SRDC SRSA
Social Research and Demonstration Corporation
Société de recherche sociale appliquée

READINESS TO LEARN IN MINORITY FRANCOPHONE COMMUNITIES

FIRST COHORT FINDINGS REPORT

Authors

Glenn Thompson, Ph.D.

Louise Legault, Ph.D.

Paul Lalonde

Annie Bérubé, Ph.D.

The Social Research and Demonstration Corporation (SRDC) is a non-profit research organization, created specifically to develop, field test, and rigorously evaluate new programs. SRDC's two-part mission is to help policy-makers and practitioners identify policies and programs that improve the well-being of all Canadians, with a special concern for the effects on the disadvantaged, and to raise the standards of evidence that are used in assessing these policies.

Since its establishment in December 1991, SRDC has completed over 200 projects and studies for various federal and provincial departments, municipalities, as well as other public and non-profit organizations. SRDC has offices located in Ottawa, Toronto, and Vancouver.

Readiness to Learn Sponsor

The Readiness to Learn project was funded by Employment and Social Development Canada.

Canada

The analysis report in this paper is the responsibility of the authors and does not represent the views of the Government of Canada.

Published in 2014 by the Social Research and Demonstration Corporation.

Table of Contents

Tables and Figures

Report Summary

The purpose of this report is to document the effects of the new preschool program on the children and families participating in the Readiness to Learn in Minority Francophone Communities project (henceforth referred to by its abbreviated title: Readiness to Learn project; formerly known as the Child Care Pilot Project), a demonstration project funded by Human Resources and Skills Development Canada (HRSDC). The project tests a preschool child care program[1] which pairs a child care component specifically developed to meet the needs of francophone children in minority settings with a family literacy component targeting the children's parents. The program aims to develop a child's language skills, knowledge and use of French, awareness of and identification with the francophone culture as well as foster his or her school readiness and overall development. The program is evaluated using a quasi-experimental research design with non-equivalent comparison groups. The research design comprises three treatment groups: the Program Daycare group made up of children enrolled in francophone daycare centres offering the new preschool program; the Comparison Daycare group consisting of children enrolled in francophone daycares that do not offer the new program; and the Informal Care group consisting of children who are cared for during the day at home or in an unregulated family daycare setting. The Comparison Daycare group aims to control for the influence of a formal daycare centre on child development, a treatment in itself. The Informal Care group controls for the influence of an informal care setting on child development. The project includes two participating cohorts — the first enrolled in 2007 and the second enrolled in 2008.

This report concerns data collected from May 2007 to October 2009 from the first cohort of participants. This period corresponds to a time when the children were, on average, between the ages of three and five. A mixed research methodology was used as part of the Readiness to Learn project. This approach entails the use of a range of tools, both quantitative and qualitative in nature, from several information sources, all selected based on the research objectives. The wealth of the information collected facilitates the triangulation of research findings, thus ensuring the rigour of the conclusions. Moreover, the complementarity of the data collected provides a more complete, more nuanced picture of the phenomenon under study. Analyses were thus conducted using data from, among other sources, child assessments, parent surveys, observations in daycare classes and during family literacy workshops, as well as administrative data (e.g., the record of children's presences and absences from daycare).

The impact analysis plan provides for two main series of analyses: the first aims to establish the effects of the new preschool child care program on child development, while the second examines the effects of the family literacy workshop component on parents' attitudes and behaviours. The results of all analyses suggest that the program has a positive impact on child and parent outcomes. Specifically, one observes significant benefits for all dimensions related to school readiness for children in the Program Daycare group compared with those observed in the Comparison Daycare group[2]. These benefits are significant soon after the program begins at the first post-test and half-way through the project's second year. In terms of the family literacy

[1] Officially known as *enriched child care services* in initial HRSDC documents, SRDC, in agreement with HRSDC, has, since 2007, referred to the program as the "preschool child care program."

[2] The sole exception was the domain of Physical development, for which no effect of the tested program was anticipated.

workshop component, the results of the analyses show positive impacts of the program on certain attitudinal dimensions. Thus, after the workshops, parents report significant benefits in terms of their knowledge of child development and their sense of self-efficacy. These findings suggest that part of the impacts observed among children stem from the parents, and would therefore be attributable to the family literacy workshop component of the intervention. The picture sketched by all the analyses allows us to conclude that the tested program has a modest impact on school readiness as well as on language development for children in the Program Daycare group.

Two study limitations allow us to put these results into perspective. First, the main tool for measuring a children's school readiness was an issue. In its original form, the Early Years Evaluation – Direct Assessment (EYE-DA, Willms, 2007) effectively meets the objective of identifying children with a developmental delay or those who may benefit from targeted help in learning the skills necessary for certain tasks usually observed among their peers. However, in the context of a program evaluation, this type of tool is not sensitive enough to measure subtle differences in skills development among children. Specifically, the outcomes measured by the EYE-DA do not provide information about the *developmental processes* identified in the literature as good predictors of educational outcomes in the longer term (Hirsh-Pasek, Kochanoff, Newcombe, & de Villiers, 2005). As a result, the program's effects on the children's school readiness and language skills are, in fact, underestimated in this study.

Second in the series of analyses: the relatively small sample of the Readiness to Learn project, in terms of both the sample (about 250 children) and the treatment groups (less than 100 children per treatment group) translates into low statistical power for certain extended analyses that would otherwise have been interesting to conduct. Moreover, the quasi-experimental design itself involves the risk that the results observed may be influenced by the effect of a variable not controlled for in the analyses. Analyses combining data from both cohorts are recommended in order to confirm the results observed with the first cohort and increase the statistical strength power to identify the program's more subtle effects on the outcomes of children and parents.

Note that this report is one in a series of reports by the Social Research and Demonstration Corporation (SRDC). It follows the *Readiness to Learn in Minority Francophone Communities: Reference Report* (Legault, Mák, Verstraete, & Bérubé, 2014), the final version of which was submitted to HRSDC on October 13, 2009. This first report profiled the children, families and communities participating in the Readiness to Learn project. Further, this report supplements the *Readiness to Learn in Minority Francophone Communities: Project Implementation Report* (Bérubé, Legault, Janisse, Carson, Saucier, & Lefebvre, 2014) submitted to HRSDC on May 31, 2010. Three additional reports are planned dealing respectively with: impact analyses and an implementation study of the preschool child care program comprising both cohorts of participants; the 12-month impacts following the end of the intervention; and the 24-month impacts post-intervention.

1. Presentation of the Readiness to Learn in Minority Francophone Communities project (Readiness to Learn project)

1.1. PROJECT BACKGROUND

The Readiness to Learn in Minority Francophone Communities project (Readiness to Learn project) is part of the Government of Canada's 2003–2008 Action Plan for Official Languages and continues under the 2008–2013 Roadmap for Canada's Linguistic Duality. The guiding principles that shaped the project include a desire to support minority francophone communities by providing children with a good start in life and encouraging parents to participate actively in their child's education (HRSDC, 2006). At the community level, the Readiness to Learn project is intended to be a rigorous assessment of a promising intervention whose goal is to maintain and even to renew the ethnolinguistic vitality of the minority francophone community[3].

As such, this project recognizes that the development of linguistic and identity-related dimensions is more important in a minority context than it might be otherwise. Indeed, members of the linguistic majority can take this developmental process for granted, but its true complexity is exposed within a minority-language setting. Specifically, being Francophone is conceived as an ongoing process that includes learning the French language, constructing a francophone identity and culture, and integrating into a francophone community.

The development of linguistic and identity-related dimensions is the end result of a socialization process that spans multiple settings, including the family environment, school or preschool and other socio-institutional settings (Landry & Allard, 1997). Pioneering studies on the importance of culture to child development were conducted by Vygotsky (1978). The culture in which a child is raised influences the development of his or her language skills and learning in general via the integration of social symbols to which he or she is exposed. Thus, the social environment that surrounds a child is inseparable from the construction of his or her cultural and linguistic identity, as well as his or her overall development.

