems theory (Bronfenbrenner, 1979; van Lier, 1997) recognizes the influence of nested systems on individual’s learning including language development. It considers the reciprocal relationships between the individual, their immediate environment (such as family and school), and the broader social and cultural contexts. Lastly, sociocultural theory, developed by Vygotsky (1978) and further expanded by Lantolf (2000), highlights the significance of social interaction and cultural context in language development. It emphasizes that learning occurs through collaborative interactions with more knowledgeable individuals within a socio-cultural setting. Sociocultural theory underscores the role of language as a tool for communication and meaning-making, emphasizing the significance of social and cultural factors in the language development. Despite their unique perspectives and areas of focus, these theories share a common emphasis on dynamic and complex contexts, social interaction, and individual differences in shaping language learning processes and development, highlighting the importance of considering multiple factors when examining the bilingual development of LCD students.

Assessing LCD Students’ Language and Literacy Skills

Given the complexities inherent in the language development of LCD students as discussed earlier, assessing their language and literacy development in the school system poses a significant challenge. Ideally, the assessment process should consider the student’s language abilities in all the languages they are learning, including their L1 and the societal language. In some cases, schools may use bi/multilingual assessments that measure a student’s proficiency in each and all of their languages. For example, the Woodcock-Muñoz Language Survey III (WMLS-III; Woodcock et al., 2017) is a norm-referenced assessment that measures students’ academic language proficiency in the areas of listening, speaking, reading, and writing in both Spanish and English. As assessments administered in different languages allow for skill development comparison across languages (Duffy et al., 2018), they can offer a more
comprehensive picture of the student’s unique language profile and overall language development and help identify specific areas where additional support student may be required to become a more proficient bilingual.

Moreover, it is worth noting recent research and discussions regarding multilingual tests that allow students to use multiple languages or linguistic resources during the testing processes (Shohamy, 2011). Shohamy et al. (2022) aptly contend that multilingual assessments offer advantages not only to first-generation students who have recently started learning the dominant language relatively but also domestic LCD students whose L2 proficiency may exceed their proficiency in their heritage language. Incorporating multiple languages in assessments communicate the education system’s acknowledgment of diverse languages present within its student population and help LCD students perceive their own languages more positively. Additionally, a multilingual testing policy aligns with the overarching education policy in many educational jurisdictions that promotes multilingualism in classrooms, addressing the disconnect between learning and assessment (Shohamy et al., 2022).

In the school system, however, standardized testing administered solely in the language of instruction is the most common approach to identify, place, or monitor progress of LCD students. When an LCD student is still in the early stage of developing the societal language and is designated as a language learner, language proficiency measures designed for school-aged language learners (e.g., WIDA ACCESS) are often used by teachers in language support programs. School psychologists and other practitioners frequently administer academic achievement tests such as the Wechsler Individual Achievement Test III (WIAT-III; Wechsler, 2009) and the Woodcock-Johnson IV Tests of Achievement (WJ-IV ACH; Schrank et al., 2014) to assess reading, writing, and mathematics skills of LCD students. As LCD students become more proficient in the language of instruction, they are required to write mandated assessments that are designed to serve each educational jurisdiction’s accountability or graduation purposes all in the language of instruction (e.g., Education Quality and Accountability Office Assessments in Ontario, Canada; Foundation Skills Assessment in British Columbia, Canada; Smarter Balanced Assessment System in California, Connecticut, Michigan, and others states in the United States; Key Stage Assessments in United Kingdom).
Yet, it is crucial to exercise sensitivity towards linguistic and cultural differences when interpreting results obtained from standardized tests, particularly for students with limited exposure to the language being assessed. These test scores may not accurately reflect a student’s overall language abilities or development, as they are often designed and normed on monolingual students (Cheng et al., 2007). Such tests may be based on the assumption that all students have had equal exposure to the language of the test, and they may not take into account differences in language backgrounds, dialects, or accents. Test stimuli or items that include complex sentence structures, high-level vocabulary, or idiomatic expressions may also be challenging for LCD students to comprehend. Additionally, if the test is timed, students who are not accustomed to working under time constraints may experience anxiety or stress, leading to inaccurate test results. While accommodations such as dictionary use and additional time can be provided, they may not fully level the playing field (Fox & Cheng, 2007).

Moreover, standardized language or literacy tests may not reflect the cultural knowledge and experiences of LCD students, as the item contents are often based on dominant culture norms. The content and format of these tests often reflect dominant culture norms, including unfamiliar content for students from diverse cultural backgrounds and culturally biased items that are only relevant to the experiences of monolingual students. For example, ice hockey, one of the most popular sports in Canada, may be a foreign topic for students from a different culture. Such culturally sensitive topics can put LCD students at a disadvantage, as research consistently shows that their reading performance improves when the reading stimuli are culturally familiar to them (August & Shanahan, 2006).

For the reasons discussed, the interpretation and use of scores from standardized testing without caution when assessing the language and literacy skills of LCD students can have detrimental consequences. Blindly relying on these scores may result in misidentifying their learning difficulties and subsequently placing them in instructional programs that are not suitable for their needs (Geva & Wiener, 2015). For instance, over-identification can happen when the IQ-achievement discrepancy model (Fuchs & Fuchs, 2006) is applied to identify LCD students with learning disabilities without considering the influence of their language and cultural differences. English language learners who struggle academically due to linguistic or cultural barriers may be incorrectly identified as having learning disabilities based on their lower
academic achievement. On the other hand, more recently, some education systems have adopted a “wait and see” approach (Geva & Xi, 2016; Kangas, 2019), delaying testing for LCD students until they have acquired a certain level of language proficiency. This policy recognizes the importance of language proficiency in academic achievement and suggests that seemingly lower academic achievement may stem from language barriers rather than inherent disabilities. While delay policies aim to prevent misidentification, they raise concerns about the under-identification of genuine learning disabilities and potential delays in providing appropriate interventions and support for students in need.

This issue of over- or under-identification of LCD students with learning disabilities, along with the resulting over- or under-representation in special education programs, has been a contentious matter for decades. Extensive research highlights the harm caused to misclassified students (Counts et al., 2018; Limbos & Geva, 2001), emphasizing the importance of a cautious and comprehensive assessment process. Once labeled as having special education needs, even those who are misclassified are more likely to remain in such programs for an extended period with lower curriculum expectations and limited access to regular classroom environments (Harry & Klingner, 2007; Skiba et al., 2016). Under-identification can also impede academic progress of those who were unidentified as they may not receive the necessary resources and support required to succeed (Sullivan, 2011; Duran, 2008). Thus, it is imperative to accurately identify learning disabilities and provide appropriate programming and placement to ensure equitable access to education for LCD students.

To obtain a comprehensive and accurate evaluation of the language and literacy skills of LCD students, it is crucial that teachers and other practitioners do not solely rely on traditional tests but also utilize a variety of classroom-based performance assessments (e.g., portfolio, interviews, oral presentations, group discussions, role plays, informal observations) with cultural and linguistic sensitivity (Baker & Wright, 2021; Geva & Wiener, 2015; Jang, 2014). For instance, Jang (2014) suggests that portfolio assessment reflects the shifting culture of assessment as it collects students’ work samples (e.g., journals, presentations) and documents their learning process. Teachers are also encouraged to conduct formal and informal observations while students are interacting with teachers or peers in a group setting (Jang, 2014). These alternative assessment methods can provide a more comprehensive and authentic profile of LCD
students regarding their language development by allowing them to demonstrate their language competence in real-life situations.

One way to enhance teacher-led classroom assessments is by incorporating descriptor-based language proficiency scales, which provide detailed descriptions of language proficiency levels across various language domains such as speaking, listening, reading, and writing (Jang, 2014). For example, the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001, 2020; Piccardo et al., 2011) is commonly used for a wide range of language learners in Europe and elsewhere. In Ontario, Canada, the context of this dissertation, the Steps to English Proficiency (STEP; Ontario Ministry of Education, 2015) plays a critical role in assessing, tracking, and supporting English learners in K-12 schools as it was designed specifically for Canadian students and aligned with the Ontario curriculum (Jang et al., 2015). These scales enable teachers to identify areas of strength and weakness for each student and adjust their instruction accordingly (Ontario Ministry of Education, 2015).

Another type of assessment that is particularly useful in assessing LCD students is dynamic assessment (Lantolf & Poehner, 2011; Swanson & Lussier, 2001). Dynamic assessment is an interactive and individualized assessment approach that aims to evaluate an individual’s learning potential and the level of support they need to achieve their full potential (Jang, 2014). Based on Vygotsky’s (1987) zone of proximal development theory, the assessors—mostly teachers—actively engage with the individuals being assessed to provide feedback, guidance, and scaffolding as needed. Similar to portfolio assessment, dynamic assessment emphasizes the learning process rather than just the final product, and teachers can gain insights into the learning potential of their students, identify their strengths and weaknesses, and provide targeted support and intervention to help improve their language and literacy skills.

Relatedly, the response to intervention (RTI) approach, combined with multi-tiered instruction, has garnered significant interest in the assessment of students, including LCD students who may have potential learning disabilities (Xu & Drame, 2008). RTI is considered a form of dynamic assessment (Fuchs & Fuchs, 2006; Geva & Wiener, 2015) as it focuses on monitoring students’ progress in response to instruction and provides tiered interventions and supports based on their individual needs. In the RTI model, interventions become progressively more intensive as students move from Tier 1 to higher tiers, such as Tier 3, through approaches
such as explicit instruction, increased frequency, longer duration, or smaller group settings (Fuchs & Fuchs, 2006). Both dynamic assessment and the RTI approach can mitigate the risk of bias in assessment and offer a more equitable evaluation of students’ language skills by considering not only their language abilities but also their cultural and linguistic background (Klingner et al., 2005).

As noted in the description of various assessment approaches, assessing the language development of LCD students is a complex task that requires trained and qualified teachers who understand the intricacies of bi/multilingualism and the linguistic and cultural diversity of LCD students. Teachers must recognize that assessment should not be a one-time event but rather an ongoing process that involves regular monitoring of students’ language development. However, many schools and programs lack the necessary resources and expertise to provide accurate and reliable assessments for LCD students (Hill, 2017). This is particularly challenging in heritage language education, where teachers struggle to find appropriate assessment materials or adequate training opportunities on assessment literacy (Morgan et al., 2021).

**Overview of the Dissertation**

This dissertation is comprised of three studies that address three high-level research questions regarding LCD students in the K–12 public education system:

1. How do LCD students’ home language environments change over time and how are these changes associated with their literacy achievement?

2. To what extent do large-scale literacy assessments provide equitable opportunities for LCD students to demonstrate their learning?

3. How are LCD students assessed and supported for their heritage language development in school?

Study 1 addressed the first research question regarding academic performance LCD students. The study aimed to understand changes in elementary school students’ home language environments over time, and how longitudinal patterns of English literacy achievement differ among students with different home language shift patterns. The study analyzed longitudinal cohort data from nearly 90,000 students in grades 3, 6, and 10 from the provincial assessments in
Ontario to identify changes in their home language environment. A subsample of 18,000 students was used to examine different patterns of relative literacy performance over time and their associations with immigration background and early intervention programming using multi-group latent growth curve modeling.

Study 2 addressed the second research question concerning equitable access to fair assessments and the potential disadvantage experienced by LCD students in large-scale assessments. The study investigated the extent to which items in a large-scale reading test function differentially across multiple groups of students with varying levels of familiarity with mainstream Canadian culture, using length of residence in Canada as a proxy variable. This study analyzed provincial literacy assessment data from approximately 3,400 grade 3 students, using differential item functioning analysis, a statistical approach instrumental to investigating item bias.

Study 3 addressed the third research question about the support provided for LCD students’ heritage language development. Focusing on assessment practices, the study was designed as a case study of Korean language classes to evaluate a publicly funded heritage language program in Ontario. Through individual interviews with five teachers and nine parents of heritage learners in Korean language classes, the study investigated how recent provincial assessment policy changes were integrated into the program.

In the last chapter, the summary of overall implications and contributions of the three studies are presented, along with future research directions.
Study 1. Change in Home Language Environment and English Literacy Achievement Over Time: A Multi-group Latent Growth Curve Modeling Investigation

The educational and literacy outcomes of students from diverse home language environments have been an important concern for educators and policymakers. In most studies investigating this issue, variables that identify this population have been considered as static. In other words, such labels as English language learners, English-as-a-second-language students, or language minority students, tend to serve as a time-invariant variable, reflecting the view of students’ status as fixed. In reality, however, students’ language use and their home language environment change over time, as a result of ecological interactions among students, their families, and communities. Although it is common for students from immigrant families to change their primary language from that spoken at home (L1) to the socially dominant language (L2) (Wong-Fillmore, 1991), the extent to which they maintain their L1 in their home environments widely varies. For example, some students continue to hear and speak the L1 at home, while others gradually increase the use of the L2 with family members and can even lose aspects of their L1 competence.

In one study (Hammer et al., 2009) on children’s language and literacy acquisition, the authors acknowledged that students’ home language environment can change over time—for example, the use of the L1 in their homes can increase or decrease—and that this can influence students’ L2 literacy development. However, large-scale research is scant regarding how students and their family members in multilingual homes change their language use over time. Under-researched as well is the association of such changes with students’ literacy development over time.

In the present study, we build on Hammer et al.’s (2009) findings and contribute to the current literature by recognizing home language environments as a dynamic, rather than static, variable. While Hammer et al. (2009) used a small-scale dataset covering two years, we used population-level literacy performance data from Ontario, Canada over a seven-year time span to understand how students’ home language environments change, and how their relative achievement in English literacy differ among multiple subgroups of dynamic home language environments.
Literature Review

Home language environment and language shift

For many children from immigrant families, reaching the same level of language proficiency in both their L1 and L2 is often challenging. Children who were once proficient in their L1 have been reported to go through a process called language shift once they begin to receive intensive input in the L2 and become linguistically assimilated into the larger society (Flores, 2015; Wong-Fillmore, 1991). Traditionally, language shift was thought to be a process that takes generations; yet, researchers have begun to address how language shift can, in fact, take place during a single generation. Stevens and Ishizawa (2007) argued that, based on their findings that the language repertoires of siblings in the same households may differ, the inter-generation perspective of L1 shift is too simplistic. In the same vein, Pease-Alvarez (2002) acknowledged an intra-generation language shift and criticized the traditional view for characterizing bilingualism as “a temporary inter-generational bridge between monolingualisms” without taking “the complex and dynamic network of sociocultural processes” into account (p. 115).

Research on language shift in the home environment concerns families’ daily interactions, their language ideologies and beliefs, and their goals for shaping language learning and use (King & Fogle, 2017). One of the earliest studies was by Scheff (1965), who examined changes in immigrant families’ home and community language use. Scheff (1965) found home language use was more resistant to change than community language use, but this varied by age, education, length of residence, and income. Although many researchers have focused on parents’ influence on the prevalence of languages spoken at home, children are agentive in influencing their parents’ language behavior, actively shaping their own home language environment. The process may be intuitively viewed as a negotiation between parents and their children (Fogle & King, 2013). Children have been shown to be more likely to use, prefer, and maintain their L1 if they view their parents positively and also see their family as cohesive and egalitarian (Tannenbaum & Howie, 2002). The younger the age of language-minority children when they enroll in majority-language schools, the faster the attrition of the home language, which can hold negative implications for within-family communication (Wong-Fillmore, 1991).
Although the phenomenon of language shift among multilingual children and their families itself has been extensively researched, the dynamic nature of the home language environment has not drawn much attention among educators and policymakers as a factor to consider in relation to academic achievement. In the context of schools, Slavkov (2018) examined registration forms in three Canadian provinces and found that only a small portion conceptualized home language environments as dynamic by asking multiple language-profiling questions (e.g., first language, primary language spoken at home, other languages); the other forms operationalized home language environment as static (e.g., home language). Furthermore, in the current literature, research that examines the association between changes in the home language environment and educational outcomes over time is scarce.

Related to this intra-generation view of language shift is the idea of *complex systems*, or *dynamic systems*, which is gaining traction among many researchers in developmental psychology and education (Fusella, 2013; Larsen-Freeman & Cameron, 2008). These theories focus on changes, dynamism, and interdependencies within a complicated system (e.g., process, person, society). Larsen-Freeman and Cameron (2008) argued that focusing on both individual growth and variability as well as stability can help researchers and educators gain a better understanding of the developing system. They further encourage researchers to use appropriate analysis techniques that can capture variability at different levels and timescales such as multivariate time-series modeling and growth curve analysis. As such, examining language shift and literacy development from a complex systems perspective can provide valuable insights into multilingual students’ learning processes and help educators construe ways to better support the students.

**Variation in literacy achievement by home language**

The language shift within individual students and their dynamic home language environments deserve attention in research on literacy development and academic achievement in general. This line of inquiry has the potential not only to further demonstrate the benefits of bi/multilingualism but, more importantly, to inform researchers about the effects of changes in the home language environment on literacy development from a longitudinal perspective.
It is well established that language and literacy skills transfer across languages and that bilingualism and biliteracy have substantial cognitive benefits (Bialystok, 2011; Chen & Schwartz, 2018; Cummins, 1993). At a minimum, L1 maintenance shows no correlation with L2 literacy development (Nguyen et al., 2001), but a more common finding is that a rich home language environment and parental use of the L1 are associated with high academic achievement, as Sneddon (2000) found in a study of Gujarati-speaking families in London, England, and Dolson (1985) found for Spanish speakers in Los Angeles, the United States.

Similarly, in the context of Ontario, Jang et al. (2013) found that elementary school students who use multiple languages at home transcend early reading achievement gaps and have the most competent reading comprehension skill profile of all home language subgroups (including domestic students) after living in Canada for five years. Findings by Sinclair et al. (2019) also suggested students who use multiple languages at home in grade 6 are likely to have strong literacy profiles and retain those strengths in high school. However, in the latter study, students who exclusively use languages other than English at home (a group that showed high achievement in elementary school) had a tendency toward literacy skill attrition in high school. This implies that such students may benefit from intervention in middle school and early high school years. Indeed, Fox and Cheng (2007) and Han and Cheng (2011) found that high school students who immigrated to Ontario and who were learning English as an L2 experienced a complex set of challenges while attempting to meet the literacy standards required for graduation.

Many factors interact to impact immigrant students’ literacy and academic achievement, including their multilingual proficiencies, national origin, educational ambition, school context, their families’ social, human, and economic capital, and the level, length, and nature of acculturation (Gunderson, 2007; Portes & Rumbaut, 2001). Relatedly, Jia et al. (2014) investigated the relationship between English reading comprehension, length of residence, acculturation to mainstream Canadian culture, and heritage enculturation (maintenance of one’s own culture) within a sample of Ontario-based Chinese immigrant adolescents. Length of residence and acculturation significantly predicted increases in reading comprehension, but heritage enculturation did not demonstrate a significant effect in the model. This suggested that
acculturation positively impacts reading comprehension in L2, but retaining one’s culture has neither a positive nor negative impact.

Both cross-sectional and longitudinal studies have examined the association between home language environment and literacy development. Some cross-sectional studies showed no statistically significant correlation between parental use of English in a multilingual home environment and English vocabulary in grade 2 (Gutiérrez-Clellen & Kreiter, 2003) or English grammatical ability in grade 5 (Duursma et al., 2007). With regard to longitudinal studies, Mancilla-Martinez and Lesaux (2011) investigated the relationship between the home language environment of Spanish-speaking language-minority students in preschool, and their vocabulary growth in Spanish and English up to the age of 12 years. Intuitively, preschool students from homes where English was used significantly more than Spanish demonstrated the highest level of English vocabulary, followed by students who used equal amounts of both languages and those who used mostly Spanish at home. However, these latter two groups’ rates of growth were significantly higher, and their rates of deceleration significantly lower, than that of the group who spoke mostly English, significantly reducing the vocabulary gap by the age of 12 years.

The longitudinal study by Hammer et al. (2009), mentioned briefly above, uniquely conceptualized home language as a dynamic, time-variant factor. These authors examined how changes in bilingual (Spanish-English) maternal language usage influenced young children’s bilingual vocabulary development. Mothers’ increased English usage did not significantly impact the children’s English vocabulary growth or emergent English literacy abilities, but it was associated with a decreased rate of Spanish vocabulary growth. Although this study offered insightful findings, it was limited to a relatively small sample size \( n = 72 \), covered a short period of time (2.5 years), and had a narrow focus on language usage by mothers. With the current study, we aimed to overcome such limitations.

**Latent growth curve modeling and its applications**

As already noted, the focus on changes and dynamism within and between individuals necessitates the use of innovative longitudinal analytic methods. Latent growth curve modeling (LGCM) is a flexible modeling approach appropriate for tracking changes over time. Unlike analysis of variance (ANOVA), which treats within-group differences as error variance, LGCM
evaluates both inter-individual and intra-individual variability (i.e., differences among individuals and changes within individuals over time) (Preacher, 2010). Specifically, LGCM can be used to examine linear and curvilinear trends of change for individuals as well as the overall sample’s change (Duncan et al., 2013). Under the structural equation modeling framework, LGCM can account for measurement error, evaluate the influence of time-invariant and time-varying covariates, and determine the variation in longitudinal patterns among multiple subpopulations (Duncan & Duncan, 2009).

Figure 1.1 displays a sample path diagram of an unconditional linear LGCM with three time points. Two latent factors in circles (intercept and slope) are estimated through three manifest variables in rectangles at Time 0, 1, and 2 (with the same time intervals). The model depicted in Figure 1 can be described in the following equations:

\[
\begin{align*}
    y_{0i} &= \pi_{0i} + \pi_{1i} \times 0 + \varepsilon_{0i} \\
    y_{1i} &= \pi_{0i} + \pi_{1i} \times 1 + \varepsilon_{1i} \\
    y_{2i} &= \pi_{0i} + \pi_{1i} \times 2 + \varepsilon_{2i}
\end{align*}
\]

where \(y_{0i}, y_{1i}, \) and \(y_{2i}\) represent observed outcomes measured at three different time points for individual \(i; \varepsilon_{0i}, \varepsilon_{1i}, \) and \(\varepsilon_{2i}\) are composite error terms that are composed of both random measurement error and time-specific influence of individual \(i; \pi_{0i}\) is the fixed intercept factor representing the initial level of the outcome; \(\pi_{1i}\) is the fixed slope factor representing the rate of change in the outcome over time. The variances of the random effect intercept (\(\zeta_{0i}\)) and random effect slope (\(\zeta_{1i}\)) in Figure 1.1 indicate inter-individual differences in initial status and rate of growth or change, respectively. A statistically significant slope factor mean would indicate a presence of change over time on average while its statistically significant variance would indicate significant variability in their growth rate among individuals. The covariance between two latent factors (\(\psi_i\)) demonstrates the association between initial status and rate of change. For multi-group analysis, the same latent growth model is simultaneously fitted for different subpopulations, and the initial status and rate of change for each group is estimated separately. Of substantive interest to the present study is whether the parameters related to the latent factors (i.e., \(\pi_0, \pi_1, \zeta_0, \zeta_1, \psi\)) significantly differ across groups.
LGCM has been implemented to evaluate the development of literacy skills, for instance, the growth in morphological awareness and receptive vocabulary (Kieffer & Lesaux, 2012), word reading and productive vocabulary (Mancilla-Martinez & Lesaux, 2010), oral reading fluency (Jimerson et al., 2013; Yeo & Park, 2014), and reading comprehension (Guglielmi, 2008; Lervåg & Aukrust, 2010). The extent to which mediating factors predict initial status and rate of growth can be evaluated by the inclusion of covariates. For instance, Kieffer and Lesaux (2012) found that phonological awareness did not statistically significantly predict the initial status or rate of growth in morphological awareness and vocabulary among Spanish-speaking language-minority students in grades 4 to 7.

LGCM allows the evaluation of how the initial status, growth trajectories, and influence of covariates vary across groups through multi-group LGCM or dummy-coded covariates. For example, previous studies utilizing LGCM have found that language-minority students often demonstrate lower initial oral reading fluency (Jimerson et al., 2013; Yeo & Park, 2014) and reading comprehension (Lervåg & Aukrust, 2010) than their language-majority peers, although differences in growth trajectories remain inconclusive. The inclusion of covariates in multi-group LGCM allows for the investigation of variables that predict initial status and rate of growth.
among groups differently. For instance, self-reported L1 proficiency has significantly predicted initial English reading achievement among Hispanic, but not Asian students in grades 8 to 12 (Guglielmi, 2008). Similarly, vocabulary knowledge has been shown to have a significantly stronger relationship with reading achievement growth among language-minority than language-majority students (Lervåg & Aukrust, 2010).

Current LGCM research on literacy exemplifies the capacity for this method to answer longitudinal research questions and to determine differences in growth trajectories or change patterns among multiple subgroups (Guglielmi, 2008; Halle et al., 2012). Few studies to date have specifically investigated longitudinal literacy achievement while incorporating dynamically measured home language environment as a grouping variable.

The Present Study

The purpose of this seven-year longitudinal cohort study was three-fold: (1) to explore the patterns of changes in students’ home language environment over time; (2) to model longitudinal literacy achievement patterns among multiple subgroups with different home language shifts; and (3) to determine the influence of immigration status and English-as-a-second-language (ESL) program support on the relative literacy achievement pattern of each group. The last set of variables—immigration status (Canadian-born vs. first-generation immigrant) and ESL program support (whether or not the student was receiving ESL instruction at G3)—was included as they have been reported to predict literacy outcomes in the previous studies focusing on multilingual students in Ontario (Jang et al., 2013; Sinclair et al., 2019). The present study, inspired by a complex systems perspective (Larsen-Freeman & Cameron, 2008) and Hammer et al.’s (2009) conceptualization of home language as a dynamic variable, answers the following research questions:

1. How do students and their families change their home language use between grades 3 and 6?
2. How are changes in home language environment associated with students’ relative literacy achievement patterns between grades 3 and 10?
3. Do immigration status and ESL program enrollment in grade 3 predict the relative literacy achievement patterns of students with different home language environments?
Method

Participants

We used longitudinal cohort data of students who participated in Ontario’s provincial assessment program administered by the Education Quality and Accountability Office (EQAO). The literacy achievement of students in Ontario is measured by this assessment in the spring of grades 3, 6, and 10 (G3, G6, and G10, respectively). The data utilized in this study are from students who took the G3 assessment in 2010, G6 in 2013, and G10 in 2017. The original datasets included approximately 130,000 students in G3 and G6, and 210,000 students in G10. Students were included in the present study if they satisfied these three criteria: (1) they wrote the assessments in all three years; (2) they took the English version throughout (as opposed to in French); and (3) their home language information was available in both G3 and G6. This resulted in 89,609 students (49.6% female) in the final dataset.

Among these 89,609 students, 8.6% were born outside Canada, among which 60.0% had been in Canada for at least five years, and an additional 15.4% had been in Canada for at least three years at the time of the G3 assessment administration. By the G6 administration, 93.0% of non-Canadian-born students had been in Canada for five years or more. With regard to early intervention, 9.5% of all students included in the analysis were enrolled in an ESL program in G3. Among these students, 66.2% were Canadian-born, comprising the early-identified “domestic language learners” (Jang et al., 2013, p. 425). Owing to its longitudinal nature, the data did not include students who arrived in Canada or entered the Ontario school system after G3.

---

1 As Canada is an English–French bilingual country, students in French-language schools complete the French version of the provincial test. These students (3.5–5.0% in each grade) were excluded from the analysis in order to limit the focus of the study to the English language.
Measures

During the G3 and G6 EQAO literacy assessment administrations, students were asked to complete a questionnaire that included two questions about home language environment: “In which language(s) do people speak to you at home?” and “Which language(s) do you speak at home?” with options of “only English,” “mostly English,” “another language(s) as often as English,” “mostly another language(s),” or “only another language(s).” To be consistent with the classification and notations used by Jang et al. (2013) and Sinclair et al. (2019), students’ home language status was coded in this paper as language(s) heard at home followed by language(s) spoken at home; for example, EnEn (English–English) means hearing mostly or only English and speaking mostly or only English, and OtEq (Other–Equal) means hearing mostly or only another language(s) and speaking both English and the other language(s) equally. Each student was assigned to one of the nine groups (i.e., EnEn, EnEq, EnOt, EqEn, EqEq, EqOt, OtEn, OtEq, OtOt) in G3 and G6, separately.

The EQAO literacy assessment administered between 2010 and 2017 elicited the same literacy skills at G3, G6, and G10: in reading, explicit comprehension, implicit comprehension, and making connections; in writing, topic development and writing conventions. The G3 and G6 reading measures consisted of 26 multiple-choice and 10 open-response questions using four reading passages (Cronbach’s $\alpha = .84, .88$). The G3 and G6 writing measures ($\alpha = .79, .81$) included eight multiple-choice and six open-response questions. The G10 literacy measure, also known as the Ontario Secondary School Language Test, integrated reading and writing measures into five sections. The reading measure incorporated 31 multiple-choice and five open-response questions ($\alpha = .81$) while the writing measure included eight multiple-choice and eight open-response questions ($\alpha = .77$). The EQAO reports a composite score as overall literacy skill on a scale of 0.1 to 4.9 in G3 and G6, and 200 to 400 in G10. Due to the different metrics used between G3–6 and G10, the composite scores in each grade were transformed to z-scores using all students who completed the English version of the assessment. With a mean of 0 and a standard deviation of 1, a student’s z-score expresses their relative standing of overall literacy skills among the test taker population.
Analysis

In order to examine changes in home language environment over time (RQ1), each student’s home language status at G3 and G6 was determined separately. Changes in students’ home language environment between two grades were analyzed descriptively and visually.

To investigate relative literacy performance among multiple groups of home language environment (RQ2) and the predictive power of two covariates (i.e., immigration status and ESL enrollment in G3) (RQ3), a multigroup LGCM approach was deployed. The group membership for each student was determined using the home language status at both G3 and G6. At each time point, students were identified as to whether they were living in an English-dominant (EnEn) or multilingual home (all other eight categories other than EnEn). Then, a dynamic variable that focuses on the type of changes in home language environment between G3 and G6 was created, resulting in four groups: (a) consistently English-dominant (Eng to Eng), (b) English-dominant to multilingual (Eng to Multi), (c) multilingual to English-dominant (Multi to Eng), and (d) consistently multilingual (Multi to Multi).

LGCM analyses were conducted using Mplus Version 8.2 (Muthén & Muthén, 1998–2017). All analyses used maximum likelihood estimation with robust standard errors and a Satorra-Bentler scaled test statistic, or MLM (Satorra & Bentler, 2001), which is robust to the violation of the normality assumption (Muthén & Muthén, 1998–2017). For model building, we started by testing three baseline models: intercept-only, linear, and curvilinear models (Geiser, 2012). Models were evaluated based on multiple fit indices: chi-square goodness-of-fit index, the comparative fit index (CFI, > .95), Tucker-Lewis index (TLI, > .95), root mean square error of approximation (RMSEA, < .08), and standardized root mean square residual (SRMR, < .10) (Cangur & Ercan, 2015). After identifying the best-fitting functional form, the parameters for the latent factors were allowed to differ by group (Curran et al., 2010). We then tested whether the means of the latent intercept factor (initial status in G3) and latent slope factor (rate of change) were invariant across four groups. All 12 possible pairs of the parameter estimates (six pairs for intercept, six pairs for slope) were tested using a series of Wald tests (Wang & Wang, 2012). An a priori significance level of $\alpha = .05$ was applied. Finally, two covariates (immigration status and ESL enrollment in G3) were incorporated into the model to predict the latent growth factors. The path diagram of the final hypothesized model in a linear form is illustrated in Figure 1.2. The
factor loadings (or, time scores) of the slope factor, which represent values of the time metric, were fixed at 0, 0.43, and 1 (instead of 0, 1, and 2.33) to relate the interpretation of the slope parameters to the entire seven years rather than every three years.

**Figure 1.2**

*Final Hypothesized Model as a Linear Function with Two Covariates*

---

**Results**

**RQ1: How do students and their families change their home language use between grades 3 and 6?**

We first report students’ home language use at G3 and G6, cross-sectionally, and then track changes in home language environment for individual students during this time period. Figure 1.3 displays the compositions of home language environment of 89,609 students in G3 and G6. For each grade, nine boxes represent different home language environments. The boxes...
in the top row are homes where students hear only or mostly English, whereas the boxes in the
bottom row are homes where they hear only or mostly another language(s). Similarly, the boxes
on the far-right side are homes where students speak only or mostly English, whereas the boxes
on the far-left side are homes where they speak only or mostly another language(s). Boxes
located higher and to the right represent students and their families who use more English at
home.

Figure 1.3
Home Language Use of Students and Their Families at G3 and G6 (N = 89,609)

In G3, the majority of students (66.1%) heard and spoke only or mostly English (EnEn).
The second largest group was EqEq (8.7%), where students heard and spoke English and another
language(s) equally, followed by OtOt (8.2%). For these three groups (74.0%), the language(s)
they spoke corresponded with the language(s) they heard. These three groups were followed by
EqEn (4.8%) and OtEq (4.3%), who spoke more English than they heard. Fewer students were
identified as EnEq (3.1%) or OtEn (3.0%), whereas only 1.0% and 0.9% of students were in
EqOt and EnOt, respectively.

Moving on to G6, the order of the home language environment groups in terms of the
group size remained largely the same. Compared to G3, more students were living in English-
dominant homes (EnEn, 70.7%), followed by EqEq (6.8%) with a lower proportion. The
proportion of students in EqEn (6.3%) increased by over 30%, becoming the third largest group; on the contrary, the proportion of OtOt (5.6%) decreased by 32%.

Subsequently, we tracked changes in home language environment for individual students by comparing their self-reported home language use between G3 and G6. Among 81 possible combinations of home language status at two time points (nine groups at G3 times nine groups at G6), the majority of students (54,613, 61.0%) reported their status as EnEn in both G3 and G6, suggesting that their home language environment remained as English-dominant throughout the time period.

In order to focus our investigation on multilingual students, we excluded these constant English-monolingual students and re-examined the data comprised of 34,996 students whose home was multilingual at one or both time points. Figure 1.4 illustrates the ten most prominent patterns of changes in home language environment, with the thickness of the arrows representing the relative size of groups with the corresponding pattern. Although the home language environment of some multilingual students remained unchanged (7.3% for OtOt, 4.9% for EqEq), many students tended to hear and speak more English than another language(s) at home in G6 compared to G3. For example, 6.6% of the students reported hearing and speaking another language(s) as often as English in G3, but changed to hearing and speaking mostly English in G6 (EqEq to EnEn, 6.6%). Similarly, 4.0% of the students used to hear and speak mostly another language(s) in G3, but ended up hearing and speaking English as often as another language(s) in G6 (OtOt to EqEq, 4.0%). Many students who reported hearing or speaking another language(s) as much as English in G3 found themselves in a mostly English-speaking home environment in G6 (EqEn to EnEn, 6.1%; EnEq to EnEn, 5.1%). Although some students reported their home language environment changed from most English to multilingual (EnEn to EqEn, 5.4%; EnEn to EqEq, 2.7%), students and their families, in general, tended to increase their use of English at home.
Figure 1.4

Most Prominent Patterns of Changes in Home Language Environment of Multilingual Children from G3 to G6 (n = 34,996)

RQ2: How are changes in home language environment associated with students’ relative literacy achievement patterns between grades 3 and 10?

As mentioned earlier, students were classified into four groups based on the patterns of changes in home language environment between G3 and G6: (a) Eng to Eng, (b) Eng to Multi, (c) Multi to Eng, and (d) Multi to Multi. The left side of Table 1 provides the means and standard deviations (SDs) of literacy assessment scores (in standardized z-scores) with the sample size by group. Owing to the substantial differences in the number of students across the four groups, 4,500 students from each group were randomly selected to comprise a subsample that was used in subsequent LGCM analyses (n = 18,000). The minimal discrepancies of the descriptive statistics between the population and the subsample (on the right side of Table 1.1) reflect the
representativeness of the subsample. The observed literacy scores between any two time points for each group were strongly correlated ($r = .62 – .73$).

### Table 1.1

Sample Sizes, Means, and Standard Deviations of Literacy Assessment Performance by Group

<table>
<thead>
<tr>
<th>Group</th>
<th>Population ($N = 89,609$)</th>
<th>Subsample ($n = 18,000$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>G3</td>
</tr>
<tr>
<td>Eng to Eng</td>
<td>54,613</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>(61.0%)</td>
<td>(0.932)</td>
</tr>
<tr>
<td>Eng to Multi</td>
<td>4,586</td>
<td>−0.145</td>
</tr>
<tr>
<td></td>
<td>(5.1%)</td>
<td>(0.990)</td>
</tr>
<tr>
<td>Multi to Eng</td>
<td>8,701</td>
<td>−0.086</td>
</tr>
<tr>
<td></td>
<td>(9.7%)</td>
<td>(0.952)</td>
</tr>
<tr>
<td>Multi to Multi</td>
<td>21,709</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>(24.2%)</td>
<td>(0.994)</td>
</tr>
</tbody>
</table>

*Note.* Standard deviations are shown in parentheses.

We used multi-group LGCM to examine differences in comparative literacy achievement patterns among students with different patterns of change in their home language environment. Literacy scores showed high kurtosis at all three points (3.14–3.88) and thus violated univariate and multivariate normality. To compensate, an estimator robust to the violation of multivariate normality was used, as mentioned in the method section. Although homoscedasticity was also violated ($p < .001$), LGCM can estimate different time-point-specific variances and does not require this assumption (Preacher et al., 2008). As a preliminary step to evaluate whether there was enough variability across students, we calculated the intra-class correlation (ICC) coefficient, the proportion of variance that is explained by student demographics rather than within-student longitudinal growth. The ICC was .220, which is large enough to warrant the use of LGCM (Hox, 2010). To determine the appropriate shape of the change-in-rank curve, we fitted three unconditional models with different functional forms to the data before applying multi-group analysis. Although the intercept-only model showed adequate fit, except for RMSEA, $\chi^2(4) = 1088.168$ ($p < .001$), CFI = .954, TLI = .966, RMSEA = .123, CI$_{90}$ [.117, .129], SRMR = .054, the linear model demonstrated better fit, $\chi^2(1) = 10.219$ ($p = .001$), CFI = 1.000, TLI = .999, RMSEA = .023, CI$_{90}$ [012, .036], SRMR = .004. The curvilinear model with a latent
quadratic factor could not be identified due to the limited number of time points in the data (Preacher, 2010).

Based on the linear model, multi-group LGCM modeling was conducted by allowing the estimated latent intercept and slope factors to vary by group. The model fit the data well, $\chi^2(4) = 110.595$ ($p < .001$), CFI = .995, TLI = .986, RMSEA = .077, CI90 [.065, .090], SRMR = .015. This resulting model was used to investigate possible differences in relative literacy achievement patterns among multiple subgroups with various home language shift patterns.

The left side of Table 1.2 presents the results of the fitted multi-group model with parameter estimates for each group (see Figure 1.5 for the results embedded in a diagram). The estimated mean intercepts ($\pi_0$) indicate that in G3, Multi to Multi outperformed all other three groups with their literacy performance estimated as being 0.078 SD higher than the population average. While the achievement of Eng to Eng (0.053) was close to that of Multi to Multi, the other two groups, Eng to Multi (–0.157) and Multi to Eng (–0.076), performed lower than the average. With regard to the rate of change, the estimated mean slopes ($\pi_1$) suggest that literacy achievement of Eng to Eng increased, on average, by 0.099 SD between G3 and G10. This estimate of the rate of relative performance change doubled for Eng to Multi and Multi to Eng (0.195 and 0.184, respectively) and almost tripled for Multi to Multi (0.280), suggesting that the rate of change in relative literacy achievement of Multi to Multi over time was the highest among all four groups.
Table 1.2
Results from the Fitted Multi-group LGCM without Covariates (Left) and with Covariates (Right) (n = 18,000)

<table>
<thead>
<tr>
<th>Factors</th>
<th>Linear Model without Covariates</th>
<th>Linear Model with Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eng to Eng</td>
<td>Eng to Multi</td>
</tr>
<tr>
<td>Intercept mean ($\pi_0$)</td>
<td>0.053 ***</td>
<td>–0.157 ***</td>
</tr>
<tr>
<td>Slope mean ($\pi_1$)</td>
<td>0.099 ***</td>
<td>0.195 ***</td>
</tr>
<tr>
<td>Intercept variance ($\zeta_0$)</td>
<td>0.591 ***</td>
<td>0.697 ***</td>
</tr>
<tr>
<td>Slope variance ($\zeta_1$)</td>
<td>0.141 ***</td>
<td>0.194 ***</td>
</tr>
<tr>
<td>Intercept-slope covariance ($\psi$)</td>
<td>–0.057 **</td>
<td>–0.099 ***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Eng to Eng</th>
<th>Eng to Multi</th>
<th>Multi to Eng</th>
<th>Multi to Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept on <em>Born Outside</em> ($\alpha_0$)</td>
<td>0.078</td>
<td>0.159 **</td>
<td>0.257 ***</td>
<td>0.120 ***</td>
</tr>
<tr>
<td>Slope on <em>Born Outside</em> ($\alpha_1$)</td>
<td>0.135</td>
<td>0.136 **</td>
<td>0.066</td>
<td>0.090 **</td>
</tr>
<tr>
<td>Intercept on <em>ESL in G3</em> ($\beta_0$)</td>
<td>–0.025</td>
<td>–0.257 ***</td>
<td>–0.302 ***</td>
<td>–0.343 ***</td>
</tr>
<tr>
<td>Slope on <em>ESL in G3</em> ($\beta_1$)</td>
<td>0.094</td>
<td>0.247 ***</td>
<td>0.182 ***</td>
<td>0.174 ***</td>
</tr>
<tr>
<td>$R^2$ in Intercept</td>
<td>0.000</td>
<td>0.010</td>
<td>0.017</td>
<td>0.040</td>
</tr>
<tr>
<td>$R^2$ in Slope</td>
<td>0.004</td>
<td>0.054</td>
<td>0.029</td>
<td>0.411</td>
</tr>
<tr>
<td>Effective error</td>
<td>0.010</td>
<td>0.008</td>
<td>0.010</td>
<td>0.012</td>
</tr>
<tr>
<td>Effective curve reliability</td>
<td>0.936</td>
<td>0.961</td>
<td>0.919</td>
<td>0.681</td>
</tr>
</tbody>
</table>

Note. ***p < .001, **p < .01, *p < .05. Intercept and slope variances for the conditional model refer to residual variances after variability in their prediction by two covariates.
Figure 1.5
A Path Diagram for the Fitted Multi-group LGCM Without Covariates, with Associated Fitted Parameter Estimates for Means (π₀, π₁), Variances (ζ₀, ζ₁), and Covariance (ψ) for Each Group (n = 18,000)

<table>
<thead>
<tr>
<th>Covariance (ψ)</th>
<th></th>
<th></th>
<th>Variance in Slope (ζ₁)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng to Eng = −0.057**</td>
<td>Eng to Eng = 0.053***</td>
<td>Eng to Eng = 0.0591***</td>
<td>Eng to Eng = 0.099***</td>
</tr>
<tr>
<td>Eng to Multi = −0.099***</td>
<td>Eng to Multi = −0.157***</td>
<td>Eng to Multi = 0.697***</td>
<td>Eng to Multi = 0.194***</td>
</tr>
<tr>
<td>Multi to Eng = −0.065***</td>
<td>Multi to Eng = −0.076***</td>
<td>Multi to Eng = 0.583***</td>
<td>Multi to Eng = 0.115**</td>
</tr>
<tr>
<td>Multi to Multi = −0.062**</td>
<td>Multi to Multi = 0.078***</td>
<td>Multi to Multi = 0.643***</td>
<td>Multi to Multi = 0.025</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intercept (π₀)</th>
<th></th>
<th></th>
<th>Variance in Intercept (ζ₀)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng to Eng = 0.053***</td>
<td>Eng to Eng = 0.053***</td>
<td>Eng to Eng = 0.0591***</td>
<td>Eng to Eng = 0.053***</td>
</tr>
<tr>
<td>Eng to Multi = −0.157***</td>
<td>Eng to Multi = −0.157***</td>
<td>Eng to Multi = 0.697***</td>
<td>Eng to Multi = −0.099***</td>
</tr>
<tr>
<td>Multi to Eng = −0.076***</td>
<td>Multi to Eng = −0.076***</td>
<td>Multi to Eng = 0.583***</td>
<td>Multi to Eng = −0.065***</td>
</tr>
<tr>
<td>Multi to Multi = 0.078***</td>
<td>Multi to Multi = 0.078***</td>
<td>Multi to Multi = 0.643***</td>
<td>Multi to Multi = −0.062**</td>
</tr>
</tbody>
</table>

Grade 3
Literacy Performance

Grade 6
Literacy Performance

Grade 10
Literacy Performance

Note. ***p < .001, **p < .01, *p < .05

In order to test invariance of the means of latent intercept and slope factors across groups, a series of Wald tests was conducted. Table 1.3 shows that most of the p-values from Wald tests comparing all 12 possible pairs were significant, except for two: (1) the intercept factor means of Eng to Eng and Multi to Multi (p = .194), and (2) the slope factor means of Eng to Multi and Multi to Eng (p = .501). These results suggest that all the differences in mean initial status in G3 (the intercept factor) and the mean rate of change between G3 and G10 (the slope factor) across groups are statistically significant, except for the initial score in G3 between Eng to Eng (0.053) and Multi to Multi (0.078), and the rate of change for Eng to Multi (0.195) and Multi to Eng.
(0.184). These findings are further illustrated visually in Figure 1.6, which displays the estimated relative performance trajectory for each group.