In the public sphere, the reality of the minority context means that young francophone children are exposed to two different cultures while their identity is formed. According to Gilbert (2003), exposure to French in all social contexts is especially important for a child raised in a highly minority francophone setting where, by virtue of demographic weight alone, English predominates in every aspect of daily life. Empirical studies have shown that access to resources and services offered only in French offsets the strong influence of the demographic and social weight of English in the daily lives of members of the francophone community. Moreover, the presence of multiple francophone environments fosters the preservation and development of ethnolinguistic identity and language (Landry, Allard, & Deveau, 2007a). In terms of children, several authors advocate the availability of French-language child care and schooling as the main vectors of community vitality (Commission nationale des parents francophones, 2005; Landry & Allard, 1997; Gilbert, 2003). Therefore, the best-case scenario would have parents with "*ayant*

[3] See Guimond (2003) for an overview of studies on ethnolinguistic vitality in minority settings.

droit" status enrolling their children in quality French-language child care and in French-language schools. The reality, however, is completely different.

Children whose parents have "*ayant droit*" status are frequently enrolled in English-language schools. In New Brunswick, for example, 83% of children with one or two francophone parents attend French-language school. In Ontario, only 51% of children with at least one French-speaking parent attend French-language school (Corbeil, Grenier, & Lafrenière, 2007). This is the case for only 26% of young Franco-Manitobans (Statistic Canada, 2004). The findings of the 2006 Survey on the Vitality of Official-Language Minorities (SVOLM; Corbeil, et al., 2007) show that only 56% of children whose parents have "*ayant droit*" status attend French-language primary schools. This proportion declines to 44% in adolescence. These data highlight that a large proportion of parents with "*ayant droit*" status choose to enrol their children in immersion schools or in English-language schools. Parents cite a number of reasons for choosing an immersion program, including school proximity, non-availability of a French-language school, and the quality of the program or school. The main reasons for choosing a regular English-language program are, in decreasing order: school proximity (27%); the fact that English is the mother tongue or the language most familiar to the child (18%) or the parent (17%); and the quality of the school or program (17%; Corbeil, et al., 2007).

With regard to young francophone children enrolled in French-language schools, several studies show that they score lower in literacy and numeracy than children who belong to the majority language group. This disparity between the two groups is observed, among other studies, in the results of international tests such as the Program for International Student Assessment (PISA) where francophone children enrolled in French-language schools in minority language settings score lower in reading than their Canadian English-speaking counterparts (Bussière, Cartwright, Crocker, Ma, Oderkirk, & Zhang, 2001; Canadian Council on Learning, 2008). The few studies on young minority Francophones show that this gap appears at a young age. For instance, when kindergarten teachers were asked to rank children based on the Ontario government's performance scale, they judged half the students as having an overall knowledge of French that was lower than the provincial standard (Masny, 2006). These results are corroborated by a more recent study which described Franco-Manitoban children aged between four and six as scoring lower than expected on tests of French vocabulary (the Peabody Picture Vocabulary Test, or PPVT, and the communication and general knowledge scales of the Early Development Instrument, or EDI). Perhaps unsurprisingly, children raised in an environment that was primarily English-speaking obtained the lowest scores. This trend continues into primary school where francophone children are found to fare better in grade three if they grew up in a francophone family and preschool environment than if their early environment was predominantly Anglophone (Chartier, Dumaine, Daudet-Mitchell, Gosselin, & Vielfaure, 2008).

However, recent data have emerged that nuance the interpretation of these findings. The latest findings show that Francophones outside Quebec now graduate with a post-secondary degree more often than their Canadian peers (D'Amours, 2010). In Ontario, the scores of young Franco-Ontarians are improving, such that in grades three and six their scores are now comparable to, if not higher than, those of Anglophone children. These trends seem to indicate that recent actions are proving effective in enhancing the situation of Francophones in minority settings (Office de la qualité et de la responsabilité en éducation, 2009). Nevertheless, francophone communities must remain vigilant about maintaining what has been achieved in

academic performance and extending these achievements to those settings where the challenges faced by Francophones are greatest.

The source of these difficulties is likely the home, where a limited exposure to French places these children at risk of delays in the development of their French-language skills. Indeed, almost two-thirds of minority francophone children are from exogamous households (67%) and most adopt English as the language spoken at home (Landry, 2010)[4]. Only 20% of exogamous couples choose to raise their children in French from the ages of zero to four (Martel, 2001). Further, the latest data from the 2006 Census indicate that almost 39% of Francophones living outside Quebec speak English at home, although French is still used (Corbeil & Blaser, 2007).

This greater use of the English language in daily life explains in part why 62% of the francophone adults outside Quebec who chose to take a literacy test in French (rather than English) failed to achieve the level of literacy deemed necessary to function properly in society (i.e. a literacy level of less than 3 on a scale of 5; HRSDC & Statistics Canada, 2005, Table 3.24). It is safe to assume that this percentage would have been higher had all Francophones outside of Quebec completed the test in French (65% of them chose to complete the test in English despite claiming French as their mother tongue; HRSDC & Statistics Canada, 2005, p. 54). Even though the 62% estimate is likely an underestimate of the true problem, it still does not compare favourably with that of English test-takers (all of Canada, including Quebec) and French test-takers in Quebec, who respectively failed to achieve a minimal level of functional literacy 44% and 55% of the time (HRSDC & Statistics Canada, 2005, Table 3.24). In other words, the functional literacy rate of French test-takers outside of Quebec was 7% lower than that of French test-takers in Quebec and 18% lower than those who completed the test in English (including Quebec). These findings make a compelling case for the need to strengthen language acquisition among minority populations and to encourage the commitment of parents to preserving the vitality of the francophone community.

For minority francophone children, the above findings are worrisome because language skills are *crucial* to academic success, which is linked to positive professional and social outcomes. An early linguistic environment of francophone children, where an intervention that targets both the home and preschool settings could positively affect the developmental trajectory of their language skills and, by extension, foster their academic success and integration into the francophone community. Such an intervention requires, among other things, a component targeted at parents that is designed to raise their awareness of the challenges associated with living in a minority setting and of the concrete actions they can take to ensure this rich cultural heritage is passed on to their children. The Readiness to Learn project assesses the impact of such an intervention.