Table 1.3

Pairwise Wald Test Statistics of Invariance of Latent Intercept Means (π₀, Upper Diagonal) and Slope Means (π₁, Lower Diagonal) Across Groups (n = 18,000)

<table>
<thead>
<tr>
<th>Group</th>
<th>Eng to Eng</th>
<th>Eng to Multi</th>
<th>Multi to Eng</th>
<th>Multi to Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng to Eng</td>
<td></td>
<td>– 115.03***</td>
<td>45.81***</td>
<td>1.69***</td>
</tr>
<tr>
<td>Eng to Multi</td>
<td>34.75***</td>
<td>–</td>
<td>17.32***</td>
<td>137.99***</td>
</tr>
<tr>
<td>Multi to Eng</td>
<td>26.52***</td>
<td>0.45***</td>
<td>–</td>
<td>62.53***</td>
</tr>
<tr>
<td>Multi to Multi</td>
<td>120.60***</td>
<td>27.06***</td>
<td>33.60***</td>
<td>–</td>
</tr>
</tbody>
</table>

Note. ***p < .001, **p < .01, *p < .05

Figure 1.6

Estimated Changes in Comparative Literacy Performance by Group (n = 18,000)

The parameter estimates for the variance of intercept and slope (ζ₀, ζ₁) provide information on inter-individual differences both in G3 and in their rate of change over time. Across groups, most of the variance estimates for both intercept and slope were significantly
different from 0 (see Table 1.2 and Figure 1.5). These results suggest significant inter-individual variability in both the initial status and rank-change rate and, thus, justify the inclusion of covariates (e.g., immigration status and ESL program enrollment in G3 in the present study) (Byrne, 2012). One exception to the significant variance was the variance in slope for Multi to Multi, which indicates the rate of change in rank for students in the Multi to Multi group was rather homogeneous. The intercept-slope covariances for all four groups are negative, indicating that, regardless of the group, the relationship between the initial status at G3 and the rate of change is negative; that is, students with higher scores in G3 tended to have lower rate of change over time and, in turn, initial gaps in achievement within groups tended to decrease over time (Seltzer et al., 2003).

Effective error and effective curve reliability values for each group were reported at the bottom of Table 1.2 as measures of the precision and reliability, respectively, of the estimated rates of change (Brandmaier, 2020; Brandmaier et al., 2018). Effective error is an unstandardized index that captures how (im)precisely an individual slope estimate measures that individual’s true rate of change in a longitudinal study design, and its inverse is proportional to the precision of measurement. On the other hand, effective curve reliability considers slope variance (ζ_i) and can be interpreted as a standardized effect size. Although the reliability value of Multi to Multi (0.681) appears substantially lower than the other groups (0.919–0.961), this seemingly unsatisfying reliability should be accounted for by relatively small slope variance (0.025), rather than the precision of the instrument (Brandmaier et al., 2018; Singer & Willett, 2003). That is, despite its comparable precision (0.012), detecting inter-individual differences was more difficult for Multi to Multi than for other groups due to its distinctively small individual differences within the group.

RQ3: Do immigration status and ESL program enrollment in grade 3 predict the relative literacy achievement patterns of students with different home language environments?

Two covariates (immigration status and ESL in G3) were added to the model as predictors, resulting in improved model fit compared to the previous model, χ^2(12) = 188.999 (p < .001), CFI = .993, TLI = .979, RMSEA = .057, CI90 [.050, .065], SRMR = .015. With regard to the effects of covariates, the intercept and slope estimates on covariates (α0, α1, β0, β1) on the
right side of Table 1.2 suggest that neither of the two covariates had significant effect on Eng to Eng. Intuitively, students in Eng to Eng are less likely to have immigrated or been enrolled in ESL; the low variability in the covariates in this group must have reduced their predictive power. Yet, for all the other three groups, both covariates were found to be statistically significantly associated with initial status in G3 and rate of rank change. Born Outside was positively associated with literacy achievement in G3 ($a_0$) across three groups, and these significant, positive associations were also found for the rate of change ($a_1$) for Eng to Multi and Multi to Multi. These results indicate that, on average, students from (at least at one point) multilingual homes and who were born outside Canada outperformed their Canadian-born peers in G3 and showed a higher rate of change in relative performance over time. Students who were enrolled in ESL in G3 also demonstrated a higher rate of change ($\beta_1$) across three groups, although their initial literacy achievement in G3 ($\beta_0$) was significantly lower than non-ESL students in G3.

The addition of the covariates to the model resulted in some changes to the means of intercept and slope ($\pi_0$, $\pi_1$) for these three groups, especially for Multi to Multi. Once the immigration status and ESL enrollment in G3 are controlled for, the advantage of Multi to Multi’s initial performance in G3 becomes stronger (0.078 to 0.155). Their rate of change in rank over time somewhat decreases compared to the previous model (0.280 to 0.203) along with the rates of change of Eng to Multi and Multi to Eng, but still maintains the advantage over the other groups. Yet, the decrease in variances ($\zeta_0$, $\zeta_1$) after adding the covariates was rather marginal across groups. Particularly for Eng to Eng, the residual variances are almost identical to those in the previous model, suggesting that, again, the two covariates explained little variance in the intercept or slope for this group.

The $R^2$ values in Table 2 indicate the explained variances in the latent intercept and slope factors by the covariates. Most of the R2 values are minimal, particularly for Eng to Eng, the group for which the covariates have little to predict. Contrariwise, for Multi to Multi, the two covariates collectively explain 41.1% of the variation of the slope factor of the previous unconditional model. This implies that the immigration status and ESL enrollment in G3 have remarkably strong prediction power on the rate of change in relative literacy achievement over time for this group.
Discussion

The present study was designed to examine changes in students’ home language environment and the association between these changes and their relative literacy achievement patterns over time. By investigating province-wide longitudinal data, we were able to portray the home language maps of a linguistically diverse province at different time points and delineate how home language use by individual students and their families changes over time. Between G3 and G6, the proportion of students with an English-dominant home environment increased, whereas students hearing or speaking other language(s) at least as often as English (i.e., OtOt, EqOt, OtEq, EqEq) declined. When comparing these two home language maps (one in G3 and the other in G6), linguistic diversity in the family domain appears to decrease as students become older.

Our findings on changes in home language environment from a longitudinal perspective are consistent with the literature that witnessed prevalent, and rapid, intra-generation language shift (Flores, 2015; Swidinsky & Swidinsky, 1997). Eight out of 10 of the most prominent patterns of changes that we found were those moving from more use of non-English language(s) towards heavier use of English at home. Overall, nearly half of the students’ homes used English to a greater extent in G6 than in G3, whereas approximately a quarter of our sample used less English during the same period. In the homes of the remaining quarter of the students, the extent to which English was used remained the same. Some interesting patterns of changes involved the direction from English-dominant to multilingual homes led by other family member(s) (e.g., EnEn to EqEn, EnEn to EqEq, considering the former part of the notation refers to the language(s) students hear). A possible explanation of these patterns may be that once parents assumed that their children had become proficient enough in English, they might attempt to support their children’s L1 development.

Our LGCM analysis with four groups suggests that students whose home language environment shifted across time regardless of the direction (i.e., Multi to Eng, Eng to Multi) showed significantly lower literacy performance in the earlier grade compared to their peers, but a higher rate of change in relative achievement over time compared to the consistently English-dominant group (Eng to Eng) whose rate of change in rank was negative. Furthermore, students who maintained their home language (Multi to Multi) had significantly greater initial relative
achievement in G3 and a higher rate of change in rank compared to the other three groups. Although their advantage (approximately 0.2 standard deviation in both the initial status and rank-change rate) could be seen as small, this difference in population data should not be considered negligible as this interpretation should be only relative rather than absolute. Our findings suggest at a minimum that maintaining home language in the school years may not impede English literacy development of linguistically diverse students; rather, although we cannot make causal inferences based on our analyses, one possible explanation could be that they may benefit from maintaining and developing multilingual competence.

Examination of the roles of the covariates reveals an interesting finding that immigrant students outperformed their domestic peers in G3 across groups. Considering that 77.3% of these non-Canadian-born students were living in multilingual homes, their higher comparative outcome in English literacy may seem counter-intuitive. Yet, it is important to note that, at the time of G3 administration, the majority of immigrant students had lived in Canada at least for five years and had received formal instruction in English in school for almost five years (two years of kindergarten and three elementary school years). These considerations make it clear that our finding is consistent with that of previous research studies (Cummins, 1981; Jang et al., 2013), which report higher performance of immigrant students after five years of residence in the same geographical context as ours, that is, Ontario, Canada. Furthermore, the fact that students who were born outside Canada and whose home was multilingual in a later year (i.e., Eng to Multi, Multi to Multi) demonstrated higher rates of change in rank, may imply the cognitive benefits of bi/multilingualism. Nonetheless, we acknowledge that this result may have been confounded by students’ socio-economic status, which was not controlled for in the current study. In fact, Sinclair et al. (2019) pointed out an important sociopolitical factor, Canada’s selective immigration policy, as a plausible explanation of higher achievement among immigrant students. Considering this policy is highly selective in approving immigration applications from applicants with higher language proficiency, education, and financial assets (Haque, 2017), some multilingual students are likely to have parents with a relatively high socio-economic background.

Another noteworthy finding is that a relatively large proportion of students receiving ESL support in G3 was Canadian-born (66.2%). When examining by group, 9.3% of Canadian-born,
Eng to Multi students were enrolled in an ESL program in G3. That these students who heard and spoke mostly English at home in G3 were identified as having limited English proficiency highlights that not all Canadian-born students from English-speaking homes are fully prepared for academic language demands at school. Early identification and intervention for these students, who have a lower chance of being identified as English learners than their non-Canadian-born peers (McGloin, 2011), might effectively support them in the long term, as evidenced in the current study by the positive relationship between ESL support in G3 and a higher rate of change in comparative literacy achievement.

Due to the nature of Ontario’s provincial assessment system and items included in the student questionnaire, the current study is limited in that the measures were administered at three time points only; students’ home language environment was self-reported by arguably immature students and was not followed in G10; students’ socio-economic status was not controlled for; and outcome measures were standardized via $z$-scores in order to place the scores from the three different time points in the same metric. As for the last point, the $z$-score metric captured students’ relative growth compared to peers at the time of testing but is not truly sufficient for measuring growth as, to measure each student’s growth independent of their peer-to-peer comparison, the underlying assessments across the years must be vertically scaled. Psychometrically, the metric scale utilized in LGCM analyses can greatly influence the results associated with growth trajectories (Goldschmidt et al., 2010). Modeling individual students’ growth trajectories in academic achievement requires vertically scaled test scores for consistent interpretation across time points (Briggs et al., 2008). Item response theory (IRT)-based vertical scaling is recommended, owing to its increased measurement accuracy (Seltzer et al., 1994); however, it requires common items between grades for either separate or concurrent calibration (Kolen & Brennan, 2004). Unless the assessments assess students in adjacent grades, it is often practically challenging to use common items for achievement testing. IRT-based vertical scaling was not feasible in the current study because the assessments used in the study are not vertically scaled. As such, the rate of changes in relative achievement reported here should be interpreted with caution until the assessments themselves are vertically scaled and the results of this study are replicated with an IRT-based, vertically scaled metric.
The present study contributes to the literature in several meaningful ways. One of the contributions of this study is our specific focus on the complex, dynamic nature of students’ home language status, using large-scale longitudinal data. Investigating relative achievement patterns longitudinally, together with changes in home language environment, provides important evidence for identifying how such contextual factors influence educational outcomes over time. Moreover, our application of LGCM to literacy assessments highlights the importance of longitudinal modeling of both intra-personal and inter-personal variations in relating individual student differences to differential learning trajectories.

Finally, this paper draws implications for promoting students’ home language, instilling a positive view of multilingual competence. Our LGCM analysis demonstrates that students whose homes maintain a multilingual environment (Multi to Multi) show the highest academic achievement in the early grade as well as the highest rate of change in relative performance over time. These students were found to have a clear, although not large, advantage over other groups, especially when they have an immigrant background and received early intervention services (ESL support in G3). Yet, caution must be applied as this study was not designed as a randomized controlled trial. For example, it is possible that parents who are more supportive of education are more invested in the maintenance of the L1 and are less apt to linguistic assimilation.

Yet, it is alarming that, ironically, our findings clearly indicate a movement toward an English-dominant home environment among multilingual students and their families in Canada—a country in which linguistic diversity is supposedly celebrated (Kim et al., 2019). Although language shift from L1 to L2 among language-minority children is viewed as highly detrimental in the existing literature, especially among families in which parents’ L2 skills are limited (Wong-Fillmore, 1991), many immigrant parents have been reported to believe that maintaining the L1 would impede their children’s academic success and, thus, encourage them to maximize the use of L2 even at home (King & Fogle, 2006). Indeed, parents’ perception of, and attitude towards, maintaining the L1 has been known to be one of the most influential factors in their children’s L1 maintenance (Park & Sarkar, 2007; Swidinsky & Swidinsky, 1997). Parents who are learners of L2 (or, learners of English in the context of this study) have a right to make an informed decision on their home language practices based on rigorous research findings. We
close the paper with a call for more research on the relationship between changes in students’
home language environment and their L2 development, the latter of which is directly linked to
their academic achievement. Of equal necessity is more attention from researchers to the factors
affecting the shift in home language use in multilingual families, their perspectives about
language and identity, and effective support strategies for publicly disseminating the positive
aspects of multilingualism.
Acknowledgments

Study 1 was funded by the Ontario Graduate Scholarship. The preliminary findings of this study were presented at the annual conference of the Canadian Association of Language Assessment in 2019. Subsequently, the study was published in the Language Testing journal (see the complete reference below). Within the study, my role encompassed research conceptualization, data cleaning and analysis, and manuscript writing. I also led the revisions and incorporated the constructive input offered by the anonymous reviewers and journal editors. My co-authors contributed to preparing the initial draft of the literature review section and further enriched the manuscript through their insightful feedback and revisions.


https://doi.org/10.1177/0265532220930348
Study 2. Differential Item Functioning Due to Cultural Familiarity on a Reading Achievement Test: Does the Length of Residence Matter?

Considering the significant roles that educational assessments play for accountability purposes in the public school system, ensuring valid score interpretations and fair score use is a matter of paramount importance for educational researchers. With increasing diversity in student populations and rapidly changing learning environments, this mission has become increasingly complex. In Canada, linguistically and culturally diverse (LCD) students make up an estimated 50% of the school-aged population in urban areas (Toronto District School Board, 2017). Previous empirical studies have indicated that standardized assessments in North America were designed primarily for students whose first language is English, potentially leading to the underestimation of abilities or knowledge among students with immigrant backgrounds (Cheng et al., 2007). These challenges underscore the need for careful consideration of the appropriateness and fairness of standardized testing in schools serving high LCD populations.

Differential item functioning (DIF) analysis, a fundamental concept in educational and language assessment, has proven instrumental in deepening our understanding of test bias (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014; Martinková et al., 2017). DIF refers to statistical differences in performance on test items between groups of test takers with comparable abilities (Holland & Thayer, 1988). This psychometric approach helps identify test items that may exhibit inherent bias, thereby favouring certain groups of test takers over others. When the performance on these DIF-flagged items is determined to be influenced by factors irrelevant to the measured construct, such items are said to exhibit substantive DIF or item bias (Roever, 2007). Items demonstrating substantive DIF may undermine test validity, especially its consequential validity (Messick, 1995), by further marginalizing specific groups of students, who often already belong to underserved populations.

There has been extensive research focusing on DIF examination between groups differentiated by gender, ethnicity, and linguistic background, among many others (Li et al., 2022). Yet, the concept of cultural familiarity has received comparatively less attention. Cultural familiarity can be defined as a person’s level of knowledge and familiarity with different cultural
aspects within a specific society, encompassing shared beliefs, values, and practices (Ziegahn, 2001). Previous studies have demonstrated the positive effects of cultural familiarity on reading comprehension (Droop & Verhoeven, 1998; Freimuth, 2008). Length of residence is a well-established indicator of acculturation among immigrants (Kuo & Roysircar, 2011), and students who have lived longer in a specific culture may outperform those new to the society on large-scale reading tests developed for use within the target culture. This performance advantage may persist even after accounting for overall reading ability, suggesting a potential relative disadvantage for newcomers to the society.

The present study aimed to investigate the extent to which certain items in a province-wide reading achievement test function differently across multiple groups of students with varying lengths of residence in Canada as a proxy measure for the level of familiarity with mainstream Canadian culture. By doing so, we sought to contribute to our understanding of the validity and fairness issues of standardized testing for LCD student populations and draw implications for future test development and practices.

**Literature Review**

**Cultural Knowledge and Familiarity as a Source of Item Bias**

Developing reading abilities in students, especially in young learners, is a multifaceted process that involves many key stages and skills (Adams, 1990; Lesaux, 2012). In particular, cultural familiarity can significantly affect the development of reading ability and assessment (Droop & Verhoeven, 1998; Freimuth, 2008). Students who are more familiar with the culture related to a reading text likely have a better understanding of the vocabulary, cultural references and inferences, and contexts within the text. When reading comprehension tests fail to consider students’ cultural familiarity with the context and content of the reading passages, they may inadvertently favor those who are more culturally familiar with the context and content of the reading stimuli, which can skew results and not accurately reflect a student’s reading ability (ElHadad et al., 2017).

An item is said to be biased when one “group of test takers (e.g., males) were less likely to get an item correct than the comparison group of test takers (e.g., females) because the item is tapping a factor over-and-above the factor of interest” (Zumbo & Gelin, 2005, p. 3), and
therefore, item bias is a concept closely related to construct validity and may jeopardize valid interpretations of test scores (AERA, APA, & NCME, 2014). Messick (1989) emphasized that construct-irrelevant variance, along with construct underrepresentation, is a major threat to construct validity as it “makes items or tasks easier or harder for some respondents in a manner irrelevant to the interpreted construct” (p. 7). Construct-irrelevant variance can be present if any features of items interact with test takers’ background variables (Abedi, 2006; Roever, 2007). Such interactions are likely to result in an underestimation of the ability level of specific groups of test takers, which can undermine consequential validity (Messick, 1995).

Several variables tended to be repeatedly used in existing research on item or test bias, such as gender (Khorramdel et al., 2020), race (Santelices & Wilson, 2010), linguistic background (Abedi et al., 1997; Chen & Henning, 1985; Jang & Roussos, 2009), age (Geranpayeh & Kunman, 2007), academic discipline (Pae, 2004), and disability (Lollis & LaSasso, 2009). Cultural knowledge and familiarity were often treated as the same variable as a first language (L1) or race (e.g., Malda et al., 2010; Roever, 2007) due to the intricate relationship of culture with language and race. For example, in order to examine the effect of cultural background knowledge on performance in a test of English as a second language (ESL) pragmalinguistics, Roever (2007) compared the performance of test takers whose L1 was an Asian language (i.e., Japanese, Chinese, Korean, Thai, Vietnamese) with that of test takers whose L1 was German. His findings suggest that some items in the test systematically advantaged the German group over the Asian group. The item that displayed the clearest difference in performance (with seven times higher chances of correct response) between the two groups matched upon their overall ability was related to culturally specific content (i.e., Christianity), and therefore, it was argued that the German test takers benefitted from their own culture, which is based on Judeo-Christian tradition (Roever, 2007).

Similarly, in Malda et al. (2010), both linguistic background and race were used as a grouping variable in a study in which the focus was cross-cultural differences. Their study aimed to investigate the impact of cultural content familiarity on cognitive test performance such as short-term memory, attention, working memory, and figural and verbal fluid reasoning. The study was theoretically grounded in the cultural complexity hypothesis (Helms-Lorenz et al., 2003), which hypothesizes that people familiar with the culture in which a test is developed are
more likely to complete the test successfully than people from a different culture. To test this hypothesis, Malda et al. (2010) had Afrikaans-speaking white children and Setswana-speaking (the language of the Tswana people) black children in South Africa write both of the two versions of cognitive tests: one version in which items with a relatively higher content familiarity for the Afrikaans than for the Tswana children (e.g., computer, shower, swimming pool, cricket); the other version in which items with a relatively higher content familiarity for the Tswana children than for the Afrikaans children (e.g., braids, porridge, wooden spoon, soccer). The results of the multivariate analysis of variance on test scores revealed that the participants generally performed better on the test version designed for their own group than they did on the other version, providing strong support for the role of cultural knowledge or familiarity as a confounding variable in cognitive tests.

However, the level of cultural knowledge or familiarity is neither binary nor static; rather, they are a matter of degree and change over time. Taking the length of residence into account is important if a test is developed and administered in a specific geographic context and taken by test takers with varying lengths of stay including recent immigrants. Immigrants usually enter a new society with their own cultural knowledge from their countries of origin, and they tend to accumulate new cultural knowledge as they become exposed to the predominant culture in the new community (Oppedal & Toppelberg, 2016). By using the length of residence in a given society as a proxy for the level of cultural familiarity, its changing nature can be taken into account in the present study.

**Previous Research on DIF for Language Learners in Large-scale Assessments**

As noted earlier, a variety of student characteristics have been used in previous studies of DIF in large-scale academic assessments (e.g., gender, race/ethnicity, age, academic discipline); yet, language learner status has been one of the most common variables investigated (Banks et al., 2016; Buono & Jang, 2021; Koo et al., 2014; Liu & Bradley, 2021; Mahoney, 2008; Martiniello, 2009; Munist, 2011).

Kim and Jang (2009) examined whether items measuring different reading subskills on the Ontario Secondary School Literacy Test functioned differentially for English language learners (ELL) versus English-as-first-language (EL1) students. Students’ group membership
was assigned based on their self-reported home language use: EL1 if speaking only or mostly English at home, and ELL if speaking only or mostly another language(s) or at least as often as English at home. Adopting Roussos and Stout’s (1996) multidimensionality-based DIF analysis, they found many items assessing vocabulary knowledge favoured EL1 students, while items measuring grammatical knowledge or integrated reading and writing skills unexpectedly advantaged ELL students. Extending Kim and Jang’s (2009) work, Koo et al. (2014) used a meta-analytical approach to investigate DIF patterns across reading test items measuring different reading subskills in grades 3 and 10 students in Florida. In their study, students considered to have limited English proficiency were categorized as ELLs, using receipt of English-for-speakers-of-other-languages services as a proxy. Using the log unit of Mantel-Haenszel odds ratio as an effect size measure, Koo et al. (2014) corroborated Kim and Jang’s (2009) findings in that ELLs have lower performance on the items requiring vocabulary subskill in grade 3 and higher performance on the items concerning critical reading subskill in grade 10. In both Kim and Jang (2009) and Koo et al. (2014), the differential performance on DIF items is attributed to varied strengths and weaknesses in reading subskills between ELL and EL1 students, but not necessarily to item bias. Thus, ELL students’ needs for continued support in literacy development was highlighted for them to succeed in school even after achieving daily oral proficiency.

Other DIF studies focused on potential item bias against ELL students and raised construct validity concerns. Among frequently investigated sources of potential item bias is linguistic complexity in items in content-based assessments (Abedi, 2006; Abedi & Lord, 2001). For example, Martiniello (2009) modeled magnitude of DIF between ELL and EL1 students in a state-wide grade 4 mathematics assessment as a function of linguistic complexity and schematic representation (e.g., equations, diagrams, tables). Multiple regression results showed that the more grammatically and lexically complex items, the greater magnitude of DIF value against ELLs. However, the impact of linguistic complexity on ELL student performance was considerably mitigated by schematic representation, suggesting its potential use as a means to support ELL students in large-scale mathematics assessments. In the Ontario context, Buono and Jang (2021) conducted a similar study in which item-level DIF was investigated using the Ontario’s grade 6 provincial mathematics assessment data. Based on Abedi and Lord’s (2001)
complex language indicator scheme, 11 out of 28 items were flagged as containing linguistically complex features, a construct-irrelevant factor. A confirmatory SIBTEST DIF approach (Stout & Roussos, 1995) identified moderate DIF in six out of the initially flagged 11 items.

Another topic of interest is the cumulative impact of DIF items at the test level by estimating test scores after removing the DIF-flagged items. Munist (2011) contributed to this line of research by detecting DIF between ELL students with immigrant background and EL1 students in Massachusetts’ state-wide assessments. After conducting initial DIF analyses using logistic regression method on the mathematics tests for grades 3 through 5 across multiple administration years, a specific test with a relatively large number of items exhibiting moderate or high levels of DIF was selected for further analysis. When the purified version of this test after removing DIF-flagged items were compared to the original version, the test scores for immigrant students exhibited the most significant improvement. This finding suggests that the cumulative impact of DIF items on immigrant students can have detrimental effects.

Regarding the fluid nature of many characteristics of immigrant students, it is worth noting that Koo et al. (2014) and Munist (2011) explicitly considered the time factor when identifying ELL students. Koo et al. (2014) labeled as ELLs those who had received language support “in the immediately preceding two years” (p. 92). Likewise, in Munist’s (2011) study, the criteria for the membership in the focal group, or “likely immigrant students,” included being in Massachusetts schools for less than five years and had current or previous official limited English proficiency status. However, as these studies did not further categorize ELL students into multiple subgroups, the trend of DIF magnitude as a function of length of residence could not be investigated.

As such, the influence of time that allows test takers to accumulate cultural knowledge or familiarity in a society is currently not well represented in the existing DIF literature. Most previous studies have used language background as the main variable to distinguish groups of students with diverse cultural backgrounds. This approach is commonly employed even when the primary research focus was the different levels of cultural familiarity (Malda et al., 2010; Roever, 2007). However, using only L1 as a proxy for cultural familiarity and knowledge of mainstream culture may not be ideal as it fails to differentiate between first- and second-generation immigrants and does not account for the time-varying aspect of cultural familiarity.
To address these limitations, the current study expands on previous research by using the length of residence in a target-language-speaking society as an indicator of the level of cultural familiarity. This approach aims to examine how the length of residence plays a role in LCD students’ performance on items in a provincial reading achievement assessment.

**DIF Analysis Approaches**

A critical issue in DIF research is to match subgroups on the similar ability (Maller & Pei, 2017). To determine test takers’ ability level, a set of items believed to be DIF-free (called “matching items”) are specified through content experts’ judgment or a purification method. Test takers from reference and focal groups are matched based on their performance on these matching items, and when the probabilities of answering an item under study correctly between the reference and focal groups are significantly different, the item is flagged as exhibiting DIF. When this difference is in favour of one group over the other irrespective of ability level, it is called uniform or unidirectional DIF. If the difference in the probabilities of a correct response is not constant across the ability level and the disadvantaged group can vary along the ability continuum, it is called non-uniform or crossing DIF (Magis et al., 2010; Zumbo, 1999).

Depending on the approach used for ability matching, DIF detection methods can be categorized into two groups: total-score-based and item-response-theory (IRT)-based methods (Maller & Pei, 2017). Among the most widely used total-score-based methods are Mantel-Haenszel procedure (Holland & Thayer 1988), logistic regression method (Zumbo 1999), and SIBTEST (Shealy & Stout 1993), while IRT-based methods such as Lord’s chi-square test (Lord, 1980), Raju’s area method (Raju, 1988), and IRT likelihood ratio test (Thissen et al., 1993) have gained more popularity with advancement in statistical tools (Maller & Pei, 2017). It is also common to use multiple DIF methods in a DIF study (Li et al., 2022) as each of the above-mentioned methods has its own advantages and disadvantages with its unique capabilities.

In DIF analyses, the comparison is made for one or more focal group(s) such as females, immigrants, ELLs, and test takers with disabilities against the reference group. While the statistical methods introduced in the early era of DIF studies were limited to detect DIF between two groups, some of these methods have been extended to detect DIF for multiple groups simultaneously (Magis et al., 2010). Although one can conduct a separate DIF test for each focal
group against the reference group, some drawbacks of this multiple testing practice include a higher Type I error rate and lower statistical power (Magis et al., 2010; Penfield, 2001). Currently, three statistical methods have been extended to examine DIF for multiple groups simultaneously: the generalized Mantel-Haenszel method (Penfield, 2001), generalized logistic regression procedure (Magis et al., 2011), and the generalized Lord’s chi-square test (Kim et al., 1995). Using these extended procedures, researchers can now more easily identify items that have different probabilities of correct responses across multiple groups with the same estimated ability. These procedures allow for identifying multiple focal groups (in this study, LCD students with differing length of stay) in addition to the reference group (in this study, domestic EL1 students).

Nonetheless, not all items flagged by DIF detection methods are subject to test bias because DIF itself is a relatively neutral term that refers to statistical differences in item performance between groups of examinees of comparable ability (AERA, APA, & NCME, 2014; Holland & Thayer, 1988). When differential performance between groups is legitimate given the construct being measured, items exhibit statistical DIF only. On the other hand, items are said to have substantive DIF if item performance interacts with factors not related to the construct of interest (Penfield & Lam, 2000). To determine whether or not items are indeed biased, analyses of item content or test taker-task interactions are often conducted before or after statistical analyses (e.g., Buono & Jang, 2021; Jang & Roussos, 2009; Koo et al., 2014; Martiniello, 2009; Roth et al., 2013). For example, Young et al. (2013) detected a moderate number of items as significant DIF in the Test of English as a Foreign Language (TOEFL) Junior Standard Test. Through follow-up qualitative linguistic analyses, however, it was concluded that most of these items were construct-relevant, and, therefore, the overall test was acceptable in terms of construct validity. Likewise, Jang and Roussos (2009) utilized the think-aloud verbal report data collected from test takers to identify the causes of DIF and substantiate the DIF results obtained from the statistical analyses. As such, qualitative analysis such as item content analysis or interviews with test takers can be highly instrumental in examining sources of DIF and determining its construct (ir)relevance (Ercikan et al., 2010).
The Present Study

The current study investigated the presence of DIF among the provincial reading achievement test items across subgroups of young readers with different levels of exposure to the predominant culture in the society, using the length of stay in the country as a proxy variable. Considering contamination issues with a single-item-at-a-time DIF approach, a confirmatory DIF analytic approach was applied to answer the two research questions:

RQ1. To what extent do items in Ontario’s provincial reading achievement test require knowledge about or familiarity with mainstream Canadian culture?

RQ2. How do the magnitudes of DIF, if any, for these items vary as a function of student’s length of residence in Canada?

Method

Participants

The test item response data of 101,046 students from the 2007 grade 3 English reading comprehension test of Ontario’s provincial assessment program\(^2\) were obtained from the test developer (Education Quality and Accountability Office, 2007b). Among the three grades subjected to a provincial reading test (grades 3, 6, and 10), grade 3 was chosen for analysis given that students who immigrated at an early age may require a longer time to acquire cultural knowledge and become familiar with mainstream culture than those who arrive later in life.

To classify students into the reference (non-LCD students) or one of the LCD focal groups, the study considered two variables: (1) the number of years of residence in Canada, and (2) whether or not English is the student’s first language. The reference group, called Domestic EL1s, consisted of students born in Canada and who learned English as their L1. The focal groups were comprised of students born outside of Canada and identified by parents upon enrollment as not having learned English as L1 at home. These groups were further categorized based on the length of residence in Canada (i.e., Less than 1 year, 1–2 years, 2–3 years, 3–

\(^2\) Unfortunately, we were unable to utilize data from a more recent year due to the test developer’s decision to withhold certain items for the purpose of creating the item pool.
5 years, and 5 years or more). The last focal group, referred to as Domestic ELLs, included students born in Canada but who did not learn English as their L1 at home, and thus likely had less familiarity with mainstream culture than students in Domestic EL1s, formed the last focal group called Domestic ELLs.

Students with missing data on time in Canada or L1 (n = 18,124) were excluded from the study. Students included in the study and those excluded were comparable in terms of gender distribution (49.7% female for both groups; chi-square test p-value < .001) and the observed total score of the 26 multiple-choice items (18.9 and 19.0, respectively; t-test p-value = .004; Cohen’s d effect size = 0.02). Due to the marked differences in group size in the population data, a sample of 484 students was randomly selected from each group, totalling 3,388 students. Table 2.1 provides a summary of the sample sizes of each group in both the population data and the study sample.

Table 2.1
Sample Size by DIF Group

<table>
<thead>
<tr>
<th>Category</th>
<th>L1</th>
<th>Time in Canada</th>
<th>Population data</th>
<th>Study sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Less than 1 year</td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>Focal groups</td>
<td>Not</td>
<td>Less than 1 year</td>
<td>642</td>
<td>0.8%</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>1–2 years</td>
<td>484</td>
<td>0.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2–3 years</td>
<td>697</td>
<td>0.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3–5 years</td>
<td>857</td>
<td>1.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 years or more</td>
<td>2,763</td>
<td>3.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Born in Canada (Domestic ELLs)</td>
<td>10,216</td>
<td>12.3%</td>
</tr>
<tr>
<td>Ref. group</td>
<td>English</td>
<td>Born in Canada (Domestic EL1s)</td>
<td>67,263</td>
<td>81.1%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>82,922</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

For the item content review, six experienced educators were recruited as content experts. The expert group comprised of two categories: (1) Ontario public school teachers and (2) ESL instructors from various educational settings. In the first group, the three experts held Ontario teaching certificates and had extensive experience teaching K–8 students in Ontario schools with diverse community demographics and educational needs. Two of them possessed specialist qualifications in Teaching English Language Learners or Special Education. The second group consisted of three experts with extensive teaching experience in ESL, holding certificates as
Teachers of English as a Second (and Other) Language. They had taught students ranging from elementary school to university level, both in Canada and internationally. Among the six experts, four either held or were pursuing a doctoral degree in education, one possessed a master’s degree in education, and one held a bachelor’s degree in applied linguistics. The experts’ teaching experience ranged from five to 21 years.

Measure

The grade 3 reading achievement test consisted of 26 multiple-choice items with four options each and 10 open-response items (Cronbach’s \( \alpha = .88 \), as reported in Education Quality and Accountability Office, 2007a). As indicated in Table 2.2, a total of five reading passages of different genres were provided, with each passage followed by four to 10 multiple-choice items and two open-response items. While the multiple-choice items were scored dichotomously (either correct or incorrect), the open-response items were scored polytomously with partial credit scoring. As these two types of items require distinct statistical models, for the sake of maintaining simplicity in the analysis, only the multiple-choice items were included in the DIF analysis. Table 2.3 shows the mean and standard deviation of the observed total score for 26 multiple-choice items by DIF group. Furthermore, the proportion correct by DIF group and item-to-total point-biserial correlation coefficients for each item are presented in Table 2.4. They serve as measures of item difficulty and item discrimination, respectively.

Table 2.2

<table>
<thead>
<tr>
<th>Genre</th>
<th>Title</th>
<th>No. of MC items</th>
<th>Corresponding MC items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphic text</td>
<td>My Café Menu</td>
<td>4</td>
<td>Items 1–4</td>
</tr>
<tr>
<td>Informational text</td>
<td>Do You Need More Sleep?</td>
<td>4</td>
<td>Items 5–8</td>
</tr>
<tr>
<td>Long narrative</td>
<td>The Scarecrow’s Hat</td>
<td>10</td>
<td>Items 9–18</td>
</tr>
<tr>
<td>Poem</td>
<td>Hands</td>
<td>4</td>
<td>Items 19–22</td>
</tr>
<tr>
<td>Short narrative</td>
<td>A Better Place to Be</td>
<td>4</td>
<td>Items 23–26</td>
</tr>
</tbody>
</table>

*Note. MC = multiple-choice*
<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focal groups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(LCD students)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than 1 year</td>
<td>18.43</td>
<td>5.07</td>
</tr>
<tr>
<td>1–2 years</td>
<td>16.52</td>
<td>5.11</td>
</tr>
<tr>
<td>2–3 years</td>
<td>17.69</td>
<td>4.95</td>
</tr>
<tr>
<td>3–5 years</td>
<td>18.20</td>
<td>5.09</td>
</tr>
<tr>
<td>5 years or more</td>
<td>19.41</td>
<td>4.70</td>
</tr>
<tr>
<td>Domestic ELLs</td>
<td>18.18</td>
<td>4.91</td>
</tr>
<tr>
<td>Reference group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Domestic EL1s)</td>
<td>19.03</td>
<td>4.82</td>
</tr>
<tr>
<td>Total</td>
<td>18.90</td>
<td>4.85</td>
</tr>
</tbody>
</table>
Table 2.4

Item Difficulty and Discrimination (N = 82,922)

<table>
<thead>
<tr>
<th>Item</th>
<th>Item difficulty (Proportion correct)</th>
<th>Item discrimination (Item-total correlation)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt; 1 year</td>
<td>1–2 years</td>
</tr>
<tr>
<td>1</td>
<td>.64</td>
<td>.48</td>
</tr>
<tr>
<td>2</td>
<td>.70</td>
<td>.67</td>
</tr>
<tr>
<td>3</td>
<td>.61</td>
<td>.60</td>
</tr>
<tr>
<td>4</td>
<td>.56</td>
<td>.46</td>
</tr>
<tr>
<td>5</td>
<td>.68</td>
<td>.61</td>
</tr>
<tr>
<td>6</td>
<td>.96</td>
<td>.91</td>
</tr>
<tr>
<td>7</td>
<td>.63</td>
<td>.56</td>
</tr>
<tr>
<td>8</td>
<td>.69</td>
<td>.56</td>
</tr>
<tr>
<td>9</td>
<td>.83</td>
<td>.78</td>
</tr>
<tr>
<td>10</td>
<td>.82</td>
<td>.76</td>
</tr>
<tr>
<td>11</td>
<td>.62</td>
<td>.49</td>
</tr>
<tr>
<td>12</td>
<td>.58</td>
<td>.56</td>
</tr>
<tr>
<td>13</td>
<td>.71</td>
<td>.55</td>
</tr>
<tr>
<td>14</td>
<td>.69</td>
<td>.65</td>
</tr>
<tr>
<td>15</td>
<td>.84</td>
<td>.77</td>
</tr>
<tr>
<td>16</td>
<td>.82</td>
<td>.74</td>
</tr>
<tr>
<td>17</td>
<td>.74</td>
<td>.73</td>
</tr>
<tr>
<td>18</td>
<td>.75</td>
<td>.70</td>
</tr>
<tr>
<td>19</td>
<td>.58</td>
<td>.52</td>
</tr>
<tr>
<td>20</td>
<td>.73</td>
<td>.62</td>
</tr>
<tr>
<td>21</td>
<td>.87</td>
<td>.78</td>
</tr>
<tr>
<td>22</td>
<td>.76</td>
<td>.68</td>
</tr>
<tr>
<td>23</td>
<td>.93</td>
<td>.92</td>
</tr>
<tr>
<td>24</td>
<td>.50</td>
<td>.42</td>
</tr>
<tr>
<td>25</td>
<td>.68</td>
<td>.63</td>
</tr>
<tr>
<td>26</td>
<td>.51</td>
<td>.37</td>
</tr>
</tbody>
</table>

Analysis Procedure

Item Content Review

The content experts were individually tasked with rating each of the 26 items based on the degree to which the items required or assumed students’ cultural knowledge or familiarity with mainstream Canadian culture on a five-point Likert scale (i.e., Not at all (1), Slightly (2), Moderately (3), Very (4), Extremely (5)). In addition to the ratings, the experts were requested to provide a brief written rationale for each of their ratings. Subsequently, a two-hour virtual focus
A group interview was conducted, involving all six experts. The purpose of the focus group was not to establish a consensus on the ratings but rather to provide an opportunity for the experts to consider different perspectives that they may not have previously considered. During or after the focus group, the experts were encouraged to revise their initial ratings and rationales if they deemed it necessary.

To identify items to be tested for DIF, the means and standard deviations of the modified ratings provided by the six content experts were calculated. The selection criteria for the test items included: (a) high mean ratings, indicating a high level of required cultural knowledge; (b) low standard deviation, suggesting a high level of agreement among the content experts; and (c) a limit of approximately 40% of items (or 10 items) to ensure sufficient items could be included in the matching item set. This criterion allows for the matching of ability levels among students in different groups and helps control Type I errors (Craig, 2017).

**DIF Detection Methods**

To determine the most suitable DIF detection method, three multi-group testing methods were evaluated: the generalized Mantel-Haenszel method (generalized MH; Penfield, 2001), generalized logistic regression procedure (generalized LR; Magis et al., 2011), and the generalized Lord’s chi-square method (generalized Lord; Kim et al., 1995). Given that IRT-based methods estimate the relationship between item responses and latent traits while accounting for the specific item properties in DIF analysis, the generalized Lord method was preferred to the other two approaches. IRT-based detection methods, including the generalized Lord method, requires the assumption of local independence among test items. A comparison between two confirmatory factor analysis models—a one-factor model vs. a five-factor model grouped by items that shared the reading passage—showed that the unidimensionality assumption could be considered being met (CFI > .90, TLI > .90, RMSEA < .08, SRMR < .10) although the goodness-of-fit indices of the one-factor model were slightly inferior to those for the five-factor model (see Table 2.5 for detail).
Table 2.5

Goodness of Fit Indices of Two Confirmatory Factor Analysis Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Chi-square Test</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA Estimate</th>
<th>90% CI Lower</th>
<th>90% CI Upper</th>
<th>SRMR</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 factor</td>
<td>973.85</td>
<td>5</td>
<td>&lt; .001</td>
<td>.940</td>
<td>.026</td>
<td>.028</td>
<td>.025</td>
<td>89675</td>
<td>89412</td>
</tr>
<tr>
<td>5 factor</td>
<td>690.91</td>
<td>289</td>
<td>&lt; .001</td>
<td>.964</td>
<td>.020</td>
<td>.018</td>
<td>.022</td>
<td>89993</td>
<td>89792</td>
</tr>
</tbody>
</table>

Note. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval; SRMR = standardized root mean square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion.

The Lord method was also deemed particularly useful for the current study as its delta difference statistic under the one parameter logistic (1PL) model can be used as a measure of DIF magnitude (Penfield & Camilli, 2007). Moreover, Lord’s two parameter logistic (2PL) model can provide more insight by detecting non-uniform DIF effects through item discrimination parameter estimation. Thus, the generalized Lord (Kim et al., 1995) and the traditional Lord (Lord, 1980) were selected as the main DIF detection methods in this study.

**DIF Analysis Procedure**

Prior to conducting multi-group DIF analysis, item purification was performed using the generalized Lord method with a 2PL model (Kim et al., 1995). This purification process improves the accuracy of the analysis by reducing Type I error (Magis et al., 2010). It ensures that only DIF-free items are included when calculating total test scores, which are subsequently used to match ability levels (Maller & Pei, 2017). Any items flagged as DIF during the purification step were removed from the matching item set.

The generalized Lord method with a 2PL model (Kim et al., 1995) was again used for detecting DIF items between the reference group and multiple focal groups simultaneously. Items that passed the purification step but were not selected for testing comprised the matching item set. Items flagged in multi-group analysis underwent pairwise comparisons between each individual focal group and the reference group using the traditional Lord chi-square test (Lord,
1980). Pairwise comparisons help identify where differences in performance exist and compare the magnitudes of DIF estimates among different pairs of groups under the Lord 1PL model.

To investigate the changes in the magnitude of DIF of any DIF-flagged items as a function of students’ length of residence, the difference between item difficulties of the reference group and the focal group, multiplied by –2.35 (or, ΔLord), was used as an effect size measure (Penfield & Camilli, 2007). The degrees of DIF were interpreted using the delta classification scheme developed by Educational Testing Service: negligible if |ΔLord| ≤ 1, moderate if 1 < |ΔLord| ≤ 1.5, and large if |ΔLord| > 1.5 (Holland & Thayer, 1985; Zieky, 1993; Zwick, 2012). A positive ΔLord value indicates an item favouring the focal group while a negative value indicates an item favouring the reference group. Lastly, DIF-flagged items in earlier steps were further tested using the Lord 2PL model to investigate non-uniform DIF effects. The item characteristics curves between the reference and focal groups were compared visually across the length of residence in Canada for each flagged item. All DIF analyses were conducted using the difR package in R (Magis et al., 2010). The a priori significance level set for this study was α = .05.