1.2. THE READINESS TO LEARN PROJECT

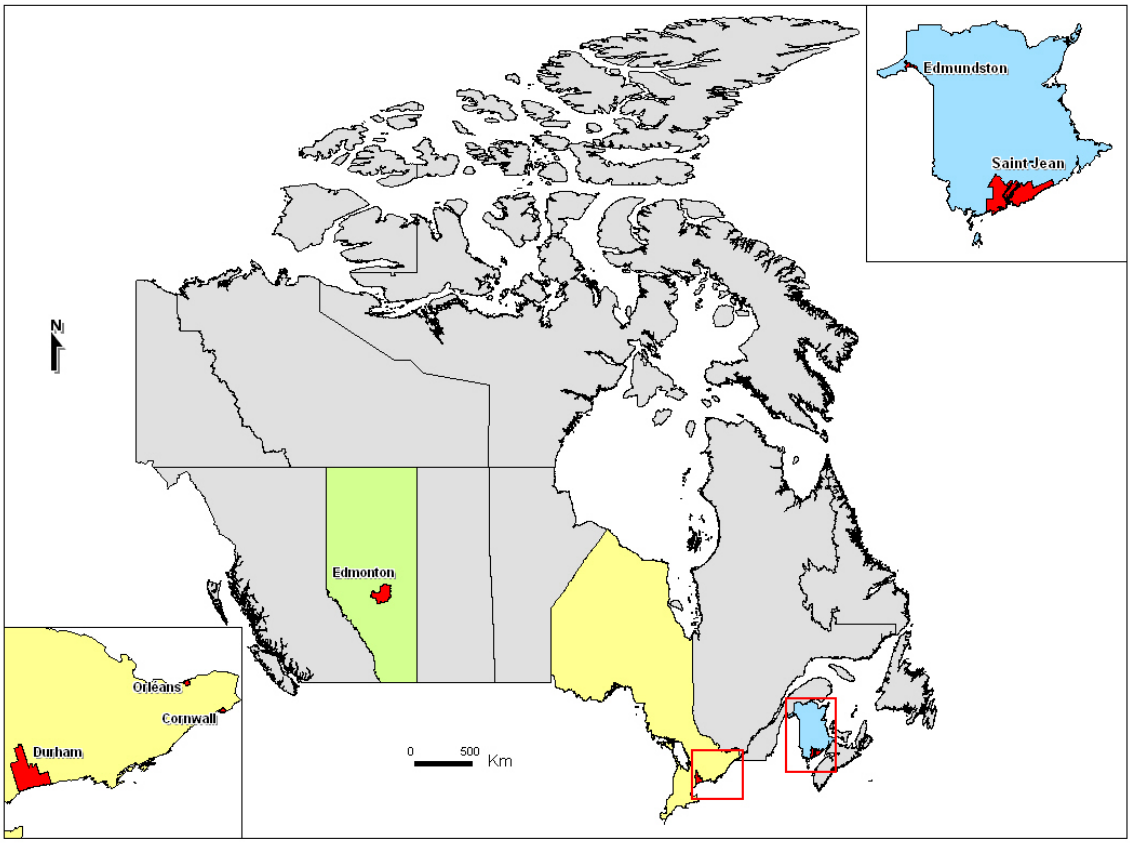
The Readiness to Learn project tests a preschool child care program and aims to identify the benefits of this preschool program for children living in minority francophone settings in comparison with other children who are not exposed to the program. The program's effect on child development is identified by comparing a group of participants who were exposed to the

[4] Landry, R. (2010). *Petite enfance et autonomie culturelle : Là ou le nombre le justifie… V.* Research report for the Commission nationale des parents francophones. Canadian Institute for Research on Linguistic Minorities. Moncton, New Brunswick.

new program (referred to as the Program Daycare group) to comparison groups comprised of study participants who were *not* exposed to the program (referred to as the Comparison Daycare group and Informal Care group). A research advisory committee, consisting of academics specializing in francophone early childhood education and representatives of francophone communities, was created to help design, implement, monitor and evaluate the pilot project. The Social Research and Demonstration Corporation manages, implements, and evaluates the program, in addition to participating actively in project development.

The program has been delivered to two cohorts of participants. The first cohort began the program in fall 2007 in six minority francophone communities (Saint John and Edmundston in New Brunswick; Orléans, Cornwall and Durham in Ontario; and Edmonton in Alberta). Program delivery to the second cohort began in fall 2008 in two communities (Orléans and Cornwall in Ontario; Figure 1.1 shows the six communities at the national level). The study takes place over a period of four years. Children are followed from age three to age seven, when they begin the second year of primary school. The span of the study allows us to chart the development of minority francophone children from preschool until their second year of primary school. The final waves of data collection are planned for the fall of 2011 and 2012 for the first and second cohort respectively.

Figure 1.1: Location of the Six Participating Communities Based on the Geographic Boundaries of Statistics Canada, 2006



In its first phase, the Readiness to Learn project aims to answer the following research question: *Does the new preschool program, which includes a daycare component and a parent–*

child workshop component, have a significant impact on children's language skills, francophone cultural identity and school readiness beyond the development that would take place in its absence, and independently of any other external factors that may come into play? Related questions are also investigated, including: *Who benefits the most from this program? Is the program cost-effective? Can the new program be replicated?* In a second phase, the Readiness to Learn project addresses the new research question: *Does the new preschool program better prepare francophone children raised in minority settings to succeed in the tasks essential to academic success that are reading and mathematics?* While the first research question focuses on the preschool period, the second research question examines the schooling period from grade one to grade two, when children are aged six and seven.

1.3. THE PROJECT'S CONTRIBUTION

This project's contribution is that it takes into account the specific environment of minority francophone communities by including a francization component, a cultural identity component and a parental awareness and involvement component. The project focuses on young francophone children and is inspired by Bronfenbrenner's ecological model (1979), one premise of which is that children are influenced by all the environmental contexts that surround them. The tested program aims to strengthen the ties between the main environments in which the child is raised — ensuring that the actions of the daycare environment support those of the family environment and vice versa — so as to optimize the child's learning in terms of both his or her general development and the development of his or her school readiness, as well as the development of French-language skills, and a francophone cultural identity.

The many benefits of programs that modify a child's environment simultaneously at daycare and at home have been established through several studies on other so-called vulnerable populations (see the literature review by Reese, Sparks, & Leyva, 2010; Engle, et al., 2007). As shown by Pelletier and Corter (2005), the effect of such programs can be enhanced if the parent and the educator use the same approaches with the child. These authors assessed a school readiness program involving children aged four, their parents and their educators. Their findings showed that a program based on early literacy activities with two components (one for parents, one for educators) fosters child literacy more effectively than similar programs that make use of only one of these components. Children whose environments changed both at home and at daycare were more advanced in terms of vocabulary development, early reading and numeracy. The results of other studies are described in Section 2.5.4.

Although the benefits of a program involving parents and educators appear to be relatively well established, it remains to be seen whether similar effects will be obtained for a program targeting both early literacy activities and the development of a child's francophone cultural identity.

1.4. THE TESTED PROGRAM

The tested preschool program pairs a child care component specifically developed to meet the needs of francophone children in minority settings with a family literacy component targeting the children's parents. The latter component is intended to stimulate the parents' active

participation in their child's development and school readiness, and in passing on French language and culture. The project itself is one of many studies on the development of preschool age children and the vitality of the French language in minority environments. The following sections describe both components of the tested program and how they are harmonized.

1.4.1. The Daycare Component

The program uses a preschool, or school readiness, approach. In this type of program, the children are led to achieve specific development objectives directly related to school readiness and intended to facilitate their academic success. This approach stands opposed to a so-called "social" learning approach in which the program has general orientations and in which each environment adopts elements based on the specificities of their community.

In addition, the daycare program advocates a play-based approach. According to this approach:

[TRANSLATION]

"[…] children are seen as independent beings who can actively contribute to shaping their learning environments. The objective is to enrich and broaden new learning opportunities, based on the educator's knowledge of child development, observations and documentation of the child's activities, as well as on the child's family and community context." (Bertrand, 2007, p. 4)

The preschool program offered in daycares is based on the *Programme des prématernelles fransaskoises* [Fransaskois junior kindergarten program] (for four-year-olds) developed by Ministry of Education of Saskatchewan, the French Education Branch, and is adapted for three-year-olds. Educators are assisted in delivering the program with basic training and regular follow-up training. They are also given French-language teaching and activity material to foster children's development of their French language skills, readiness for French-language school, knowledge of francophone culture as well as their sense of belonging to the francophone community.

1.4.2. The Family Literacy Workshop Component

Unlike some other family literacy programs, the Readiness to Learn project's family literacy workshops are intended to prepare parents to support their child's development rather than to improve their own level of literacy. The family literacy workshops program uses an adult education approach and has several objectives, the most important of which is to raise parental awareness concerning their role as their child's first educator. The program's other objectives include preparing parents to support their child's development with respect to French language, culture and identity in a family context that is unilingual, bilingual, trilingual or multicultural and raising parental awareness regarding the work of educators as well as the importance of the complementary parent–educator relationship in supporting their child's learning.