**Results**

**Identification of Items to Be Tested for DIF**

The required level of cultural knowledge, as rated by experts, was used as criteria for the selection of items to be under examination. Figure 2.1 illustrates the means and standard deviations of the ratings from the six content experts for all 26 items in descending order of mean rating. A general tendency of lower standard deviation for items with higher means was observed ($r = -.50$), suggesting the expert ratings tended to converge more for items with a higher level of cultural familiarity required. Given the first five items showing an average rating above 4.0 (“Very”) and the relatively large drop in the mean rating between the fifth (Item 17) and sixth (Item 3), the first five items (Items 4, 2, 22, 1, and 17) were selected to be tested for DIF. The contents of these five items, as appeared in the test, are provided in Appendix A, along with the corresponding reading passages.
DIF Analysis

To ensure that the matching criterion was free of any DIF items, item purification was performed on all 26 items. The generalized Lord analysis with the six focal groups indicated that in addition to the items by content expert ratings, Items 13 and 14 also exhibited DIF. Thus, besides the five items selected to be tested, these two additional items (Items 13 and 14) were removed from the matching item set for further analysis, lowering the number of anchor items to 19 out of 26.

With these 19 anchor items as matching criterion, the five items selected through the expert ratings were tested for DIF for multiple focus groups simultaneously using generalized Lord. The analysis flagged three out of the five items tested (Items 2, 4, and 22). Figure 2.2 provides a visual representation of the generalized Lord chi-square statistics by item with DIF-flagged item numbers in red. These results indicate that these three items have different probabilities of correct response between the reference group and one or more focal groups,
controlling for ability estimated based on their responses to the purified 19 items. It is apparent from the figure that the generalized Lord chi-square statistics considerably varied among items.

Figure 2.2

*Results of Simultaneous Multi-group Comparison Analysis (Generalized Lord)*

![Graph showing generalized Lord's $\chi^2$ statistic](image)

*Note.* The horizontal line indicates the detection threshold of $\chi^2 = 21.03$ at the .05 significance level. DIF-flagged items are indicated in red.

For the three DIF-flagged items, the pairwise comparisons were conducted using the Lord method to compare performances of each of the six focal groups (i.e., *Less than 1 year, 1–2 years, 2–3 years, 3–5 years, 5 years or more, Domestic ELLs*) with the reference group. Table 2.6 presents the results obtained from these analyses by pair of groups compared: item numbers, Lord chi-square statistics, $p$-values, and $\Delta$Lord (the absolute value of which indicates the magnitude of DIF or the effect size). In each pair comparison, at least two out of three items were identified as DIF. Specifically, compared to the reference group, the *Less than 1 year* and
1–2 years groups had three DIF items while the remaining groups showed two DIF items. For all DIF-flagged items, $\Delta$Lord values were negative, suggesting that all the DIF-flagged items favoured the reference group and disadvantaged the focal groups.

### Table 2.6

Results of Pairwise Comparison Analysis (Lord)

<table>
<thead>
<tr>
<th>Groups compared</th>
<th>Item</th>
<th>Lord $\chi^2$</th>
<th>$p$</th>
<th>$\Delta$Lord (effect size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 1 year</td>
<td>2</td>
<td>27.48</td>
<td>&lt;.001***</td>
<td>–2.11</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>4</td>
<td>4.30</td>
<td>.038*</td>
<td>–0.64</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>22</td>
<td>5.87</td>
<td>.015*</td>
<td>–0.91</td>
</tr>
<tr>
<td>1–2 years</td>
<td>2</td>
<td>13.24</td>
<td>&lt;.001***</td>
<td>–1.42</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>4</td>
<td>4.43</td>
<td>.035*</td>
<td>–0.66</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>22</td>
<td>7.28</td>
<td>.007**</td>
<td>–1.00</td>
</tr>
<tr>
<td>2–3 years</td>
<td>2</td>
<td>19.15</td>
<td>&lt;.001***</td>
<td>–1.73</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>4</td>
<td>6.77</td>
<td>.009**</td>
<td>–0.83</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>22</td>
<td>2.35</td>
<td>.126</td>
<td>–0.52</td>
</tr>
<tr>
<td>3–5 years</td>
<td>2</td>
<td>28.87</td>
<td>&lt;.001***</td>
<td>–2.10</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>4</td>
<td>19.19</td>
<td>&lt;.001***</td>
<td>–1.43</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>22</td>
<td>0.68</td>
<td>.410</td>
<td>–0.17</td>
</tr>
<tr>
<td>5 years or more</td>
<td>2</td>
<td>23.86</td>
<td>&lt;.001***</td>
<td>–1.98</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>4</td>
<td>15.46</td>
<td>&lt;.001***</td>
<td>–1.31</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>22</td>
<td>0.97</td>
<td>.326</td>
<td>–0.29</td>
</tr>
<tr>
<td>Domestic ELLs</td>
<td>2</td>
<td>19.21</td>
<td>&lt;.001***</td>
<td>–1.73</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>4</td>
<td>6.49</td>
<td>.011*</td>
<td>–0.80</td>
</tr>
<tr>
<td>vs. Reference Group</td>
<td>22</td>
<td>1.11</td>
<td>.293</td>
<td>–0.31</td>
</tr>
</tbody>
</table>

Note. The detection threshold was $\chi^2 = 3.84$ at the .05 significance level. ***$p < .001$, **$p < .01$, *$p < .05$.

Figure 2.3 visualizes the changes in the DIF magnitude as a function of the length of residence based on the $\Delta$Lord values reported in Table 3. The $\Delta$Lord values of only the items identified as DIF in each pairwise comparison analysis were displayed. Although not apparent, the trend of DIF magnitude for the flagged items, suggested by the absolute value of $\Delta$Lord, tended to eventually decrease or even disappeared as they stay longer in Canada. Item 22, which was flagged as negligible DIF for the Less than 1 year and 1–2 years, no longer exhibited DIF for the students who had lived in Canada for two years or more. For Items 2 and 4, the absolute
values of ΔLord fluctuated or increased for their first five years of residence in Canada, but started to drop for students who had been exposed to the predominant culture for five years or more. While Item 4 was flagged mostly as negligible to moderate DIF, Item 2 exhibited large DIF against almost all focal groups when compared to the performance of the reference group.

**Figure 2.3**

*Changes in the DIF Magnitude by Item as a Function of Length of Residence*

Non-uniform DIF effects were examined using the Lord 2PL model. Figure 2.4 illustrates the item characteristics curves for the reference and focal group for each pair comparison for all DIF-flagged items. Overall, the plots show that the DIF effect was the highest for Item 2, followed by Item 4 and Item 22. For Item 2, the DIF effect was smaller or even reversed at the low ability level (θ < –2) for students who had lived in Canada for less than three years. Yet, students who had lived in Canada for longer than three years had much lower probability of a correct response than their domestic EL1 peers, or the reference group, regardless of their ability level. Item 4 functioned differentially against all focal groups, but more so for mid-to-high performers (–1 < θ < 3) as demonstrated by the larger gaps between two lines. For Item 22, low performers (θ < –1) who had lived in Canada for less than one year clearly underperformed their
EL1 peers, but the performance gap quickly disappeared as the two lines cross and significantly overlap for other focal groups.
Figure 2.4

*Item Characteristics Curves for Three DIF-flagged Items*

<table>
<thead>
<tr>
<th></th>
<th>Less than 1 year</th>
<th>1–2 years</th>
<th>2–3 years</th>
<th>3–5 years</th>
<th>5 years and more</th>
<th>Domestic ELLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 2</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
<tr>
<td>Item 4</td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
<td><img src="image9" alt="Graph" /></td>
<td><img src="image10" alt="Graph" /></td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
</tr>
<tr>
<td>Item 22</td>
<td><img src="image13" alt="Graph" /></td>
<td><img src="image14" alt="Graph" /></td>
<td><img src="image15" alt="Graph" /></td>
<td><img src="image16" alt="Graph" /></td>
<td><img src="image17" alt="Graph" /></td>
<td><img src="image18" alt="Graph" /></td>
</tr>
</tbody>
</table>

*Note.* In each plot, the $x$-axis represents the ability level ($\theta$) while the $y$-axis represents the probability of a correct response. The solid (darker) line represents the reference group, and the dotted (lighter) line represents the focal group.
Discussion

The present study set out to investigate the relationship between students’ length of stay in Canada and their performance on items in a reading achievement test, after controlling for their reading ability. Through the content experts’ item review, five out of 26 items that require the highest level of cultural familiarity were selected for DIF testing. The simultaneous multi-group DIF analysis, using the generalized Lord procedure, flagged three items (Items 2, 4, and 22) as exhibiting DIF against the focal groups. This indicates that these items systematically disadvantage LCD students with limited knowledge on mainstream culture. A subsequent examination of these flagged items revealed a decreasing trend in the magnitude of DIF as a function of students’ length of stay, although two items (Items 2 and 4) exhibited moderate to large DIF even for ELL students who had lived in Canada for more than five years or were born in the country.

A close examination of the item content and the rationales for the content experts’ ratings is conducive to making sense of the results from the DIF analysis. Both Items 2 and 4 are linked to the same reading passage—a menu that would be typical in a North American diner. At the item level, Item 2 is intended to assess students’ vocabulary skill by asking for the antonym of “chilly” in relation to the food chili. The content experts rated the level of cultural knowledge required as an average of 4.7 out of 5. All the six experts pointed out that this item certainly draws on culturally specific knowledge of the food item chili, the idiomatic word “chilly”, and the pun of chili-chilly as homonyms, favouring domestic EL1 students over LCD students. Similarly, Item 4 asks why some dishes are grouped separately under the subtitle “On the Side” on the menu. All the content experts agreed that this item may present a greater challenge to students who are unfamiliar with the concept of North American “side dishes” with an average rating of 4.8. One expert’s rationales nicely summarized the points discussed and agreed upon during the focus group:

First, the question and possible answers assume children understand the sequencing of meals commonly eaten in Canada. In many cultures, salad and “sides” may follow a main dish. Second, it implies students understand the language and context of “sides” in a restaurant, as this word can have multiple meanings. Lastly, this question also assumes children belong to a certain socioeconomic class that frequents restaurants. There are
children who have not been to restaurants and therefore may not understand menus and restaurant culture.

The DIF analysis results from the present study provide compelling evidence that the ability to answer these items correctly may have interacted with each student’s cultural background, which is irrelevant to the construct measured in a reading achievement test. These two items disadvantaged students who are less familiar with the predominant culture in Canada including those who had lived in Canada for five years or more.

On the other hand, the results also suggest that LCD students acquire knowledge about the predominant culture and close the performance gap with their domestic EL1 peers as they are exposed to mainstream culture for a longer period of time and learn about shared beliefs, values, and practices. Item 22 follows a poem titled “Hands” and are intended to assess students’ understanding of the (implicitly stated) main idea. Although the poem promotes diversity by describing different hands, the content experts found this item presuming Canadian-specific cultural knowledge because some descriptions such as “freckled and tanned” are in relation to whiteness, resulting in the mean rating of 4.7. One content expert wrote:

This question is culturally loaded for the immigrant students who are from monoethnic cultures/countries. This whole poem is about diversity, but this particular question about different hands (especially colours) is probably specific to Canada, a society where multiculturalism is encouraged and fostered.

This item exhibited moderate DIF for students newer to Canada, but the DIF flag disappeared once students who lived in the country for two years or more. This indicates that performance of LCD students on certain items becomes comparable to that of domestic EL1 students as they stay in the country longer. In the case of the current study, as part of official discourse in public schools, LCD students may learn diversity as one of the shared values in Canadian society during their first years of residence, letting alone any contentious issues regarding how it is taught (Gérin-Lajoie, 2008; Tuters, 2015). Thus, the different DIF pattern in Item 22 compared to that in Items 2 and 4 may be attributed to the different likelihood of becoming familiar with the notion of diversity during the initial period of stay in Canada compared to the predominant food culture.
By examining cultural familiarity as a time-varying construct and employing the less common approach of the extended DIF detection method (Magis et al., 2010), this study contributes to our understanding of the role of length of residence in school-based assessments for LCD students. The findings reveal a general trend of decreasing DIF as a function of length of residence, which was used as an indicator of the level of cultural familiarity with the predominant culture. Furthermore, the study highlights that culturally biased items, depending on their cultural content, can impact not only recent immigrant students with limited language proficiency but also those who have resided in the society long enough to acquire language and literacy skills comparable to their EL1 peers.

Reading comprehension involves not only decoding text but also understanding the cultural nuances and references embedded within the text (Steffensen et al., 1979). From this perspective, cultural familiarity can be considered as an essential component of reading ability, and the observed DIF phenomenon in this study reflects the inherent complexity in disentangling cultural knowledge from reading ability, rather than item bias. Indeed, attempting to separate cultural knowledge from reading comprehension as a measured construct becomes complex, as both elements are intertwined and influence each other. However, it is crucial to consider the broader implications of incorporating cultural knowledge into reading assessments, particularly in standardized testing. Standardized tests aim to “create conditions, questions, scoring procedures, and interpretations that are consistent across schools” (Morris, 2011, p. 5), regardless of student characteristics. The growing cultural diversity among students poses challenges in developing reading test items that are culturally unbiased yet representative of the construct being measured. By including items that require specific cultural knowledge, there is a risk of introducing cultural bias into the assessment process (ElHadad et al., 2017; Reynolds & Carson, 2005). Students from different cultural backgrounds may be disadvantaged if the assessment assumes knowledge that is not universally shared.

As such, the discussion surrounding the inclusion of cultural knowledge in reading assessments and its impact on construct definition and fairness in standardized testing requires nuanced considerations. Striking a balance between cultural relevance and fairness in standardized testing remains a crucial area of exploration and improvement in educational assessment practices (APA, AERA, & NCME, 2014). The International Test Commission’s
(2018) guidelines for large-scale assessments administered to LCD populations recommend reviewing test items during the item development stage to identify any content that may be “sensitive or unacquainted” (p. 13) for diverse test takers. To ensure a thorough item review process, it is advised to engage external panels of experts and target test takers who can identify any test content that may be “differentially familiar or interpreted differently by members of different groups” (AERA, APA, & NCME, 2014, p. 54). While many test developers have already adopted this recommendation (e.g., Education Quality and Accountability Office, n.d.), this practice has become even more important with the increasing use of automated item generation through artificial intelligence (Ercikan et al., 2023). Consequently, test developers may consider reallocating their human resources from item generation to item review, reflecting the evolving practices in the field.

Regarding the study’s methodological approach, some limitations should be acknowledged. First, due to the cross-sectional design, actual changes in DIF magnitude as students experience mainstream culture over time can only be inferred from different student groups with varying lengths of residence. In terms of the data, although the overall reading ability was matched using the purified matching item set, the DIF results may have been affected by other confounding factors such as socio-economic status, urban or rural areas, and country or culture of origin, all of which were unavailable in the dataset. Additionally, the DIF investigation was limited to multiple-choice items, resulting in a reduced number of items included in the analysis. The interpretation of the DIF analysis relied solely on content experts’ item review and focus group. Future research incorporating additional item types (e.g., constructed response items) and data sources (e.g., student cognitive interviews, think-aloud protocols as suggested by Ercikan (2010)) would enhance our understanding of DIF patterns and causes.³

A natural progression of the current study is to examine the impact of DIF-flagged items on minority student groups if all assessment items can be included in the studied dataset. As mentioned earlier, such investigation of the impact of DIF between immigrant and non-

---

³ Initially, our plan was to collect qualitative data through think-aloud protocols with grade 3 students in Ontario schools. However, due to the COVID-19 pandemic, it was pivoted to an expert content review followed by a virtual focus group interview.
immigrant students was conducted for a state-wide mathematics test by Munist (2011), which calculated corrected score estimates for each test taker and demonstrated the impact of removing DIF items on schools or school districts with higher concentration of black, low-income, and likely immigrant students. A similar study in reading or literacy assessments with multiple focal groups can be a fruitful area of research. Recent methodological advancements in the DIF impact literature include differential test functioning statistics (Chalmers et al., 2016), performance comparison among effect size measures (Joo et al., 2022), and the impact estimates on group-level mean score (Finch & French, 2023).

Altogether, the present study holds significant implications for future test development and practices, aiming to address issues of test validity and fairness for all students, irrespective of their linguistic and cultural backgrounds. It highlights that language proficiency alone is not the sole obstacle for LCD students to fully demonstrate their learning in standardized achievement tests (Fox & Cheng, 2007). These tests, when used in schools with substantial LCD student populations, can inevitably present uneven opportunities for students to demonstrate their knowledge and skills. The present study suggests the possible presence of culturally biased items in large-scale assessments in Canada and elsewhere, which potentially disadvantage minority student subgroups (van de Vijver & Poortinga, 1997). The inclusion of culturally biased items threatens the validity of test score interpretation for students with diverse cultural backgrounds (International Test Commission, 2018), as these items pose greater difficulties for LCD students compared to their mainstream peers.

As both Fox and Cheng (2007) and Solano-Flores and Trumbull (2003) stated, developing an absolute culture-free or culturally neutral test may be impossible because tests are “inevitably cultural devices” (Solano-Flores & Trumbull, 2003, p. 9). Nevertheless, test developers should pay greater attention in related to items that require specific knowledge on mainstream culture so that score interpretation and use can be valid for every student, including students with distinctive cultures of their own such as language learners, Indigenous students, and students with immigration or refugee background.
Acknowledgments

Study 2 was funded by the Social Sciences and Humanities Research Council Doctoral Fellowship, Cambridge Doctoral Dissertation Grant from the International Research Foundation for English Language Education, Language Learning Dissertation Grant from the Language Learning journal, and Duolingo Doctoral Dissertation Award from Duolingo English Test. The preliminary findings of this study were presented at the Annual Conference of the American Educational Research Association in 2018 (with Drs. Jeanne Sinclair and Eunice Eunhee Jang as co-authors) and the Language Testing Research Colloquium of the International Language Testing Association in 2021 (as a sole author). I was responsible for the development of the research questions, data cleaning and analysis, and preparing for the initial draft of the manuscript. The second and third authors provided feedback and revisions to an earlier version of this manuscript prepared as a conference proposal. The manuscript has been submitted to a journal and is currently under review for publication.
Study 3. Assessing Heritage Language Learners in Ontario’s International Languages Elementary Program: A Case Study of Korean Classes

Since its adoption of official multiculturalism in 1971, Canada has formulated and implemented policies that can support students with diverse linguistic backgrounds to maintain and develop their first language (Early et al., 2017). In the province of Ontario, the government introduced a publicly funded language education program called Heritage Languages Program\(^4\) in 1977 (Ontario Ministry of Education [MOE], 1991). Over the past four decades, the program has provided invaluable opportunities for K-8 students to learn their ethnic languages in the public-school system. Nearly 100,000 students across the province are receiving language instruction every year under this now-called International Languages Elementary (ILE) program (Cummins, 2014).

Although seemingly well-intended, there has been a constant concern, since its inception, about the implementation quality of the ILE program. Previous studies point out that the program is not adequately supported in terms of curriculum development, teaching materials, physical facility, implementation guidelines, and professional training, resulting in unintended consequences such as poor program quality and marginalization of heritage language education (Berryman, 1986; Feueverger, 1997; Mycek, 2015). The major source of these issues that Berryman (1986) identified pertained to the policy that restricts the program to be offered only outside regular school hours. Because the ILE program is not part of the regular curriculum and most classes are offered as after-school or weekend programs, it receives less support from the government and school boards and, thereby, often perceived as inferior in terms of quality and status. Considering that no changes have been made to this outside-regular-school-hours-only policy since the introduction of the program (Kim et al., 2020), the issues appear to remain unresolved as evidenced in more recent studies (e.g., Mycek, 2015).

This case study critically examined the extent to which Ontario’s ILE program is operating as intended in its Korean classes in the Greater Toronto Area by evaluating the

\(^4\) The term “heritage languages” is often used in Canada to refer to languages other than two official languages (English and French).
program activities, resources, implementation procedures, implementation barriers, and program modifications made to meet local program needs. The ultimate goal of the current study was to provide thick descriptions of the program operation in its Korean classes, find areas in need of improvement, and provide suggestions to enhance the quality of the program (Mertens & Wilson, 2012). We took an improvement-focused approach (Posavac, 2015; Stufflebeam, 2001) and aimed to examine the core activities in the program, discover discrepancies between what was planned and what is being implemented, and assess the program operation in the Korean classes based on its stakeholder needs and experience. Specifically, we narrowed the focus of the evaluation into assessment practice within the program because the Ontario Ministry of Education has brought significant changes to its assessment and reporting policy over the past decade, mainly shifting from assessment of learning to assessment for learning (Ontario MOE, 2010). It was our intention to investigate how these external assessment policy changes were integrated into the program, specifically in the case of Korean classes. After individual interviews with teachers and parents of Korean language classes, the participants’ experiences and perceptions were systematically analyzed based on grounded theory to reveal any causal relationships among the themes identified. Implications on how to improve the current assessment and reporting system as well as the program in general are also discussed.

Literature Review

Evaluation of Language Programs

Norris (2006) described program evaluation as “the gathering of information about any of the variety of elements that constitute educational programs, for a variety of purposes that primarily include understanding, demonstrating, improving, and judging program value” (p. 579). When a program evaluation aims to find areas in need of improvement by examining how the program is being implemented and any discrepancies between planned and actual implementations, this type of evaluation is called process evaluation (Mertens & Wilson, 2012; Posavac, 2015). The importance of understanding program implementation—as opposed to program outcome—and providing usual information for local decision makers has been widely recognized, drawing attention to qualitative data collection approaches including stakeholder
interviews, observation protocols, and focus groups proficiency tests (Norris, 2016; Norris & Watanabe, 2013).

In the field of language education and applied linguistics, it was not until the late 1990s that the context of evaluation studies was diversified from the typical English-as-a-second-language programs to other settings such as foreign or heritage language settings with a non-English language as a target language (Norris, 2016). For example, Elder (2009) synthesized three evaluation studies for Australia’s bilingual programs (Vietnamese–English, Chinese–English, and Arabic–English, respectively) conducted in the late 1990s. Sohn and Merrill (2008), possibly the only published program evaluation work for Korean as a heritage language, presented findings from two outcome evaluative studies based on Korean–English dual language programs in Los Angeles, California. For future directions, Norris (2016) encourages researchers to use a program logic model, which is a process visualization of the program’s theory of change with identified program components (see the following section for details). Norris also emphasized the possibility of using program evaluation as a vehicle for assessment validation and suggested researchers reconceptualize assessments as entire programs.