The workshops are an opportunity to discuss in French with other parents and resource people. The focus is on the pleasure of learning together. Activities are designed to support the various experiences and socialization periods of children, either with their parents, siblings, daycare friends or other community members. Parents are given advice, resources (material lent to parents and children) and training on various aspects of their child's development (e.g.,

multiple intelligences, discipline, nutrition, etc.). The participants are asked to exchange information on the resources and services available to francophone families in their community.

The family literacy workshop component consists of ten weekly parent/children workshops led by literacy practitioners. It was developed specifically for the Readiness to Learn project by Éduk and is based on the best practices identified by the Centre for Family Literacy (2002). Workshop content is inspired by recognized programs that include:

- *Grandir avec son enfant* (2002) and its adaptations, including the Nova Scotian program *J'apprends en famille*, particularly with respect to activities that concern parenting skills and children's needs;
- *Chansons, contes et comptines*, as well as *Grandir avec des livres*, given the focus on pre-reading skills and the francophone culture component;
- The English-language *Learning Together* program, which was the subject of a longitudinal study in Alberta (from 2001 to 2005) and whose effectiveness with children and families are well documented (Phillips, Hayden, & Norris, 2006); and
- The *Programme des prématernelles fransaskoises* (2001).

1.4.3. Harmonization of the Daycare and the Family Literacy Workshop Components

To maximize the Readiness to Learn project's impact on children, its two main components (the preschool program and the family literacy program) are harmonized. The two components share the objectives of optimizing children's French language skills, readiness for French-language school, development of their francophone cultural knowledge and their sense of belonging to the francophone community. The preschool program emphasizes francization and introduction to pre-literacy skills for preschoolers (including the subthemes of reading, writing and numeracy). Family literacy workshops complete the preschool program with activities that inform parents about a range of subjects, such as their role as their child's first educator (support for child development, learning stimulation), the range of community resources available in French and ways to pass on their language and culture to their child.

To link these components, the designers of the family literacy workshops worked closely with an early childhood consultant specializing in the daycare program to ensure consistency among the various aspects of the Readiness to Learn project and continuity in terms of the child's and the parent's learning. The designers factored in:

- The approach and values underlying the program implemented at the daycare;
- Francization strategies for use with children;
- The themes explored every month at daycare;
- The list of resources — games, toys and books purchased for the daycare component (the resources proposed for the family literacy workshops round out those of the daycare component); and
- The developmental stage of preschoolers.

Both components are offered at the Readiness to Learn project's program daycare, and their harmonization fosters a partnership among the people working with the child (educator, parent and other) and reinforces the child's learning in the various spheres of life, such as daycare, home and community. Creating ties between the preschool and family environments is important because studies show that doing so contributes to children's early literacy skills. Parents who get involved by talking with the educator, asking questions about their child's day and participating in daycare activities have children with larger vocabularies, better phonological awareness and better early writing skills (Arnold, Zeljo, Doctoroff, & Ortiz, 2008).

1.4.4. Course of the Program During the First Year of the Readiness to Learn project

The first year of the program ran from September 2007 to August 2008, during which time the tested program was delivered to participants in the first cohort. The daycare component program was inaugurated September 1, 2007, in the communities of Cornwall, Edmonton, Edmundston and Saint John. It was launched in October 2007 in the communities of Durham and Orléans. All the educators involved in delivering the daycare program received training in April or May 2007. Afterward, follow-up training was provided every two months in the six communities beginning in August 2007 and ending in June 2008. At each follow-up, the trainer observed classroom activities, gave educators feedback and provided targeted training. Among the themes addressed during the follow-ups were thematic planning, creative art, the children's portfolios, the development of a literacy centre, the integration of literacy and numeracy into daily activities and development of the child's drawing abilities. Note that the daycare program was not applied faithfully in summertime, which was the case in program daycares and comparison daycares delivering another program.

Only families in the Program Daycare group were asked to participate in the family literacy component. The workshops were presented in two series: a first series of four workshops in November 2007 and a second series of six workshops in January and February 2008. At the end of each workshop, families were given a family kit consisting of a book for the parent and an audiovisual resource or game. The families would return the kit at the next workshop and be given another. The kit also contained material and written instructions for a creative activity to do at home with the child. Moreover, the families were invited to visit the Resource Centre established specifically for workshops and to choose among the 300 French-language resources available to them, including books for parents. As of March 2008, once the family literacy workshops ended, the Readiness to Learn project community coordinator visited the children in the program daycare once a week so that they could borrow a book from the Resource Centre. Also once a week, the coordinator visited the program daycare at day's end so that parents could visit the centre with their child and choose among the complete range of resources including books for parents, CDs, DVDs and family games.

1.4.5. Course of the Program During the Second Year of the Readiness to Learn project

The second year of the program ran from September 2008 to August 2009. The daycare program was delivered in four of the six communities. Children in the communities of Edmonton, Edmundston and Saint John were exposed to the daycare program full-time while children in the community of Cornwall were exposed to it part-time. Children in the

communities of Durham and Orléans were not exposed to the daycare program in the second year. In these communities, parents enrolled their children in full-time junior kindergarten, which was offered free of charge.

In September or October 2008, the trainer began follow-up training again in the communities of Cornwall, Edmundston and Saint John to support educators in implementing the new daycare program. The new educators involved in delivering the daycare program received brief basic training in September or October 2008. The instructor gave the last follow-up training in June 2009. As in the first year, we observed that the tested program was not applied faithfully in summertime, either at daycares delivering the Readiness to Learn project program or at comparison daycares delivering another program. Note that implementation of the daycare program failed in the community of Edmonton[5]. The last visit by the trainer to this community was in September 2008.

No new series of family literacy workshops was given during the second year of the Readiness to Learn project. However, the activities of the Resource Centre continued in the second year in each of the six communities in the project, whether or not children were enrolled in school. The Resource Centres were also open once a week to allow children and parents to borrow a wide range of French-language resources.

In short, the entire preschool program is based on an integrative perspective and uses various approaches to influence the main vectors for shaping child development. In the sections that follow, we begin by presenting the factors that are known to influence the development of preschoolers with special regard for the particularities of raising a child in a minority francophone setting. We then define a key concept for this report: school readiness. The final section of the chapter details the report's objectives and content.

1.5. SHAPING CHILD DEVELOPMENT IN MINORITY SETTINGS

The Readiness to Learn project is derived from an ecological framework in which factors at various levels influence child development. This vision is clearly represented in Bronfenbrenner's ecological model (1979). This researcher was the first to put into words and images the entire system of influences on child development. Bronfenbrenner's model is based on three premises:

a) The child is at the centre of the model;

b) The central importance of the child's experiences (which are considered the "drivers" of development); and

c) The nature of the relationships among the child's various environments.

The Bronfenbrenner model consists of five systems (Figure 1.2):

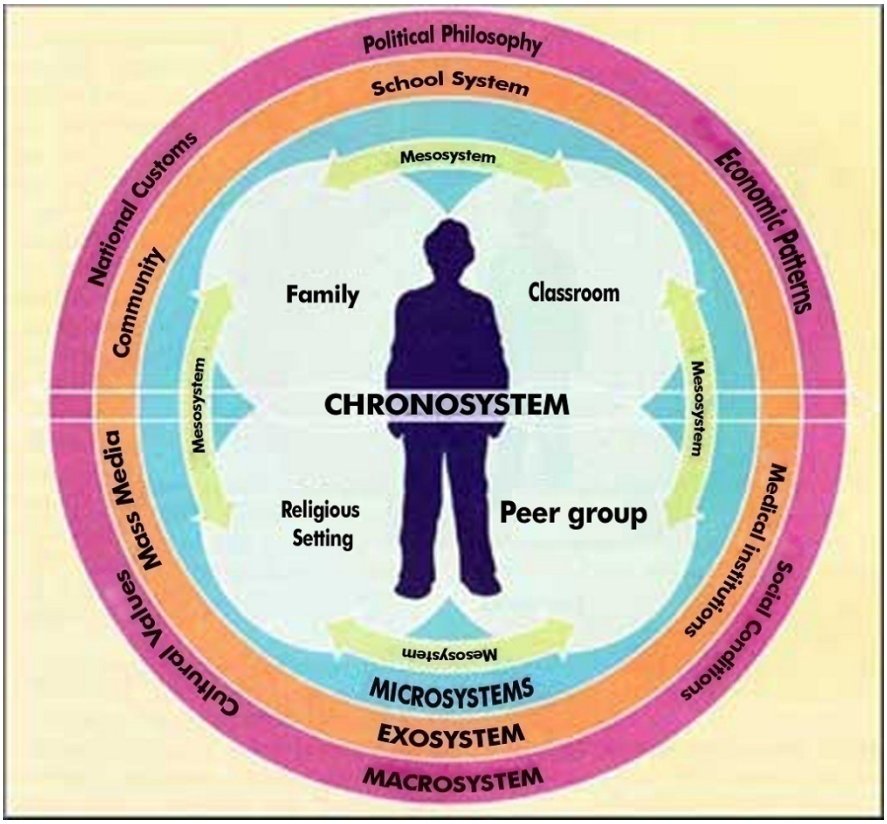
- Microsystem: Immediate environment (family, school, type of child care, peers, neighbours).

[5] An analysis of this failed implementation can be found in the *Readiness to Learn in Minority Francophone Communities: Project Implementation Report* submitted to HRSDC on May 31, 2010.

- Mesosystem: Interactions among the immediate environments (e.g., between home and school).
- Exosystem: External environment that indirectly affects the child (such as the parents' work).
- Macrosystem: Broader cultural context (western culture versus eastern culture, national economy, political culture, subculture).
- Chronosystem: Structure of events affecting the environment and transitions during existence.

In the context of the Readiness to Learn project, where the focus is on young children in minority language settings, three systems of the Bronfenbrenner model are of particular importance. First, the microsystem influences child development through the characteristics of the child's family and child care environments. In terms of family characteristics, we distinguish between the contextual variables and family processes (this division is based on the National Longitudinal Study of Children and Youth, or NLSCY; Statistics Canada, 2006). The contextual variables refer to "factual" data known to be influential in children's development (e.g., family composition). Next, the mesosystem, such as ties between the family and child care environments, also plays a role in preschooler development. Lastly, the macrosystem is among the influences of interest for the project because it consists of the community in which the child is raised and especially its language characteristics.

Figure 1.2: Bronfenbrenner's Complete Ecological Model (1979)



Source: In *UW-Extension ABC Project*, Appendix B (November 2004).

1.5.1. The Microsystem: The Family Environment (Contextual Variables)

Child development is influenced by several factors intrinsic to the child, such as his or her foetal history, birth weight and gestation period. In addition to the child's inherent characteristics, over the years research has identified a series of environmental factors that can affect the child's development. The contextual variables of the family environment are among the factors that contribute most to child development (Sanders & Morawska, 2006). These include family composition, income, the parents' level of schooling, as well as the languages spoken at home.

Family Composition

In the context of the Readiness to Learn project, we are interested in the family variables that influence child development, particularly those variables that influence a child's language trajectory. Thus, birth order is an important variable. In fact, studies show that a family's eldest child has a wider range of vocabulary on average than his or her siblings (Tamis-LeMonda & Rodriguez, 2008).

Type of family (intact two-parent, single-parent, blended family, etc.) is also a variable to consider. A study using longitudinal data from the NLSCY showed that family type (two-parent or single-parent) had a concrete influence on several aspects of child development. In fact, for children ages six to eleven, the authors concluded that [TRANSLATION] "children in two-parent families were less hyperactive, had better school results, were less anxious or depressed and their teachers rated their level of knowledge as good" (Adams & Ryan, 2000, p. iii). However, a child is not affected by family composition to the same degree as by factors associated with single-parenthood, such as the mother's stress or a lower family income, hence the importance of caution when interpreting results.

Gross Family Income

Hundreds of studies have shown an association between family poverty and a child's health, school readiness, academic success and behaviour. However, few of these studies have examined the effect of the time, duration and intensity of poverty. An insufficient family income affects child development in many ways: poor nutrition, fewer learning situations, instability in place of residence, enrolment in schools with few resources, family violence, etc. (see Brooks-Gunn & Duncan, 1997). Some studies, like that by Berger and colleagues (2005), have confirmed the hypothesis that low income influences child development outcomes through its many effects on home environment. Though the causal relationship between low income and child outcomes (intellectual and behavioural outcomes) is very clear, the interpretation of these associations remains debatable, as do the policy implications[6] Recent studies on the subject tend to show two main "paths" by which a low income can affect a child: the physical environment and the quality of parenting (Berger, Paxson, & Waldfogel, 2005)[7]. In other words, poverty will affect the purchase of material resources for the family, as well as affect the family stress level, which in turn influences child development. Several studies based on national data have shown that the

[6] For example, direct monetary transfers to families would be sufficient if there was a clear causal relationship linking income, quality of the environment and better outcomes for children (Berger, et al., 2005).

[7] The first theory was developed in economics literature (see Becker, 1993) and the second in the developmental psychology literature (e.g., Dearing, Berry, & Zaslow, 2006).

level of cognitive stimulation in the home environment (measured by the learning material and parenting as it relates to learning) accounts for 33% to 50% of the association between income and various outcomes for the child's cognitive and language development (Dearing, Berry, & Zaslow, 2006).

Studies also tend to show that the harmful effects of poverty on a child's cognitive development and academic success are greater during preschool years than at any other time (Dearing, et al., 2006). We must therefore account for this factor when assessing the program's effect within the framework of the Readiness to Learn project.

The Parents' Level of Schooling

As with family income, the parents' level of schooling is an important factor in a child's success (see, among others, Haveman & Wolfe, 1995). However, exactly how parents' education influences child development is less well studied. Klebanov and colleagues (1994) have shown that a mother's education and family income are important factors when it comes to the existence of a physical environment conducive to learning, but that only education is a major factor in terms of "warm" parenting. A series of Davis-Kean studies (2005) concludes that parental education influences child development not only through the parents' social success, but also through their beliefs and their interactions with their child. We can therefore see that level of schooling is important for child development, beyond the socioeconomic context with which it is often associated.

Mother's Age at the First Child's Birth

The mother's age at the birth of her first child is another factor with multiple effects on child development. Studies show that a mother's age when her first child is born affects the child's development and that of subsequent children, even once parenting style and family functioning is taken into account (Tremblay, Nagin, Séguin, Boivin, Pérusse, & Japel, 2004). In the case at hand, the significant fact is that very young mothers provide their child with less complex language stimulation. They use a more limited range of vocabulary and verbally stimulate their children less (Tamis-LeMonda & Rodriguez, 2008). This behaviour affects the child's language development which may be insufficient to ensure a successful start at school.

Language(s) Spoken at Home

In minority contexts, the languages most often spoken at home by the parents of the targeted child are linked to the transmission of language and the vitality of French at home (Forgues & Landry, 2006). Thus, the family's language environment is associated with the child's level of school readiness and academic success (Chartier, et al., 2008). In order to capture this important dimension, one must know which languages are usually spoken in the child's environment, particularly by his or her parents and older siblings at home and elsewhere. It is also interesting to know which languages the child usually speaks. This linguistic behaviour is a concrete indication of the languages with which the child is most comfortable. All this information paints a global picture of the child's language exposure, a primary linguistic influence.

1.5.2. The Microsystem: The Family Environment (Family Processes)

In addition to the descriptive characteristics of families, family processes undoubtedly influence child development. Among the main elements that affect child development are family functioning, parenting and family literacy activities.