The International Languages Elementary Program in Ontario

The main objective of the ILE program is to “encourage students to maintain, recover, or acquire a degree of fluency in their chosen languages of study” and, in turn, help them “function more effectively both in Canada’s multicultural society and in the international community” and “strengthen their awareness of their own ethnocultural heritage” (Ontario MOE, 1991, p. 4). All publicly funded school boards in Ontario are mandated to offer courses of any languages other than English and French as part of their continuing education program upon request from parents of 23 or more students (Ontario MOE, 2012). The program was established as “Heritage Languages Program” in 1977, but in early 1990s, the provincial government changed the name of the program to “International Languages Elementary” to better communicate that developing these languages has economic value, and that any students, regardless of their ethnic background, can be enrolled in the program to acquire language skills in any language other than English and French (Cummins, 2014; Duff, 2008). Yet, most classes, especially those for languages with less economic value, are comprised of students with heritage ties to the language taught.
Under the program, students are entitled to receive up to 2.5 hours of language and culture instructions per week outside the regular school hours. As to the timing of class offerings, a variety of program models, such as an integrated extended-day model can be supported in theory (Ontario MOE, 2012), but in reality, most classes are offered weekly through after-school, evening, or weekend program operation. The program is funded by the provincial government based on a rate of 59 Canadian dollars per classroom hour as of the 2022–2023 academic year (Ontario MOE, 2022) although parents are asked to pay a minimal fee during the yearly registration in some school boards. The program is quite sizable in terms of both budget granted and student enrolment; in the 2018–19 school year, $28 million was invested to teach over 80,000 students across Ontario (Ontario MOE, 2018).\(^5\) Given a large portion of recent immigrants reside in urban areas, most of the student enrolment is concentrated in large urban school boards. For example, in Toronto alone, nearly 40,000 students—50% of the total enrolment in Ontario—are studying over 50 different languages through the ILE program (Toronto Catholic District School Board, 2019; Toronto District School Board, 2019).

While the Ministry sets relevant policies and provides funding for the program, each school board is responsible for “establishing, introducing, designing, administering, and supervising all aspects of these programs” (Ontario MOE, 2012, p. 18). School boards should develop procedures and guidelines; hire and train instructors; provide resources, supplies, and facilities; and monitor the implementation. Since the program is considered a form of continuing education, instructors are not required to hold Ontario teaching certificate (Ontario MOE, 2012).\(^6\) Instead, they often speak the language taught as their first language and have teacher training in their home country (Feueverger, 1997; Mycek, 2015). Instructors are responsible for planning instruction and directly supporting student learning.

Several studies have examined the implementation of the ILE program. The first, and most comprehensive thus far, is an evaluation study by Berryman (1986). Using Fullan’s (1982)

\(^5\) As of the 2021–2022 year, the number of elementary students (K–8) was approximately 1.4 million.

\(^6\) This is not the case for credit-bearing courses of Classical Studies and International Languages at the secondary level (grades 9–12). Only the international language classes at the elementary level (K–8) are excluded from the regular curriculum and, thus, housed under Continuing Education through the ILE program.
factors affecting implementation as the theoretical framework, Berryman evaluated the extent to which related policies were being implemented into the integrated extended-day-school programs in two Toronto school boards (i.e., Toronto Board of Education, Metropolitan Separate School Board). Based on the interviews and survey data collected from administrators, school staff, students, and parents, Berryman pointed out many problems that occurred due to poor implementation planning and offered specific recommendations for the Ministry of Education to consider. Other relevant studies, although not labeled as program evaluation, conducted survey, interviews, or class observations (or a combination of these methods) with the purpose of, to varying degrees, improve programs and offer valuable suggestions (Feuerverger, 1997; Fiorucci-Nichollas, 1998; Mercurio, 1997; Mercurio-Berrafati, 2009; Mycek, 2015). Most of these studies were contextualized in the City of Toronto geographically (except for Mercurio, 1997; and Mercurio-Berrafati, 2009) and linguistically focused on one language—mostly Italian (except for two teacher-focused studies; Feuerverger, 1997, Mycek, 2015). The issues and recommendations identified by these studies conducted over the last three decades overlap in many aspects and point to the marginalized status of the program and its poor implementation quality as a result.

**Paradigm Shift in Educational Assessment**

In language education, assessment refers to the process of gathering information that supports making judgments about individual student’s language development (Cheng et al., 2004). Yet, this value-neutral definition of assessment has not been widely spread among laypeople, many of whom would mistakenly identify assessment with testing. In classroom settings, the primary purpose of assessment should be not only to collect information that helps teachers monitor the academic progress their students make (assessment of learning, or summative assessment) but also to identify their educational needs for further improvement (assessment for learning, or formative assessment) (Birenbaum et al., 2015; Stiggins, 2005). Formative on-going assessment encompasses informal teacher-student interactions and systematic pedagogical documentation through which teachers can better understand each student’s level of understanding of learning materials and the nature of engagement in learning activities. Formal diagnostic and end-of-unit tests are used to judge the level of achievement and further monitor the rate of progress over time. There is an increasing recognition of importance for supporting student autonomy through self-regulated learning (assessment as learning). Thus,
a variety of assessment methods are expected to be used by teachers for assessment of learning, for learning, and as learning. This expansion of classroom assessment purposes has gained considerable interest amongst researchers and practitioners in K–12 education across the curriculum as well as additional language education (Black & Wiliam, 1998; Lynch, 2001).

The Ontario Ministry of Education acknowledges assessment as one of the most important pedagogical approaches in improving students’ learning, and its most recent assessment policy document *Growing Success: Assessment, Evaluation, and Reporting in Ontario Schools* (Ontario MOE, 2010) reflects this paradigm shift from assessment of learning to assessment for learning. The ILE program has not been an exception to these changes; the ILE resource guide published by the Ministry (Ontario MOE, 2012) cites this *Growing Success* policy document and emphasizes that the primary purpose of assessment and evaluation should be “to improve student learning” (p. 68) as a fundamental principle when protocols for assessing and reporting on ILE student progress are designed by school boards. The ILE resource guide clearly states that instructors are responsible for using diverse assessment strategies to identify students’ learning needs, providing each student with ongoing, descriptive feedback, and reporting on each student’s progress informally and formally (Ontario MOE, 2012). Furthermore, it contains useful resources specific to assessing and reporting on student progress that teachers and administrators can draw upon (Ontario MOE, 2012).

Despite the assessment and reporting guidelines provided by the Ministry of Education, it remains unclear how well the assessment policy is enacted locally, especially in after-school or weekend classes offered under the ILE program (Lara & Volante, 2019). Considering that there is no mandated accountability requirement at the provincial level, the assessment and reporting practices in the ILE program are likely to depend on the protocols developed by each school board or individual teachers. Despite the fact the ILE program is outside of the regular K–12 curriculum, it is still necessary to investigate if individual students’ educational needs and learning development are appropriately identified and reported in order to ensure that the 2.5 hours per week in the ILE program are well spent. Therefore, the current study sought to evaluate the program operation in Korean classes with a focus on the assessment practice and provide ways to improve the quality of the program.
Methods

Research Context

The present study focused on Korean language classes offered as part of the ILE program. We decided to narrow our focus to this specific language rather than various languages taught in the program because this contextual focus would allow us to gain a deeper understanding of the program operation for a specific ethnic group and to provide more practical suggestions to improve the related classes. As such, the primary stakeholders sought in the present study were teachers of Korean classes and parents of students enrolled in one of these classes; yet, as elaborated in the discussion section, we believe that there is considerable potential for some of our findings to be generalized to many other language groups.

The Korean community in Ontario is one of the largest and most established Korean communities in North America with approximately 100,000 Koreans (Statistics Canada, 2022a). Similar to other immigrant communities, it is primarily concentrated in the Greater Toronto Area, which includes the City of Toronto and its surrounding suburbs. As of 2021, 82% of Koreans residing in Ontario identified the Korean language as their mother tongue, alone or with other languages (Statistics Canada, 2022b). Despite being considered as one of the less commonly taught languages in North America, Korean is frequently demanded and taught under the ILE program in many urban school boards. Although only 151 students were enrolled in ILE Korean classes in 1978, this number continued to increase and, in 2019, approximately two thousand K–8 students learned Korean language and culture through the ILE program in 36 school locations (Korean Education Centre in Canada, 2020). Instead of, or in addition to, the publicly funded ILE program, many Korean parents choose to enrol their children in community-based language schools—mostly operated by Korean churches. Those classes are often taught by volunteers and, as of 2019, attended by another two thousand students in 38 organizations (Korean Education Centre in Canada, 2020), complementing the Korean classes offered through the ILE program.

As the first step of evaluation planning, a logic model of the ILE program was constructed in accordance with the Kellogg Foundation’s (2004) guidelines (Figure 3.1). As mentioned earlier, a logic model is useful in articulating a program theory, or how elements of
the program are planned to contribute to the intended outcomes and long-term impact (Kellogg Foundation, 2004; Yampolskaya et al., 2004). The model is comprised of inputs needed for program operation, key program activities, program outputs, and short-term outcomes, and long-term impact (Mertens & Wilson, 2012). Among the activities and short-term outcomes listed, we focused on those pertinent to assessment (underlined in Figure 3.1) because we intended to evaluate the program implementation of student assessment given the recent paradigm shift in educational assessment and the following changes in the provincial assessment policy (Ontario MOE, 2010, 2012; see the following section for detail).

Figure 3.1

Logic Model of Ontario’s International Languages Elementary Program

| Participants |

A total of five in-service teachers of ILE Korean classes and nine parents of students enrolled in ILE Korean classes participated in the study. Ideally, this type of study would also involve students as the main participants; however, considering that the majority of students enrolled in ILE Korean classes are younger elementary students (mostly grade 3 and under),
students were not recruited as participants but were involved indirectly through class observation.

Participants were recruited using purposeful sampling (Creswell, 2012). We first approached executive members of local associations for teachers of Korean in Ontario, who passed our recruitment flyers to their colleagues and students’ parents. We also visited several ILE Korean classrooms and recruited additional parent participants in person before and after class. We purposefully diversified the participant profiles in terms of teachers’ teaching certification types and teaching experience and students’ age range in case of parents. Tables 3.1 and 3.2 summarize detailed profiles of the study participants. All teacher participants were female teachers certified in either Canada or South Korea, or both, with 6–12 years of teaching experience mainly for Korean heritage learners in Canada. Most of the parent participants were mothers of students, except one parent. The parent participants’ children were in kindergarten to grade 7. In most of the families, at least one parent (mostly the mother) was proficient in Korean. All participants were associated with school boards located in the Greater Toronto Area. This geographical concentration, although not ideal, was considered acceptable given that 72% of student enrolment are from this area (Korean Education Centre in Canada, 2020).

Table 3.1

<table>
<thead>
<tr>
<th>Code</th>
<th>Teacher certificate</th>
<th>Experience in teaching Korean</th>
<th>School board</th>
</tr>
</thead>
</table>
| T1   | • Teacher Certificate in Music (South Korea)  
     | • Certified Korean Language Teacher (South Korea)  
     | • Early Child Educator (Ontario) | 10 years             | Toronto Catholic   |
| T2   | • Teacher Certificate in History (South Korea) | 10 years             | Toronto Catholic   |
| T3   | • Ontario Certified Teacher | 6 years             | York Public        |
| T4   | • Certified Korean Language Teacher (South Korea)  
     | • Early Child Educator (Ontario) | 11 years             | Toronto Public     |
| T5   | • Teacher Certificate in Korean Language Art | 12 years             | Toronto Public, Toronto Catholic |
Table 3.2

Parent Participant Profiles

<table>
<thead>
<tr>
<th>Code</th>
<th>Relationship with student</th>
<th>Student’s grade</th>
<th>Korean-speaking parent</th>
<th>School board under which student’s Korean class is offered</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Father</td>
<td>SK, G1, G3, G5</td>
<td>Mother</td>
<td>Toronto Public</td>
</tr>
<tr>
<td>P2</td>
<td>Mother</td>
<td>JK, G3</td>
<td>None</td>
<td>Toronto Public</td>
</tr>
<tr>
<td>P3</td>
<td>Mother</td>
<td>G1</td>
<td>Mother</td>
<td>Toronto Catholic</td>
</tr>
<tr>
<td>P4</td>
<td>Mother</td>
<td>G2, G5, G7</td>
<td>Mother and Father</td>
<td>Toronto Public</td>
</tr>
<tr>
<td>P5</td>
<td>Mother</td>
<td>SK</td>
<td>Mother</td>
<td>Waterloo Public, Halton Public</td>
</tr>
<tr>
<td>P6</td>
<td>Mother</td>
<td>SK</td>
<td>Mother</td>
<td>Toronto Public</td>
</tr>
<tr>
<td>P7</td>
<td>Mother</td>
<td>SK, G1, G3</td>
<td>Mother</td>
<td>Toronto Public</td>
</tr>
<tr>
<td>P8</td>
<td>Mother</td>
<td>G1</td>
<td>Mother</td>
<td>Toronto Public</td>
</tr>
<tr>
<td>P9</td>
<td>Mother</td>
<td>G3</td>
<td>Mother</td>
<td>Toronto Public</td>
</tr>
</tbody>
</table>

Note. JK = junior kindergarten, SK = senior kindergarten, G = grade

Data Collection

Data was collected through semi-structured interviews with individual participants. Two interview protocols (one for teachers and one for parents) were designed both in English and Korean with an aim to encourage the participants to openly describe their program experience, specifically related to student assessment and reporting in the ILE program. The protocol for teachers included 20 prompting questions that elicited their understanding of the definition and purpose of assessment, assessment-related teacher training or resources provided by their school boards, types and examples of assessment they frequently use, communication with parents, challenges in assessment and reporting, and suggestions for improving assessment and reporting system. The protocol for parents contained eight prompts, which included children’s experience of Korean learning, parent’s experience of communication with teachers, their understanding of assessment and reporting practice, and further suggestions for improvement. Both protocols are included as Appendix B.
We conducted all individual interviews in person, except for two cases which were carried out via phone due to logistical issues. Each interview took approximately 60–120 minutes for teachers and 30–40 minutes for parents. In the beginning of each interview, the participant was asked to choose the language they prefer throughout the interview; all participants preferred Korean except for two parent participants who chose English. All interviews, including our explanation on the purpose of the study and participants’ oral consent to participate, were audio-recorded and transcribed in the language used during the corresponding interview.

In addition to the interviews, supplementary data were collected to triangulate and elaborate our interview-based findings. Triangulation was intended to enhance the interpretive validity of evaluative claims and further address any inconsistency and contradictions arising from multiple data sources (Greene, 2015; Jick, 1979). In this study, we observed three classes, each taught by one of the teacher participants, with a duration of 90 to 120 minutes, and documented our observations in field notes. We also collected from teacher participants the report card templates used in their school boards, and from parent participants copies of the report cards their children had received from previous years. Lastly, education policy documents made available online by the Ministry and school boards in Ontario served as another data source.

Analytical Approach

This study aimed to examine what is happening in Korean language classes in the ILE program with a focus on how students are assessed as well as why this is happening. The methodology of grounded theory (Corbin & Strauss, 2015; Glaser & Strauss, 1999) was employed to analyze the characteristics and mechanism of the program operation based on the interview data collected from teachers and parents. We believe grounded theory is the most suitable and applicable methodological approach to the current study as it goes beyond thematic analysis and allows researchers to develop an explanation of phenomena by analyzing the relationships between concepts that emerged from data (Corbin & Strauss, 2015; Kolb, 2012).

The coding process in grounded theory involves abstracting concepts at various levels (i.e., lower-level concepts, categories, and core categories) (Corbin & Strauss, 2015). Many previous studies that used grounded theory followed the three-level sequential coding process
Strauss & Corbin, 1990) wherein lower-level concepts were delineated through open coding of raw data, and then, categories are derived by linking and combining related concepts. Yet, Corbin and Strauss (2015) underscored that these different coding steps should be taken simultaneously and iteratively rather than sequentially and independently. Respecting this perspective, our coding process as well as category abstraction was conducted in a simultaneous and iterative manner using constant comparison method (Corbin & Strauss, 2015).

Paradigms, originally proposed by Strauss and Corbin (1990) as an analytic tool for grounded-theory-based research to link the lower-level concepts and categories, are often used to “tell the original main story of the research” (Corbin & Strauss, 2015, p. 157) in more conceptual terms. Paradigms have been reported to be useful in explaining phenomena and analyzing causal relationships in qualitative research if it is established in the early stage of analysis and revised throughout the analysis process (Jeon & Park, 2018; Kennedy, 2011; Kim, 2014). Thus, our initial data analysis was based on the original three-element paradigm (Corbin & Strauss, 2015) comprised of conditions, actions-interactions, and consequences; however, to better reflect the complexity of the ILE program characteristics and stakeholder interactions, we borrowed from Kennedy (2011) the modified six-element paradigm model that additionally includes individual dispositions, tensions, and mitigating factors.

Two of the three authors started the coding by individually deriving lower-level concepts from interview transcripts. We held several in-person meetings and discussed the rationales of coding and relationships between different concepts, which allowed us to collectively revise our coding scheme, conduct category abstraction, and map categories out onto the paradigm in an iterative manner. Through this process, we derived 193 lower-level concepts and 38 categories, of the latter of which the relationships were analyzed using a six-factor paradigm model. As a final step, we validated our emergent understandings using our field notes from classroom observations, report cards collected from teachers and parents, and policy documents of the Ministry and school boards, which were also used to elaborate our findings in the following section.
Findings

Figure 3.2 summarizes the findings derived from the analysis by mapping 38 categories onto a six-element paradigm model. *Contextual conditions* are fundamental factors that affect the *individual dispositions* of teachers, students, and parents. *Actions-interactions* show how each stakeholder group responds to phenomena or problems and interacts with each other based on the *contextual conditions* and *individual dispositions*. The interactions sometimes lead to *tensions* between teachers and parents. Yet, in some cases, *mitigating factors* influence the *actions-interactions* or *tensions*. All these factors contribute to leading to *consequences*. The phenomena that emerge in the *consequences* stage summarize the issues surrounding the assessment and reporting practice of Korean language classes in the ILE programs. The remainder of the section, organized by the element of the paradigm model, provides details of the findings.
Figure 3.2

Grounded Theory Paradigm Model

<table>
<thead>
<tr>
<th>Contextual Conditions</th>
<th>Individual Dispositions</th>
<th>Actions-Interactions</th>
<th>Mitigating Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Purpose of learning HL]</td>
<td>[ILE program characteristics]</td>
<td>[Teachers]</td>
<td>[Parents]</td>
</tr>
<tr>
<td>• Family communication</td>
<td>• Outside of regular curriculum</td>
<td>• OCT not necessary</td>
<td>• Challenges in teaching at home</td>
</tr>
<tr>
<td>• Identify as Korean</td>
<td>• No curriculum provided</td>
<td>• No training for multi-level classroom mgmt. or assessment</td>
<td>• Expect Korean learning at little to no cost</td>
</tr>
<tr>
<td>• Bilingual competence</td>
<td>• Minimal monitoring/intervention</td>
<td>• Negative perception of assessment</td>
<td>• Perceive class as childcare</td>
</tr>
<tr>
<td>• Job/career opportunity</td>
<td>• Limited teacher training</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[Teachers]</th>
<th>[Students]</th>
<th>[Parents]</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Written communication, grammar-focused; textbook/worksheet oriented</td>
<td>• Attendance depending on parental commitment</td>
<td>• Class formation by grade or ability level</td>
</tr>
<tr>
<td>• Test-focused assessment; no wide variety in assessment tools used</td>
<td>• Teacher and peer group dependent</td>
<td>• Teacher competence and parental involvement</td>
</tr>
<tr>
<td>• Limited time for assessment documentation</td>
<td></td>
<td>• Availability of teaching assistants</td>
</tr>
<tr>
<td>• Not active in communicating student’s progress with parents</td>
<td></td>
<td>• Support and monitoring of school board</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Students]
| • Frequent lateness/absence | • Passive attitude in comm. with teachers | | |
| • Prefer cultural activities | | |
| • Refuse non-peer-group classmates | | |

[Parents]
| • Infer learning contents thru children | | |
| • Expect learning by attending class | | |

Tensions
| Possibility of assessment that is useful | Necessity of homework assignment |

Consequences
| Low perceived importance and expectation of ILE program |
| Quality of instructions and assessment depending on individual teachers rather than the system |
| Low interest level resulting in low attendance rate |
Contextual Conditions

Two sets of contextual conditions surrounded the assessment and reporting practices in ILE Korean classes. The first pertained to the purpose of heritage language learning. Teachers and parents perceived the ILE program as a means to enhance students’ ability to communicate with their families in their first language and maintain their ethnic identity through the use of Korean language and participation in cultural activities. Strengthening familial bonds across generations was seen as a potential outcome, as improved Korean language skills were believed to foster stronger relationships with parents and grandparents. Additionally, parents enrolled their children in the ILE program based on the language-as-a-resource mindset (Ruiz, 1984), viewing bilingual individuals as having a competitive advantage in the job market, both within Canada and internationally, compared to monolingual speakers. Teachers shared this perspective held by parents.

The second set of contextual conditions was shaped by the characteristics of the ILE program itself. All teacher participants and the majority of parent participants were aware that the ILE program operated independently from the regular curriculum and did not perceive it as a comprehensive language curriculum. This inherent policy limitation seemed to have an impact on teaching practices. Interviews with teachers indicated minimal intervention from the Ministry or school boards, resulting in limited opportunities for professional development in assessment literacy (Popham, 2009) and restricted access to curricular resources. Furthermore, due to a limited number of interested students or parents in a specific geographical area, school boards faced challenges in offering classes with students of similar age groups or proficiency levels. Consequently, most students were placed in “multi-level” classes where individuals from diverse age groups ranging from K–8 and with varying levels of Korean language proficiency coexisted.

Individual Dispositions

The aforementioned contextual conditions had an impact on the dispositions and attitudes of the three primary stakeholder groups involved in the ILE program: teachers, students, and parents. In terms of teachers, it was confirmed that out of the five teacher participants, only one held an Ontario teaching certificate. The teacher participants acknowledged that, despite the highly competitive nature of the job securing process, they were able to obtain their instructor
positions without the teaching certificate as long as they met other qualifications specified by the school boards. This flexibility exists because the Ministry does not required the hiring of certified teachers for the ILE program, which is not part of the regular school curriculum, in order to streamline and facilitate the teacher recruitment process (Berryman, 1986). Teachers reported receiving minimal training on administrative matters, but none specifically focused on assessment since their employment. However, there was a noticeable difference in the perceived importance of assessment among the teacher participants, potentially influenced by their teacher training backgrounds. For instance, the teacher participant with the Ontario teaching certificate recognized assessment as a valuable tool for identifying students’ strengths and areas for improvement in language learning.