Family Functioning

Beyond the mother's psychosocial difficulties, the general environment of the household is an element that has been linked to child development. Family functioning is defined as the quality of relationships in the family, in terms of the quality of communication, harmony among family members and support available within the family. Socially, poor family functioning is largely associated with aggressive development trajectories among children (Tremblay, et al., 2004). In language development, a child's vocabulary acquisition is positively related to his or her family's level of functioning, that is, the level of harmony and communication among family members (Desrosiers & Ducharme, 2006). The effects of family functioning on a child's life path are therefore manifold and non-trivial.

Parenting Style

As a child's first educators, parents play a key role in their child's development and functioning. Parenting style influences a child's social, intellectual, moral and emotional development (Bornstein & Bornstein, 2007). Parenting styles involve two aspects: *sensitivity*, which measures the degree to which the parent is attentive to the child and is able to respond to his or her needs and interests; and *control* (or *demand*), which refers to the level of supervision and discipline, as well as the degree to which the parent requires that the child be obedient and show self-control (Conseil canadien sur l'apprentissage, 2007). Studies on the subject indicate that children show better language skills and score higher on intelligence tests when their parents are more encouraging and less controlling (Sanders & Morawska, 2006). It is therefore important to measure these two aspects when following a child's language and cognitive development.

Literacy Activities

Early childhood experiences affect a child's language skills when he or she starts school (Doherty, 1997). According to Desrosiers and Ducharme (2006), children whose parents read to them on a regular basis are less likely to present a delay in vocabulary. Reading at home at an early age (before age three and a half) is even associated with improved verbal ability among children who show language problems. Likewise, parental participation in home learning activities is a predictor of children' long-term social and academic adjustment (Izzo, Weissberg, Kasprow, & Fendrich, 1999). Of particular interest for the Readiness to Learn project, exposure to oral and written French during preschool years is crucial for a child to develop strong language skills. Activities such as story time and borrowing books from the library allow the unique elements of francophone culture to be conveyed to children (Salerno in Lafrance, 1993).

1.5.3. The Microsystem: The Child Care Environment

The characteristics of the child care environment affect numerous facets of child development. Many years ago, psychologists and educators concluded, based on observations and experience, that non-parental care affects a child's cognitive and language development (National Institute of Child Health and Human Development Early Child Care Research Network

& Duncan, 2003). More specifically, the quality of a child care environment has an impact on children's cognitive and language development, school readiness and behaviour (Cleveland, Corter, Pelletier, Colley, Bertrand, Ontario Institute for Studies in Education/University of Toronto, & Jamieson, 2006). Moreover, this impact varies substantially based on certain family factors. For example, the effects of a quality child care environment are more important for children raised in an underprivileged socioeconomic context (Burchinal, Roberts, Riggins, Jr., Zeisel, Neebe, & Bryant, 2000), of which minority language status is sometimes cited as an example (Maltais, 2007).

Daycare can be a place of learning that leads to better school skills. For example, American studies show that children who attend child care centres are more likely to score higher in reading and mathematics when starting kindergarten at age five (see, among others, Howes, et al., 2008). Appropriate material, including quality games and books, an appropriate physical environment and affectionate educators that support child development, may be very beneficial for all children, particularly children living in difficult family situations. Quality of child care can take two forms: structural quality, which involves factors that can be changed through legislation (educator training, hours of operation, group size, etc.), and quality of processes, which refers to the child's experiences in his or her daycare environment (quality of activities and relationship with the educator; Burchinal, et al., 2000). Within the framework of the Readiness to Learn project, this information is factored in through observations and allows for comparison of the program daycare components with the activities of the comparison daycares. If the tested program has successfully changed the daycare environment, it will be observed here.

In a minority francophone context, the language aspect is of particular importance when studying the impact of a daycare program. In fact, a child's exposure to a French-language child care environment at the preschool level positively affects his or her success at school. Chartier and colleagues (2008) have demonstrated this statement in a study using longitudinal data on 217 children in Manitoba's francophone community[8]. Children who were exposed to French in their family environment and child care environment scored higher on the PPVT–R (which measures receptive vocabulary) and the EDI's communication and general knowledge tests compared with children who are exposed to French only at home.

1.5.4. The Mesosystem: Ties Between the Family and Child Care Environment

The mesosystem consists of the ties among the various systems that a child is in contact with. These ties contribute uniquely to child development. Several studies support the idea that the link between the school and family environments contributes to child development. Thus, children whose parents are involved in their school progress are better adjusted socially and scholastically than other children, in addition to having more positive attitudes toward school and showing higher aspirations for their future, regardless of family income and their parents' level of schooling (Connors & Epstein, 1995). The same parallel was established between parental involvement in the preschool environment and children's early literacy skills. Parents who get involved by talking with the educator, asking questions about their child's day and participating in daycare activities, have children who show a broader vocabulary, better phonological awareness and better early writing skills (Arnold, et al., 2008). The added value of a program implemented via the combination of daycare centres and family literacy workshops is that it

[8] The study is entitled the *Tots Study*, also referred to as the *1997 Manitoba Birth Cohort Study*.

emphasizes the importance of good collaboration between parents and educators and the adoption of a common set of approaches and/or approaches that are intended to be complementary (Corter & Pelletier, 2005).

1.5.5. The Macrosystem: Community Variables

An increasing number of authors recognize the important influence of community characteristics on child development (Hertzman & Kohen, 2003; Moore, 2005). The community environment fosters both a child's cognitive development as well as his or her physical and emotional health (Willms, 2007). One key factor that influences child development is the availability of French-only resources and services for families.

Resource Availability and Use

The availability and use of community resources for young families are important factors in child development. According to Connor and Brink (1999), certain categories of community resources are particularly important for child development, especially the education and health system, entertainment and culture, societal programs, special needs programs, and sports and recreation. In the context of the francophone minority, French-language early childhood services and resources are an important protective element for francophone identity and preservation of the French language (Commission nationale des parents francophones, 2005). The concept of protective element is based on the construct of a complete institutional infrastructure as proposed by Breton (1964). At the far end of the spectrum, a community with a *complete institutional infrastructure* would give its francophone population the opportunity to conduct all its daily activities in French. It follows that the presence of French-language institutions within a community fosters the creation of social networks and greater social cohesion within the community. Landry's writings (1994) revisit the concept of complete institutional infrastructure and concern four types of capital — demographic, political, economic and cultural — with enormous influence on the development, preservation (and even renewal) of a community's ethnolinguistic vitality[9]. This ethnolinguistic vitality allows minority communities to preserve their sense of belonging and pride, thus supporting the integration, not assimilation, of francophone language and culture within the majority community. One of these forms of capital, cultural capital, [TRANSLATION] "…refers to the resources and information which serve as agents of cultural transmission" (Landry, 1994, p. 18, cited in Guimond, 2003). The author advances the premise that an evaluation of this cultural capital is made possible in part by the range of educational institutions and access to cultural resources in the community. Landry, Allard and Deveau (2007b) suggest that French-language schools — and particularly the presence of a school system that allows young Francophones to study in French from the preschool to the post-secondary levels — play a crucial role in developing, preserving and growing a complete institutional infrastructure. At the preschool level, daycares, junior kindergartens, child care in family settings, extracurricular programs, resource centres and play groups serve as the gateway to French-language schooling (Gilbert, 2003).

[9] See Guimond (2003) for an overview of the studies on ethnolinguistic vitality in minority settings.

1.6. A DEFINITION OF SCHOOL READINESS

Like all school-age children, young Francophones in minority settings must prepare for the start of school. Field studies indicate that this transition is non-trivial as a strong association has been observed between a child's readiness to begin school and his or her academic success (e.g., Lemelin & Boivin, 2007). Some indicators of school readiness can be used to predict a child's aptitude for learning in school as early as age three (Thomas, 2006).

School readiness is a multidimensional concept, but one that, according to Doherty (1997, p. 25), refers mainly to the child's ability to handle the tasks that are commonly requested at school, such as staying seated and learning the material. It also encompasses the competencies that must be in place from birth to age six in order to ensure not only academic success, but also [TRANSLATION] "success in all aspects of adult life, particularly in the job market …" Five aspects of a child's school readiness are found in all research (Meisels & Atkins-Burnett, 2006):

1) Physical well-being and appropriate motor development;
2) Emotional health and a positive approach to new experiences;
3) Appropriate knowledge and social skills based on age;
4) Age-appropriate language skills; and
5) Age-appropriate general knowledge and cognitive skills.

Although factors 1, 4 and 5 are aspects generally associated with succeeding at the specific tasks necessary for academic success, points 2 and 3 also predict this success. Managing emotions, the child's general attitude in the classroom (staying seated all day or curiosity in learning, for example) and social skills are also elements essential to academic success. The National Education Goals Panel (Meisels & Atkins-Burnett, 2006) recognizes that these five factors cannot be dissociated from family, school and community, and that school readiness must be placed in relationship with these levels of influence.

1.7. THIS REPORT

This report concerns the results of analyses intended to identify the short-term impacts of the new preschool program of the daycare component on child development. It also aims to confirm the results of short-term impact analyses for the family literacy workshops component on parental attitudes and behaviours. Note that these analyses are based on data collected from May 2007 to October 2009 for the first cohort of participants in four of the six communities in question[10]. Chapter 3 details the project's methodological aspects, particularly the experimental design, the study sample, measurement tools and the study's hypotheses. Chapter 4 concerns the analytic strategy for the impact analyses. Chapter 5 explains the strategies implemented to ensure the quality and reliability of the data collected as well as the result of preliminary analyses for imputation of the missing data, the specification of error terms, and the identification of control

[10] A decision was made early on in the project to exclude the communities of Edmonton and Saint John from the impact analyses. These communities did not have French-language daycares to serve as a counterfactual. Rather than include these communities for some analyses and exclude them for others, we chose to report a coherent set of program impact estimates derived from a common sample of participants for all measures arising from the direct assessment of children. Consequently, we can more readily interpret aspects of the results that seem to depend on which comparison group is used to estimate the program effect.

variables. Chapters 6 and 7 cover, respectively, the results of the impact analyses on children's developmental outcomes for the first and second year of the project. Chapter 6 also presents the results of impact analyses for the family literacy workshops component with respect to various dimensions of the participating parents. In chapter 8, we conclude with a review of the main findings and a discussion of all the reported results.

Note that this report is part of an ongoing series. It follows the *Readiness to Learn in Minority Francophone Communities: Reference Report* (Legault, Mák, Verstraete, & Bérubé, 2014), the final version of which was submitted to HRSDC on October 13, 2009 (henceforth *Reference Report*). This first report profiled the children, families and communities participating in the Readiness to Learn project. Further, this report was prepared in parallel with and acts as a supplement to the *Readiness to Learn in Minority Francophone Communities: Project Implementation Report*[11] (Bérubé, Legault, Janisse, Carson, Saucier, & Lefebvre, 2014; henceforth *Project Implementation Report*). Three forthcoming reports are planned: one dealing with the impact analyses and an implementation study of the preschool child care program comprising the two cohorts of participants, another reporting the impact 12-months following the intervention, and a final report detailing the program's impact 24-months post-intervention.

[11] Formerly titled *Final Integrated Report: An Evaluation of the Implementation of the Child Care Pilot Project.*

2. Methodology

This chapter concerns the methodological aspects of the Readiness to Learn project. The first section describes the sampling strategy, including the target population and eligibility criteria, while the second presents the experimental design. In the third section, we enumerate the strategies that were employed to counter various threats to the study's internal validity. Section four details the measures used in the impact analyses. The final section provides a list of hypotheses tested by the impact analyses.

2.1. TARGET POPULATION

Participants were required to meet the following eligibility criteria:

- One of the child's parents had to be an "*ayant droit*" under the terms of the *Canadian Charter of Rights and Freedoms* (section 23), which meant that the child was eligible to attend French school. In addition, children of immigrants who were not "*ayant droit*", but whose first official language was French, were also eligible.
- The children (in the first cohort) had to be born in 2004 or in January 2005.
- The parents had to have the intention of registering their child at a French school.

The tested program aimed to enhance the language skills and school readiness of children who will be attending French-language schools. The first criterion served to reach this target population, which is to say children entitled to attend school in French. The second criterion limited the desired age cohort to those children eligible to register for kindergarten in September 2009 with the Ontario Ministry of Education and the New Brunswick Department of Education. The third criterion was intended to exclude from the study children whose parents had made a firm decision to send their children to English-language school. This screen was applied rarely because generally, parents of such young children have not necessarily decided what school they would attend. However, if the parents replied that English-language school was their final choice, SRDC did not follow through with the process of obtaining informed consent on the grounds that such children fall outside the target population of the new program.

2.2. EXPERIMENTAL DESIGN

The program is evaluated using a quasi-experimental research design with non-equivalent comparison groups. As with experimental studies in the field, a quasi-experimental design is intended to test — using comparison groups and pre-intervention measurements — the causal hypothesis that an intervention has a significant effect on the outcome variables in question over and above changes that would occur in the absence of the tested program and independently of other external factors that may come into play.

The methodology provides for three treatment groups: the Program Daycare group consisting of children enrolled in francophone daycares offering the new preschool program; the Comparison Daycare group consisting of children enrolled in francophone daycares that do not offer the new program; and the Informal Care group consisting of children whose child care is

provided at home or in an unregulated family daycare. The first comparison group aims to factor in the influence of a facility-based daycare on child development, which is in itself a treatment. The second comparison group was included to account for the influence of an informal child care setting on the outcomes of interest to the study, notably French language development. Children attending an English-language daycare were added to the Informal Care group for the impact analyses. Although these children are exposed to a formal child care setting, there is no exposure to French. By removing these children from the Comparison Daycare group, we maintain the language homogeneity of the Comparison Daycare group and its exposure to a French daycare program.

2.3. INTERNAL VALIDITY

In the absence of random assignment, inherent differences between the Program Daycare group and comparison groups likely exist from the outset. In such cases, special attention must be given to implementing methodological controls in the design stage of a study to eliminate *plausible* alternative explanations for results. Control techniques, often statistical ones, that involve adjustments after data have been collected are employed as well, but they are less advantageous.

Within the framework of the Readiness to Learn project, the conditions necessary for internal validity are:

- Sample size;
- A sampling strategy that ensures an equitable distribution between the groups;
- Pre-intervention measurements applied to the outcomes and associated factors;
- A verification of how attrition affects group distribution; and
- Safeguards against situations that may lead to contamination of the comparison groups.

2.3.1. Sample Size

The internal validity of the impact study with three groups requires a minimum sample of 156 children distributed evenly among the Program Daycare group (n = 52), the Comparison Daycare group (n = 52) and the Informal Care group (n = 52). This number of children per group produces the statistical power necessary to detect a moderate impact, with a high level of confidence that the real population value of the impact estimate falls within a specified range (i.e., 19 times out of 20, we would obtain the same results with other samples).

2.3.2. Targeted Sampling Strategy

A targeted sampling strategy was selected to create comparison groups that are highly similar to the Program Daycare group so as to neutralize as much as possible the influence of unmeasured factors on the outcomes in question. Previous studies have shown that certain characteristics, like family income and the parents' level of schooling, partly account for a child's developmental trajectory. It follows that if the Program Daycare group diverges considerably from the comparison group with respect to one of these characteristics, any

differences observed between the two groups on outcome measures may well be explained by these initial differences rather than by the program. Accordingly, we made a special effort to target participants for the comparison groups that presented a socio-demographic profile similar to that of the Program Daycare group (e.g., the socioeconomic level) and living in the same area, thus ensuring that they have access to the same French-language resources and services as the Program Daycare group[12]. A first control for group composition was conducted when families enrolled in the project, notably in terms of the location of the home for potential members of the comparison groups. A second control was the use of pre-intervention measurements, which are discussed in the next section.