*T3: In my classroom, assessments are meant to identify students who are ready for the next lesson and to guide the direction to lead them towards. They’re assessments for learning, as well as of learning. Not merely for exam-taking, but rather to inform what needs to be taught.*

Yet, most of the other teachers shared a different perspective, viewing assessment as a means to praise students and provide external motivation:

*T2: I use assessments or tests to determine who would receive prizes at the end of the year. Assessments are necessary for students to earn rewards and prizes, which in turn can increase student attendance rates.*

Students’ dispositions were derived from the concepts emerged from the interviews with teachers and parents. It was observed that students were often enrolled in classes by their parents rather than making the choice themselves, resulting in their attendance being heavily reliant on parental commitment:

*T2: Students say, “My mom told me to go [to the Korean class].” So, as students grew older and become more independent, they tend to resist their parents’ directives. Then, the number of student enrolment in Korean classes decreases.*

Several factors were found to impact students’ attitudes towards Korean language classes. These factors included having a friendly teacher, being part of a close-knit peer group, and engaging in culturally enriching and enjoyable activities. These aspects played a significant role in either
sustaining or diminishing students’ interest in learning their heritage language within the ILE program.

Despite unanimous agreement among parent participants regarding the importance of teaching Korean to their children, there was a shared concern that their children lacked sufficient exposure to Korean at home due to the parents’ limited proficiency or time constraints. Consequently, parents expressed appreciation for the ILE program as it offered their children opportunities to learn the Korean language and culture at little to no cost. However, the affordability of the program led some parents to view it primarily as a childcare service rather than a dedicated language learning program. This perception of the “Korean class as childcare service” resulted in some parents not fully recognizing the program as a high-quality educational service.

**Actions-Interactions**

The dispositions of teachers, parents, and students significantly influence their behaviours and interactions, which also shape their responses to the contextual conditions mentioned earlier. Many teachers acknowledged the difficulties of managing multi-level classes without adequate training, leading them to heavily rely on textbooks and worksheets focusing on Korean alphabet, grammar, and vocabulary, with a stronger emphasis on written text than oral communication. This pedagogical approach was evident as teachers mentioned that they could maintain classroom order with minimal disruptions. Even teachers who valued formative assessment predominantly assessed reading and writing skills using traditional test formats, rather than engaging in ongoing observations or oral interactions with students in the classroom. Some teachers expressed challenges in accurately assessing their students due to the wide range of language skill levels present within a single class and limited time for pedagogical documentation:

*T5: I struggle to assess my students accurately. The class is multi-level, and we need specific learning goals to assess against, which don’t really exist.*

Teachers expressed concerns about the lack of appropriate assessment tools and expressed hope that school boards or the South Korean government could develop assessment tools suitable for
their needs. Insufficient assessment resources and inadequate training resulted in inconsistent assessment practices.

On the other hand, parents expressed concerns about the lack of parent-teacher communication in the ILE program. One parent participant even reported that they once did not receive report cards, which hindered their ability to track their child’s progress. Interviews with parents further revealed that some parents’ passive involvement in communications with teachers could be attributed to cultural beliefs about the parental role in their children’s education (Moon & Jung, 2018); instead of directly communicating with teachers, parents relied on their children’s accounts or inferred information from the materials in their backpacks to gauge their progress. Some parents criticized others who expected their children to improve their Korean language skills solely through class attendance, without active participating in their children’s learning process:

P3: They [other parents] view Korean class as a form of daycare, but hope that their children learn something. But, (for the kids to learn,) they need to help them learn.

The limited level of parental involvement has an impact on students’ participation in the program. Students were reported as frequently being late or absent, contrasting with their relatively active engagement in cultural activities (e.g., cooking Korean food, playing traditional games). While these hands-on cultural experiences are an important aspect of learning in the ILE program; they do not automatically lead to the language acquisition that parents expect. Attendance rates tended to decrease among older students as they became more independent and had a greater sense of agency. Once contributing factor to this decline was that they were often placed in classes with peers of different ages and varying levels of Korean language proficiency.

P2: My son is in grade 3. You don’t want to be in the same class with 4-year-olds. They [students like my child] are the ones who do the things twice as fast as the younger kids.

This heterogeneous classroom composition, which is common in the ILE program across languages taught, might have created challenges in terms of meeting the specific learning needs of older students and maintaining their interest and motivation. The manifestation of theses structural limitations in the classroom reinforces the importance of addressing the contextual conditions.
Tensions

The actions-interactions described above caused considerable tensions between teachers and parents. One significant tension was over the possibility of conducting assessments that are meaningful and conducive to learning. Teachers, faced with the challenges of classroom management, expressed skepticism about the usefulness of assessments as one teacher participant acknowledged that “assessment may be important, but, in reality, it’s of little use” (T5). They observed that assessment results, such as report cards, were not taken seriously by students or parents and, thus, did not contribute to improving student learning. Additionally, school boards only offered high-level assessment guidelines but not specific assessment tools, so the entire process of assessment, such as assessment design and administration, was all left to the discretion of individual teachers. Few teachers reported that they had conducted assessments beyond the traditional test format (e.g., dictation, vocabulary test).

In contrast, parents held the belief that teachers should “be able to make detailed and accurate assessments if they focus on the Korean language ability” (P4). They desired practical measures to gauge their children's progress and expressed the importance of clear communication regarding assessment results in more practical terms:

P1: I like to see what the assessment was, so I could see specifically, here are the words they can recognize, the grammar for the books they are able to read, and at what level they can write. For example, can they comprehend Korean newspaper? How about Korean traditional novels? Obviously, Korean grammar is there [in these texts]. So, something like that would give me a better idea of how well they can navigate the Korean language.

Teachers and parents also held divergent views regarding the importance of assigning homework. Teachers expressed frustration over the lack of parental interest and support. They had requested parents to assist their children with homework as a means to supplement the limited classroom time (2.5 hours per week at maximum). However, they received negative responses from both students and parents in most cases. As a result, teachers refrained from making similar requests to parents:
T2: When I made suggestions to parents, the parents did not like it. Many parents don’t help their children with homework. (...) Parents should understand that attending classes along can’t magically make students proficient in the language.

On the other hand, many parent participants expressed their support for teachers assigning homework. They believed that without additional learning tasks assigned as homework, they would remain unaware of their child’s progress and learning in the class:

P5: I don't know what my child is learning because not everything shows up on the report card. If I look at what they bring home in their backpack, there are only a few letters written. I can only ask, “what song did you learn today?”

These conflicting perceptions between teachers and parents regarding homework assignment can be considered as a concrete example of the lack of communication mentioned in the previous section. However, there were certain factors that mitigated the level of tension between these two groups, which are discussed below.

Mitigating Factors

While identifying the tensions described above, it was observed that the severity of tensions varied based on the presence of mitigating factors. One such factor was the grouping of students in the classroom based on age group or language ability level, which was found to positively influence the effectiveness of class instruction, as perceived by both teachers and parents. A teacher participant reported that she was perceived as a better teacher by parents in such a class:

T5: Parents perceive teachers at this school to teach better and this perception spreads through word of mouth, attracting more students to this school. In reality, these same teachers teach at other schools. Parents think that teachers are better here because each class is strictly formed based on grade, without multi-grade classes.

In classrooms with more homogenous student groups, teachers were able to develop more targeted and effective lesson plans that catered to the specific needs of the students. This, in turn, provided students with more opportunities to actively participate in class, leading to increased engagement and satisfaction. Additionally, the higher level of engagement facilitated teachers in
gathering more evidence of students’ progress and effectively assessing and documenting their growth, which could then be shared with parents. As a result, parents developed more positive perceptions of the class, leading to increased satisfaction and support.

The competence of teachers and the level of parental involvement were identified as another significant mitigating factor in terms of student assessment and communication. This factor not only influenced the level of trust established between teachers and parents but also bridged the gap between classroom instruction and home learning. Some teachers took extra steps to ensure effective communication by utilizing online messaging applications or engaging in verbal communication with parents during student drop-offs or pick-ups. They went beyond the minimum requirement of report cards, which is the only official reporting method mandated by the ILE program, and sought external channels to facilitate ongoing parent-teacher communication. Both teachers and parents found smartphone-based frequent communication particularly effective, as it provided real-time updates compared to report cards that are typically received only at the end of the semester. This proactive approach to communication fostered a stronger partnership between teachers and parents, enhancing the overall learning experience for the students.

Another important mitigating factor was the availability of additional teaching staff, such as teaching assistants. Some classes were fortunate to have one or more teaching assistants, often high school students recruited by the school boards. While there were some concerns about the additional workload placed on teachers in managing the teaching assistants, the majority of teachers and parents recognized the presence of teaching assistants as a significant factor that allowed teachers to dedicate more time to assessing and interacting with students:

*T4: When there is a wide range of student levels in the classroom, the classroom feels like a bustling flea market. It feels chaotic and overwhelming. Having a volunteer is a big help. Even a student who can’t speak Korean is incredibly helpful in organizing lessons and activities. (…) Regardless of whether I’m in the eye of the surveillance or having the helping hands, I get to have time to assess your students.*

Lastly, the level of support and monitoring from school boards greatly influenced the satisfaction of parents and teachers. Some school boards offered more comprehensive support,
including teacher training programs, parent workshops, and regular classroom supervision. In contrast, other school boards, particularly those with fewer students enrolled, provided minimal support and have limited administrative staff available. A parent who had experienced different school boards over the years, encountered situations where the program’s implementation and organization varied significantly:

P5: How the program works greatly varies across school boards. Here [at this school board]. (...) when I took my child to school on the first day, I found the whole process to be so unorganized. I couldn’t find anyone to answer my questions or tell me where the classroom was.

The four mitigating factors discussed above played a significant role in shaping the actions-interactions and severity of tensions among teachers, students, and parents in the ILE program. The effectiveness of these mitigating factors depended on the level of investment and commitment from school boards, teachers, and parents towards the development of students in this seemingly less important non-official language. Given the two contrasting sets of contextual conditions (i.e., the needs for heritage language learning vs. the structural limitations of the ILE program), stakeholders made intentional efforts to improve the quality of instruction. These efforts included implementing different placement methods, adopting new communication channels, and providing additional administrative support and staffing. When these attempts were successful, they significantly contributed to the alleviation of tensions within the ILE program. Teachers were able to better address the specific learning needs of students, parents felt more informed and engaged in their child's education, and students benefited from a more supportive and conducive learning environment.

Consequences

Despite their differing perspectives on the assessment and reporting practices of Korean language classes and the ILE program in general, both teachers and parents shared low expectations for the program’s outcomes:

T5: The Korean language school here functions more like an after-school program due to structural issues. Students are learning Korean language to some extent, but I think the classes are perceived as daycare [more than a focused language learning program].
In this country, kindergarten is optional, and so is Korean language school to a greater extent. Besides, I only pay $20 a year [for Korean language school]. (...) I don’t expect much in that regard.

These low expectations for the overall program influenced more invested parents to heavily rely on individual teachers for their children’s Korean language learning. Parents and teachers acknowledged the need for a language-specific curriculum and assessment aligned with the curricular content, which neither the Ministry nor school boards provided:

I would like to see common curriculum standards for this program. Currently, there is no course outline provided, and it is left up to the discretion of individual teachers. It is very important for these teachers to maintain a good reputation among parents. The classes are identified by the teacher’s name rather than the name of the school or school board. Parents prioritize the reputation of the teacher in their decision-making.

More importantly, the combination of low expectations of teachers and parents and the overall lower quality of the program led had serious consequences for the students:

With so many students in the class with different levels of proficiency, I guess I just have to be a lecturer or show them some Korean TV shows. But then, some kids stop coming [to the Korean class]. They can easily watch these shows at home.

With students’ learning needs left unmet, their interest in learning Korean declined over time, as evidenced by decreasing attendance rates and student enrolment among older students.

Discussion

The present study aimed to explore the assessment and reporting practices in the ILE Korean classes by examining the perspectives and experiences of teachers and parents. This examination was intended to provide thick descriptions of the program operation, its activities, and perceived merits and values. Through a grounded theory approach, we developed a paradigm model that accounts for the mechanism of program operation, including contextual factors, perceived tensions, and consequences. The six-element paradigm model reveals critical issues concerning the assessment and reporting practice in the ILE program.
Overall, our findings indicate that the most recent provincial assessment policy (Ontario MOE, 2012), which emphasizes assessment for learning, was not effectively implemented in Korean classes within the ILE program. Structural barriers, such as the program being offered outside the regular curriculum, multi-level classes, and a lack of consistent professional development opportunities and assessment resources for teachers, contributed to the poor assessment practices. These findings align with Berryman’s (1986) earlier claim that the requirement for heritage languages to be studied as continuing education programs (as opposed to part of the regular curriculum) leads to various challenges. Unfortunately, limited progress has been made in addressing these issues since Berryman’s evaluation study conducted three decades ago, underscoring the continued marginalization of heritage language education in the public sphere. The consequences extend beyond wasted tax dollars to the diminished perceived importance and interest of students in learning their heritage language. As Feuerverger (1997) pointed out, students may perceive time spent in heritage language classes as not being worthwhile.

The identified problems in this study align with previous research in the field of heritage language education, particularly in the Canadian context. Issues such as lack of program coherence and organization, lack of teaching materials, inadequate teacher training, multi-level classes, low self-esteem among teachers, and unmotivated students have been cited repeatedly in earlier studies (Cummins, 2005; Cummins & Danesi, 1990; Duff, 2008; Feueverger, 1991, 1997). These challenges stem from the nature of the ILE program and its associated regulations. Therefore, it is likely that similar tensions and issues persist across various languages taught within the ILE program, as partly documented in Fiorucci-Nichollas (1998; Italian), Mercurio (1997; Italian), Mercurio-Berrafati (2009; Italian), Bascuñán (2009; Spanish), and Mycek (2015; Polish, German, and Mandarin). For this reason, our findings, while focused on Korean classes, may have far-reaching implications on classes of other languages under the ILE program. While addressing all the complex issues would require comprehensive changes, the following actions could contribute to positive improvements.
Recommendations for the ILE Program

As a starting point, it is urgent to prioritize adequate teacher training and additional staff. Given the structural limitations of the program, it may not be feasible to form classes exclusively with students in the same age group or language proficiency level. In this context, one visible short-term solution for the Ministry and local school boards is to support teacher education and training. Teacher interviews have revealed that many teachers still perceive “assessment” and “test” as interchangeable terms. It is essential to broaden their understanding of assessment beyond traditional testing methods, as testing for rewards is just one approach among many. Non-testing assessments, as highlighted by Lynch (2001), can provide valuable individualized instructions that cater to students with diverse proficiency levels.

In-service teachers would greatly benefit from professional development opportunities focused on a wide range of assessment approaches, strategies, and, as emphasized by Mycek (2015), effective multi-level class management techniques. These training sessions should equip teachers with the skills to implement differentiated instruction, provide personalized feedback, and utilize scaffolding methods to meet the needs of students with varying abilities. Moreover, each school board should consider a systematic allocation of additional teaching staff or volunteers. This approach would enhance the student-to-teacher ratio, allowing for more individualized attention and support. By increasing the level of one-on-one interaction, students can receive more targeted guidance and assistance in their language learning journey.

As a mid-term goal, it is imperative for the Ministry to take decisive actions to implement an effective assessment and reporting system that addresses the reality of multi-level classes within the ILE program. Furthermore, it is crucial to provide teachers with adequate assessment tools to support their instructional practices. At the beginning of each school year, students should undergo assessments using tools that offer diagnostic information about their individual levels and learning needs. These assessments can be utilized periodically to monitor and track students’ progress over time. By employing such meaningful assessments, teachers can tailor their instruction to meet the specific needs of each student, while students themselves gain a clear understanding of the skills they are expected to develop, as well as the methods and standards by which their progress will be evaluated. Aitken (2016) emphasizes that parents,
students, and teachers all have the right to “get accurate information about what a student knows and can do” (p. 246).

As mentioned throughout this paper, assessment for learning has already become an essential concept in public education in Ontario (Ontario MOE, 2010). The ILE program, with its non-compulsory nature, offers some flexibility in terms of accountability. Paradoxically, this out-of-regular-curriculum status allows for greater opportunities to implement innovative formative assessment practices or assessment for learning (Birenbaum et al., 2015). Formative assessment tools should be regularly used in classroom settings, utilizing methods such as student observation and oral interaction backed up by systematic pedagogical documentation and a comprehensive utilization of assessment results beyond traditional testing (Jang et al., 2011). In parallel with the development of appropriate assessment practices, it is equally important to establish a reporting and communication system that aligns with the purpose of these assessments. The current practice of assigning letter grades on report cards falls short in providing a complete and accurate depiction of student development. Therefore, it is essential to establish a more comprehensive and meaningful reporting method that effectively communicates students' progress and offers guidance to parents regarding support strategies.

In the long term, it is necessary to consider developing of curriculum standards for Korean and other languages taught in the ILE program, specifically tailored to meet the needs of heritage learners. Currently, the responsibility for curriculum development lies with the school boards, as outlined in the program resource guide (Ontario MOE, 2012). However, it appears that this crucial duty is not being fully fulfilled. The existence of a detailed curriculum is essential for accurate and meaningful assessment of student learning, as emphasized by both teachers and parents interviewed in this study (Aitken, 2016). With a well-defined curriculum in place, educators can establish clear expectations and benchmarks for student performance. This enables the confirmation of whether each student is meeting the curriculum expectations and allows for the identification of areas where additional support may be required. Given that age and language proficiency do not always align, it is imperative for the program to incorporate differentiated curricula that can accommodate variations in language skills and student maturity.

With these differentiated curricula in place, addressing the issue of multi-level classes with a wide range of language ability levels is crucial. The current practice of grouping students
at best solely by age group poses significant challenges, as highlighted in previous studies (Mercurio, 1997; Mycek, 2015). It becomes difficult for students at different proficiency levels, such as those who can read advanced news articles and those who are new to learning the alphabet, to make meaningful learning progress in the same classroom.

The Ministry and school boards can consider implementing pilot projects at a small number of school sites, including online classes. These pilot projects should prioritize grouping students not only by age group but also by proficiency level. By creating classes that consist of students with similar language abilities, teachers can provide targeted instruction that meets the specific needs of each group. For example, the integrated extended-day model of the ILE program provides a promising opportunity. This model incorporates language classes as part of the regular school day, with the school day extended by 30 minutes to accommodate these classes (Ontario MOE, 2012). As high student enrollment is guaranteed in this model, it becomes much more feasible to place students into different classes by language skill level and age group.

The integrated extended-day model offers several advantages over after-school or weekend models. First, students have more frequent exposure to the target language (i.e., multiple days a week vs. once a week). Second, the day school’s commitment to and ownership of the ILE program is strengthened. Additionally, regular interaction between day school teachers and ILE instructors promotes collaboration and coordination in supporting students' language development (Ontario MOE, 2012). All these factors can foster a sense of importance and value for heritage language learning that is on par with other subjects in the regular curriculum. Despite the empirical evidence supporting the benefits of the integrated extended-day model in the literature (Bascuñán, 2009; Feuerverger, 1989; Fiorucci-Nicholls, 1998; Larter & Cheng, 1986; Marujo, 1999), it is unfortunate that this model was affected by recent provincial budget cuts (Teotonio, 2019). To maximize the effectiveness of the ILE program, the Ministry and school boards should prioritize the re-introduction of the integrated extended-day model and strive to increase the number of programs adopting this model. This should be done in consideration of school demographics and community demands to ensure wider accessibility and participation in the program, ultimately benefiting students’ heritage language learning and overall educational experience.
Future Research

Further research is suggested to expand our understanding of the Ontario’s ILE program (or other heritage language programs) beyond the scope of the present study. Similar to the previous studies (Feuerverger, 1997; Fiorucci-Nichollas, 1998; Mycek, 2015), our study was geographically limited to Greater Toronto Areas, so it would be valuable to evaluate the assessment and reporting practices in suburban or rural settings. In terms of languages taught, this study specifically examined Korean language classes, but more research in the context of other languages offered in the program is needed to generalize our findings. Previous research has predominantly focused on European languages (e.g., Italian, Spanish, Polish), reflecting the dominant immigrant groups in the past. Yet, given the increasing diversity of the immigrant population, it is crucial to investigate classrooms teaching languages such as Chinese, Arabic, Tamil, Punjabi, and Farsi. Since Indigenous languages can be taught through this program, it might also be of keen interest to examine how the program operates to promote Indigenous languages in Northern Ontario. We also expect future studies to include students as study participants, who are the primary beneficiaries of these education services, to gain insights into their experiences, perceptions, and needs within the program.
Acknowledgments

Study 3 was funded by the Dissertation Writing Grant from the Centre for the Study of Korea, University of Toronto. The preliminary findings of this study were presented with different focuses at the American Association of Teachers of Korean in 2018 (with Wonhee Lee as a co-author), the Canadian Association of Teachers of Korean in 2020, and the Multidisciplinary Approaches in Language Policy and Planning Conference in 2021 (both with Wonhee Lee and Dr. Eunice Eunhee Jang as co-authors). While I conceptualized the research, the second author played a significant role in data collection and analysis, and contributed substantially to the writing, particularly the Findings section. The third author provided insightful feedback and edits to an earlier version of this manuscript. I would also like to express my gratitude to the two anonymous reviewers for their valuable and constructive feedback.
General Discussion

This dissertation aimed to examine the landscape of the public education system for students with diverse linguistic and cultural backgrounds through three empirical research studies. While these studies are not mutually exclusive or collectively exhaustive, together they provide important insights into the equity-related issues and level of support experienced by LCD students in the school system. The following sections summarize the findings from the three studies, offer recommendations for policymakers, suggest future directions, and conclude the dissertation with final remarks.

Summary of Findings

Study 1 focused on the changes in LCD students’ home language environment over time and their longitudinal academic achievement in literacy, compared to their monolingual peers. The analysis of the population-level cohort data evidenced a rapid intra-generational language shift towards an English-dominant home language environment (Flores, 2015), even during the three elementary years—between grades 3 and 6—among Ontario students. However, the following multi-group latent growth curve model analysis revealed that students who had a multilingual home environment both in grades 3 and 6 demonstrated the highest literacy achievement in grade 3, as well as the steepest growth in relative performance until grade 10, compared to students with monolingual homes in either grade 3, grade 6, or both. Furthermore, students born outside of Canada and those who received language support services in grade 3 also demonstrated a higher growth in relative literacy performance between grades 3 and 10. These findings have important implications for promoting the home language of multilingual students, while recognizing the benefits of multilingualism (Bialystok, 2011; Cummins, 1993).

Study 2 investigated potential item bias against LCD students in a large-scale grade 3 reading test due to their low familiarity with mainstream Canadian culture. The study applied multi-group DIF analysis and categorized students into multiple subgroups based on their L1 and length of residence in Canada, as a proxy for cultural familiarity. The study results revealed that, among the five items hypothesized to require a high level of cultural familiarity based on expert content review, three items exhibited DIF to varying degrees depending on the student subgroup. Although the performance gap between student groups tended to decrease as students gained
more exposure to mainstream Canadian culture, a substantial difference in performance persisted for some items, even among students who had lived in Canada for five years or more. These findings highlight the need for caution in interpreting test scores for LCD students and call for greater attention to test development to ensure more valid score interpretation for this group of test takers.

Study 3 turned its attention to the support provided for LCD students’ heritage language learning as LCD students’ home language maintenance and development appear to contribute to academic success given the findings from Study 1. It aimed to evaluate Ontario’s publicly funded heritage language program with a focus on its assessment policy and practice. Using the six-element paradigm model (Corbin & Strauss, 2015; Kennedy, 2011) as an analytical framework, this qualitative case study found that the provincial assessment and reporting policy is not adequately implemented in Korean classes offered through the program. This inadequate implementation was due to several structural barriers, including the program being offered outside of the regular curriculum and a lack of consistent teacher professional learning opportunities and assessment resources. These findings reflect the continued marginalized status of heritage language education in the public sphere (Berryman, 1986; Kim et al., 2020), leading to lower perceived importance and interest in learning heritage languages among students. Practical suggestions were made to improve the current assessment and reporting practices in Korean languages classes as well as the overall quality of the ILE program.

Recommendations for Policymakers

While the theoretical, methodological, and policy implications of each study have been extensively discussed within the respective study, this section provides an opportunity to delve into the broader policy implications derived from the three studies. Based on the findings presented in this dissertation, it is important to recognize that overcoming language barriers and achieving proficiency in the societal language is just one aspect of addressing the needs of LCD students. Contrary to the prevalent deficit perspective, the findings reveal that LCD students can demonstrate comparable or even higher academic achievement than their non-LCD peers, highlighting the potential benefits of bilingualism (Study 1). However, the research also uncovers significant challenges that many LCD students face, including discrimination in large-
scale assessments due to limited cultural knowledge (Study 2), reduced motivation among students to develop their heritage language due to inadequate language programs (Study 3), and rapid intra-generational language shift towards the societal language (Study 1). These findings collectively underscore the need for substantial improvements in current policies aimed at supporting LCD students within the public education system.

First and foremost, policymakers must have a thorough understanding of the heterogeneity within the LCD student population and the fluid nature of their characteristics (García, 1991). As discussed in the introduction chapter, this student population encompasses students with an immigrant or refugee background, Indigenous students, and students who learned a language other than the socially dominant language as their L1 or who come from a cultural background that is distinct from mainstream culture (Geva & Wiener, 2015). Reflecting this heterogeneity within the LCD student population, each of the three studies in this dissertation employed unique inclusion criteria or grouping variables such as the extent of English use at home (Study 1); whether English learned as L1, and length or residence (Study 2); and enrolment in a Korean heritage language class (Study 3). It is important to note that depending on the criteria used, the target population in one study was distinct to that in another study despite possible overlaps.

In addition to the heterogeneity within the LCD student population, their characteristics are dynamic and constantly changing (Larsen-Freeman & Cameron, 2008), which adds complexity to defining target populations for specific policies. Since individual students do not exist in isolation and are influenced by contextual factors such as family, peers, school or community environments, and the larger society (Bronfenbrenner, 1979), the characteristics that are commonly used to define subgroups of LCD students (e.g., language use, proficiency, acculturation, attitude) are subject to variation. This time-varying factor was reflected in all three papers: changes in home language environment between G3 and G6 (Study 1), changes in the level of cultural familiarity or knowledge (Study 2), and changes in attitudes towards heritage language learning between kindergarten and adolescent students (Study 3). As such, it is essential that from its early stages of policy development, the target population of the said policy should be clearly defined based on its specific objectives and goals.
Second, comprehensive policies that address LCD students’ needs holistically are needed. Throughout the dissertation, a wide range of research and policy topics related to LCD students have been discussed. These include intra-generational language shift, bilingualism and literacy achievement, familiarity with mainstream culture, length of residence, item bias in large-scale assessments, language support programs in L1 and L2, teacher-led classroom assessments, the marginalized status of non-official languages, and teacher education. The extensive coverage of these topics underscores the importance of adopting a more comprehensive and holistic approach to designing policies for the LCD student population (Robinson & Clardy, 2011). Currently, policies tend to focus on addressing LCD students’ limited proficiency in the language of instruction and closing the resulting achievement gap in school through, for example, ESL programs (e.g., Ontario Ministry of Education, 2016). However, taking into account their various identities—not only as L2 learners but also as students with distinct and valuable cultures, learners of heritage languages, and cultural bridges between their families and the wider society—can lead to the development of more comprehensive policies that address their cognitive, academic, linguistic, and socio-cultural needs holistically. Furthermore, given that in Ontario, LCD students are often addressed briefly in numerous related policy frameworks, such as those pertaining to special education, parent engagement, inclusive education, assessment, and student mental health, it becomes pertinent to formulate a dedicated, comprehensive policy document that specifically caters to their diverse needs (Lara & Volante, 2019).

Finally, there needs to be a stronger emphasis on policy implementation (Hess, 2013; Viennet & Pont, 2017). Upon closer examination of the findings of the three studies, it becomes evident that the main issue lies not in the lack of relevant policies but rather in their poor implementation. Despite the adoption of multiculturalism as an official federal policy in Canada in 1971, followed by the introduction of Ontario’s heritage language program (i.e., ILE program) in 1977 to support LCD students’ heritage language maintenance and development, languages other than the two official languages continue to be perceived as having lower status among students enrolled in the program, contributing to a noticeable intra-generational language shift among Ontario’s elementary students (Study 1). At the program operation level, the evaluation of policy implementation in the ILE program revealed inadequate adherence to Ontario’s assessment policy (Ontario Ministry of Education, 2010) (Study 3). Similarly, despite the
existence of a policy aimed at identifying and addressing systematic bias in student assessments (Ontario Ministry of Education, 2014), large-scale assessments may still feature items that require cultural knowledge with which many LCD students are likely unfamiliar (Study 2). This observation aligns with Lara and Volante’s (2019) findings, highlighting that while the policies, procedures, and strategies that address needs of immigrant students have been established, what is lacking are more targeted and direct recommendations as well as effective implementation and monitoring.

Viennet and Pont (2017) have identified key determinants that can either facilitate or hinder the implementation of education policies, providing valuable insights on how to effectively put specific policies into practice. Of particular relevance to the context of this dissertation is the (in)sufficient focus on implementation during the policy development stage at the system level. Education policymakers are often criticized for formulating policies without giving enough consideration to the practical mechanisms required for their successful implementation (Viennet & Pont, 2017), resulting in a significant gap between policy intentions and actual practice (Hess, 2013). In the case of Ontario, while the provincial government establishes overall directions and guidelines, the implementation details are often left for school boards to develop, following the decentralized governance arrangement (Viennet & Pont, 2017). Unfortunately, little attention is given to monitoring and evaluating how policies are being put into practice (Hess, 2013). Therefore, it is crucial to prioritize effective implementation strategies to address the identified issues and enhance the educational experiences and outcomes for LCD students.

**Future Directions**

Throughout the dissertation, the limitations of the study and suggestions for future research were discussed within each corresponding study and may not need to be repeated. However, it is important to acknowledge a general limitation of this dissertation, namely the inability to account for some important demographic factors, and to propose additional avenues for future exploration.

As previously mentioned, the LCD student population is inherently a diverse, heterogeneous group, encompassing a wide range of linguistic and cultural backgrounds. Their
academic achievement and overall experiences within the public education system are likely to vary across various demographic factors that were not considered in this dissertation, such as L1, race, ethnicity, and socio-economic status (Volante et al., 2018). Unfortunately, due to the unavailability of comprehensive data that can link these factors to individual students in the secondary data utilized in Study 1 and Study 2, this dissertation was limited in its capacity to investigate how these demographic factors play out and contribute to the observed variations among LCD students (Cheng & Yan, 2018).

The importance of the collection and accessibility of identity-based data cannot be overemphasized. A recent report by People for Education (2023) effectively highlights the significance of identity-based data, stating:

Identity-based data collection is an integral component to advancing anti-racism because it accomplishes the following: 1) provides a profile of the population in question; 2) yields valuable insights about how different groups experience the same institutions, systems, and processes; 3) compares the outcomes of different groups that can be sorted by demographic variables such as race. This evidence is critical in the process of advocating for marginalized and underserved groups, in addition to guiding improvement plans towards more equitable systems. (p. 6)

Despite the recognized importance of identity-based data, such information is currently unavailable at the provincial level for students in Ontario. However, progress is being made in this regard, as the process of collecting voluntary student demographic data is underway, with all school boards soon to be mandated to participate in this data collection as part of the forthcoming Board Improvement and Equity Plan (Ontario Ministry of Education, 2023). This new government initiative is expected to require school boards to collect, analyze, and report disaggregated data on various demographic variables, including, but not limited to, race, ethnicity, religion, disability, sexual orientation, gender identity, and parental socio-economic status (Ontario Ministry of Education, 2017).

Once this information becomes available at the population level for students, it will significantly expand opportunities for important and fruitful research studies, particularly in the context of LCD students and other equity-deserving student groups. It is highly recommended
that future research endeavours prioritize the incorporation of these essential demographic factors within the LCD student population and take an intersectional approach (Fruja Amthor, 2017; Savaş et al., 2021). By adopting this approach, researchers can gain a more comprehensive understanding of the unique challenges and experiences encountered by different subsets of LCD students. This would enable a more nuanced analysis of the factors influencing academic achievement and other educational outcomes, thereby providing valuable insights for the development of targeted interventions and support that cater to the specific needs of diverse LCD student populations.

In addition to the suggested future studies within each of the three papers, researchers are encouraged to expand upon the studies included in this dissertation, as there is a need for further research at the intersections of these studies. For instance, conducting a DIF analysis, as explored in Study 2, prior to examining the results of Study 1 could potentially yield different outcomes. By using a purified version of a test constructed through DIF analysis and corrected score estimates (e.g., Munist, 2011), researchers can compare achievement between LCD and non-LCD students, as well as explore variations within different subgroups within the LCD student population. Furthermore, the patterns of intra-generational language shift may vary depending on the level of public support provided for heritage language development. Researchers may also find it valuable to investigate the impact of LCD students’ participation in heritage language programs and their attitudes towards these programs, particularly in relation to the changes in their home language environment and language shift (e.g., Mattheoudakis et al., 2017).

Last but not least, it is crucial to recognize the importance of directly hearing the voices of the students themselves (Mitra, 2015) concerning how LCD students are experiencing and fairing in the public education system, a component that is currently missing in the dissertation. Conducting in-depth qualitative research with a holistic approach to understanding how the issues addressed in this dissertation impact LCD students is essential. By exploring their lived experiences, researchers can examine the interplay between these issues and gain insights into impacts on LCD students at both the individual and societal levels.

By undertaking further research in these suggested directions and incorporating qualitative approaches, researchers can contribute to a comprehensive understanding of the challenges and needs of these students in a more nuanced manner. Such research can provide
meaningful insights into the experiences and impacts of various factors on LCD students, ultimately informing the development of effective policies, interventions, and support systems within the K–12 public schools.

**Concluding Remarks**

In Ontario and many other regions worldwide, the population of LCD students is constantly growing, with a greater level of diversity within this student group. This dissertation unequivocally demonstrates the pressing need for enhanced encouragement and stronger support to assist these students in preserving their language and culture, while also advocating for improved access to education on an equitable basis. Consequently, this dissertation serves as a compelling call to action, urging policymakers to prioritize a thorough understanding of the LCD student population, formulate comprehensive policies, and ensure their effective implementation. By taking these pivotal steps, we can pave the way for a more inclusive, fair, and empowering educational experience for all students within the K–12 public education system, not only in Ontario but also globally.
References


Aitken, E. N. (2016). Grading and reporting student learning. In S. Scott et al. (Eds.), *Assessment in education* (pp. 231–260). Springer. https://doi.org/10.1007/978-3-319-23398-7_10


Education Quality and Accountability Office. (2007b). *Grade 3 Assessment of Reading, Writing and Mathematics, Primary Division, 2006–2007, Student demographic data and assessment and questionnaire responses* [Dataset].


Effectiveness and School Improvement, 21(3), 337–357.  
https://doi.org/10.1080/09243453.2010.496597


https://doi.org/10.1017/S0142716411000099


https://doi.org/10.1080/01434632.2019.1618318

https://doi.org/10.1177/0013164412467033

https://doi.org/10.1111/j.1745-3984.1995.tb00466.x

https://doi.org/10.1111/j.1467-9922.2009.00527.x

https://doi.org/10.2167/beb362.0


https://doi.org/10.1007/978-3-319-74063-8_12

linguistically diverse students in special education through culturally responsive
https://www.redalyc.org/pdf/2750/275020513038.pdf


https://www.naeyc.org/sites/default/files/globally-shared/downloads/PDFs/resources/position-statements/PSDIV98.PDF


Piccardo, E. (2020). “We are all (potential) plurilinguals”: Plurilingualism as an overarching, holistic concept. *OLBI Working Papers, 10*. https://doi.org/10.18192/olbiwp.v10i0.3825


https://doi.org/10.1080/08855072.1984.10668464

https://doi.org/10.17763/haer.80.1.j94675w001329270

https://doi.org/10.1007/BF02296192

https://doi.org/10.5964/jspp.5575


https://doi.org/10.1111/j.1468-2435.1965.tb00732.x


https://doi.org/10.3102/01623737025003263

https://doi.org/10.3102/01623737016001041

Shealy, R. T., & Stout, W. F. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159–194. [https://doi.org/10.1007/BF02294572](https://doi.org/10.1007/BF02294572)


Toronto Catholic District School Board. (2019). TCDSB: International Languages Elementary (ILE) after hours programs.
https://www.tdsb.on.ca/Portals/0/Elementary/docs/Complete%202019-2020%20Brochure(3).pdf

https://www.tdsb.on.ca/portals/0/research/docs/Student_Census_-_Grades_4-6_-_System_Report.pdf

https://doi.org/10.4324/9781315092997-4


https://doi.org/10.1027/1015-5759.13.1.29


https://doi.org/10.1787/19939019


Appendix A. Contents of the Five Items Selected to Be Tested for DIF (Study 2)

Items 1, 2, and 4

My Café Menu

To Start

Nachos $2.75
Our warm nachos are served with mild, medium or hot sauce.

Rolls $1.30
Baked fresh every day, our rolls are a favourite!

Bread and Jam $1.95
Our bread is baked fresh each day. Choose from strawberry, raspberry or blueberry jam.

Carrot Sticks $1.15
Cut thin or thick, these carrots are juicy either way!

Celery Sticks $1.45
Served with cream cheese or peanut butter. They’re a real crowd-pleaser.

Soup of the Day $2.00
Please ask your server about the soup of the day.

Main Course

Macaroni $3.95
Baked in a wood-burning oven, this dish always comes out right!

Stir-Fry $4.25
Choose your own three-vegetable combination to make your plate sizzle.

Chili $4.25
There is nothing chilly about this meal!

Rice and Beans $3.95
This yummy dish is the perfect blend of mild and spicy.

Spaghetti $4.25
A café favourite! Noodles come thin or thick—your choice.

Cheese Pizza $4.25
Lots of gooey goodness on a thin crust.

On the Side

Onion Rings $2.00
Made right here, these treats are crunchy and sweet.

Veggies and Dip $2.00
Choose from ranch, sour cream or Italian dip.

French Fries $2.00
Home-cut fries are crisp on the outside, soft on the inside!

Salad $2.00
It’s your choice of garden, Greek or Caesar salad.

Dessert

Different flavours are available every day!
Ask your server for details.

Pie or Cake $2.55
Ice Cream $1.55
Fruit Salad $1.75
1. The phrase “a real crowd-pleaser” means that celery sticks
   - are a healthy snack.
   - are liked by many people.
   - come with a cheese spread.
   - can be found in most restaurants.

2. The sentence “There is nothing chilly about this meal!” means that the chili on
   the menu is
   - hot.
   - red.
   - cold.
   - brown.

4. The onion rings, veggies and dip, french fries and salad are in their own box
   on the menu because they
   - should be eaten first.
   - are healthier than the desserts.
   - cost more than the other dishes.
   - can be ordered with another dish.
The Scarecrow’s Hat

“That’s a nice hat,” said Chicken to Scarecrow.

“Yes, it is,” replied Scarecrow. “But I’d rather have a walking stick. I’ve been standing here for years now, and my arms are so tired. I’d love a walking stick to lean on. I’d swap my hat for a walking stick any day.”

Now Chicken didn’t have a walking stick, but she knew someone who did.

“That’s a nice walking stick,” said Chicken to Badger.

“Yes, it is,” replied Badger. “But I’d rather have a piece of ribbon. It gets hot and stuffy underground, so I prop my door open with my stick. But I’m always tripping over it. If I had a ribbon, I could tie the door open. I’d swap my walking stick for a ribbon any day.”

Now Chicken didn’t have a ribbon, but she knew someone who did.

“That’s a nice ribbon,” said Chicken to Crow.

“Yes, it is,” said Crow. “But I’d rather have some wool. My nest is on this high stone ledge, and it’s very hard to sit on. I’d love some warm, soft wool to line it with. I’d swap this ribbon for some wool any day.”

Now Chicken didn’t have any wool, but she knew someone who did.

“That’s a nice wool coat,” said Chicken to Sheep.
“Yes, it is,” replied Sheep. “But I’d rather have a pair of glasses. I have to keep a lookout for the wolf, and my eyes are not as good as they used to be. I really need a pair of glasses. I’d swap some of my wool for a pair of glasses any day.”

Now Chicken didn’t have a pair of glasses, but she knew someone who did.

“That’s a nice pair of glasses,” said Chicken to Donkey.

“Yes, it is,” said Donkey. “But I’d rather have a few feathers. The flies drive me crazy, buzzing around my ears. My tail isn’t quite long enough to flick them away. But if I had some long feathers tied to the end of it, I could swat them easily. I’d swap my glasses for a few long feathers any day.”

Quick as a flash, Chicken pulled out one, two, three of her longest feathers and tied them to Donkey’s tail.

Donkey was delighted and, true to his word, swapped his glasses for the feathers.

Chicken took the glasses to Sheep—who swapped them for some of his wool.

She took the wool to Crow—who swapped it for her ribbon.

She took the ribbon to Badger—who swapped it for his walking stick.

Finally, she took the walking stick to Scarecrow. With a grateful sigh of relief, he leaned his tired old arms on the stick and gladly swapped it for his battered old hat.

Chicken took the hat and filled it with fresh, sweet-smelling straw...

“That’s a nice nest,” said Duck.

“Yes, it is,” said Chicken. “And I wouldn’t swap it for anything!”
In paragraph 20, Scarecrow gives “a grateful sigh of relief” because he is

- proud.
- brave.
- worried.
- thankful. *
Hands

Black, white,
Freckled, tanned.
Every hand is different.
The doctor’s hands, a patient’s trust.
The patient’s hands, a doctor’s work.

All hands have purpose.
Gardening hands are Father’s,
Landscaping hands are Mother’s,
Painting hands are Brother’s.

Each is unusual.
Smooth, soft hands belong to Baby Anna.
Wrinkly, spotted hands belong to Grandma Ruth.
Character in every one.
My hands.

Long,
Slender fingers.
Nails.
Coloured, glittered,
Painted.
My hands.

Writing my thoughts,
Telling my dreams,
Flipping the pages of my favourite
Book.

My hands.
My past.
My future.
The speaker finds hands interesting because each pair of hands
- can use garden tools.
- is freckled and tanned.
- is different from the others. *
- can flip the pages of a book.
Appendix B. Interview Protocols (Study 3)

Teacher interview protocol

1. Tell us about your educational background and teaching experiences in Korean language education, especially for K-12 heritage language learners.

2. Which school board(s) have you worked and are you currently working for as an instructor in the International Languages program? How many classes are you currently teaching?

3. Why do you think your students come (or their parents send them) to your classes? What would be the main purposes?

4. What do you think an assessment is? What is the purpose of assessment?

5. What is the school board’s guideline on assessment and reporting in the International Languages program?

6. What kind of resources and/or professional learning opportunities does the Ministry or the school board provide for you with respect to assessment?

7. What do you think about the school board’s guideline and assessment template?

8. How do you assess your students’ learning? Do you use any type of testing? Do you collect any documentation as evidence? Do you regularly record your students’ development or areas for improvement?

9. What aspects of language learning do you usually focus when you assess your students?

10. What are the challenges for you to assess students?

11. How do you communicate the assessment results with the parents?

12. What are the challenges for you to communicate the assessment results with parents?

13. How do you think the parents and/or students perceive the current practice of assessment and reporting system of the International Languages program?

14. How can the current practice of assessment be improved? What should be added or how should the report card template be revised to better serve your students’ learning?

15. How can the Ministry of Education and/or the school board(s) support the teachers to better assess the students and communicate the results with their parents?
Parent interview protocol

1. What grade is your child in?
2. How well does your child understand/speak/read/write Korean?
3. Tell us about your child’s heritage language learning experiences (e.g., home language practices, heritage language class participation).
4. What are the main purposes of sending your child to the Korean language class? What do you expect your child to learn from this class?
5. What do you think your child’s experience is like in the Korean language class? Does s/he enjoy the class?
6. How often do you communicate with your child’s teacher about his/her learning? Who initiates this parent-teacher communication? What do you talk with the teacher about?
7. Tell us about your experience of receiving any formal assessment results from the teacher about your child’s learning (such as report cards)? Was the form of communication useful? How do you use the assessment results to help your child’s Korean language learning?
8. How would you change the template (format) of the current report card so that it provides you with more useful information about your child’s achievement and areas for improvement?
9. How do you think the current practice of assessment and reporting system can be improved?