2.3.3. Pre-Intervention Measures

The causal inference arising from a quasi-experimental methodology is facilitated by the use of pre-intervention measurements taken with respect to the outcomes and associated factors. In the case of the Readiness to Learn project, the outcome in question is the children's school readiness measured using the Early Years Evaluation – Direct Assessment (EYE-DA) and its subscales for expressive and receptive vocabulary. Given that the results presented in this report are based exclusively on the French version of the scale, it will henceforth be referred to exclusively by its French acronym, the ÉPE–AD (from the French *Évaluation de la petite enfance – Appréciation directe*).

Pre-intervention measurements provide a baseline against which all subsequent measurements can be compared, allowing for a clearer understanding of the program's impact on child development. These measures also enable comparisons of the treatment groups prior to intervention that reveal whether the groups were following similar developmental trajectories before the start of the program. This information is useful because program impact estimates would be biased if one treatment group was more developmentally advanced than the others at the outset.

Pre-intervention measurements of sociodemographic and socioeconomic variables also help establish whether children in the three experimental groups experienced similar conditions before the intervention. The variables selected for pre-intervention measurement were shown by previous studies to have a significant impact on children's school readiness. Statistical control of these variables in the analyses will allow us to adjust for remaining differences among the treatment groups, thereby teasing apart the impact of the program from the effect of these variables.

The initial profile of the children, families and communities participating in the Readiness to Learn project was given in the *Reference Report* (Legault et al., 2014). This report also contains the result of analyses establishing the homogeneity among the treatment groups before the intervention for the six-community sample. These analyses are reported a second time here, this time based only on the four-community sample that was used for the impact analyses. The list of identified variables is found in Section 5.5 of this report, while the results of the comparative analyses are detailed in Section 5.4 of this report.

[12] To learn more, consult the *Revised Work Plan and Methodology Report* submitted to HRSDC on March 30, 2007.

2.3.4. Attrition

Loss of participants in the course of the study entails a threat to the external validity of the study (i.e., whether the sample is representative of the population) as well as to its internal validity (e.g., by altering the composition of the treatment groups). Participants can withdraw from a study for several reasons, including a move or a simple lack of interest. Whatever the reason for withdrawal, it must be determined whether the groups of participants who withdrew from the project differ from the group of participants who remain on socio-demographic characteristics. Differences observed between both groups may signal a sub-group of participants with particular characteristics (a threat to external validity). Moreover, a significant change in the composition of the treatment groups could potentially compromise the validity of impact estimates (i.e., a threat to internal validity). Withdrawal from the project was systematically monitored from the outset for the purpose of assessing the circumstances of withdrawal for threats to the validity of the study.

2.3.5. Preventing Contamination of the Comparison Groups

Contamination happens when changes in the scores obtained for the outcomes of the comparison groups are attributable to the application of the conditions and modalities of the program being tested. In other words, contamination happens when members of the comparison groups are exposed to components of the daycare program and/or those of the family literacy workshop program.

Contamination of the Comparison Daycare group is more likely in communities where more than one francophone daycare in the region participates in the project. In such cases, information can circulate freely between the program daycare and comparison daycare. Several strategies were implemented to minimize opportunities for such information transfer (a list is provided below). In addition, child care providers were educated regarding situations conducive to contamination of the comparison groups and to the importance for the study of avoiding or minimizing these opportunities. This awareness-raising effort was particularly focused in communities where several francophone daycares operate under the aegis of a single provider. These discussions led to close collaboration between the child care providers and SRDC.

The following strategies were implemented to minimize opportunities for contamination.

For the daycare component:

- Basic training and follow-up training were given only to the educators and assistant educators involved in delivering the daycare program.
- The movement of educators trained for the daycare program to a comparison daycare was limited. In fact, such movement happened only in one community. In one case, the educator was assigned to a group of children younger than the children of the Readiness to Learn project. In the other case, assistant educators were assigned to daycares that were not participating in the Readiness to Learn project. Any staff members leaving the daycare program were educated regarding the importance of not sharing the knowledge and methods with other educators and were required to return the training manual to the coordinator. Finally, the coordinator monitored the situation closely at the comparison daycare.

- The classes of the Comparison Daycare group were located in a building separate from that of classes for the Program Daycare group.
- The new preschool program's resources and material had to stay at the program daycares until three months after the end of the program being tested.

For the family literacy workshop component:

- Basic training and follow-up training were given only to the workshop leaders involved in delivering the workshops.
- After the family literacy workshop component was delivered, workshop practitioners were required to return the training manual to the coordinator. They were also educated regarding the importance of not sharing the knowledge and methods with other people.
- Child care providers and school boards were prohibited from delivering the family literacy workshop program or a similar program in the community in question.
- Resources and material from the Resource Centre had to remain with the program daycare until three months after the end of the program.

2.4. SAMPLE BY COMMUNITY AND TREATMENT GROUP

The project participants who were part of the impact analyses were from the minority francophone communities of Cornwall, Orleans and Durham in Ontario, and Edmundston in New Brunswick[13]. Participant recruitment began May 1, 2007, and ended October 31, 2007.

Total sample: At the time of enrolment, the project involved 254 children from 250 families. As observed in Table 3.1, the communities of Edmundston and Cornwall had the highest proportions of participants (33.5% and 28.3%, respectively), followed by the communities of Orleans (21.7%) and Durham (16.5%). The sample consisted of 244 children at the end of the project's first year, including 116 boys and 128 girls. The sample size dropped slightly to 240 children at the end of the project's second year, including 114 boys and 126 girls.

The average age of children at enrolment was 3.0 years. At this point, the Readiness to Learn project had a slightly higher number of girls (52.5%) than boys (47.5%). The mother tongue of the children in the sample (according to the person most knowledgeable of the child) was mostly French (72.4%) followed by English (16.5%).

According to the baseline survey, the mother's age when the child in question was born was, on average, 27 years. Moreover, three-quarters of these mothers had at least a college diploma, and half had a university degree. Household size was four members on average. 9.4% of families were headed by single parents. Over half of the participating families (60.7%) had an annual income of over $70,000; median annual income was between $80,000 and $99,999.

In terms of the linguistic profile of participating families, about two-thirds of the mothers (63.1%) and fathers (60.9%) spoke only French to their child. Most children were from

[13] The lack of francophone daycares for use as comparison daycares made it impossible to properly measure the program's impact in the communities of Edmonton and Saint John.

endogamous francophone households (58%) while one-quarter were from exogamous households (23.5%)[14].

Table 2.1: Breakdown of Participants since Enrolment by Community

	Enrolment	Year 1	Year 2
Community			
Cornwall	72 (28.3%)	69 (28.3%)	68 (28.3%)
Durham	42 (16.5%)	36 (14.8%)	35 (14.6%)
Edmundston	85 (33.5%)	84 (34.4%)	83 (34.6%)
Orléans	55 (21.7%)	55 (22.5%)	54 (22.5%)
Total	**254 (100%)**	**244 (100%)**	**240 (100%)**

Sample by treatment group: In October 2007, the project involved 77 children enrolled in the Program Daycare group (G1), 99 children in the Comparison Daycare group (G2) and 78 children in the Informal Care group (G3). An examination of Table 3.2 reveals an increase in G2 along with a reduction in G1 and G3 in February 2009. These changes in the composition of the treatment groups are due to school registration of young Ontarians in September 2008, that is, at the start of the project's second year. Moreover, these fluctuations were greater in Cornwall where many children from G1 and G3 were registered at a school that housed a comparison daycare[15]. Analyses of outcomes in the second year were designed to accommodate these changes in the composition of the treatment groups.

Table 2.2: Breakdown of Participants by Treatment Group

Treatment Group	Oct. 2007	Feb. 2008	June 2008	Oct. 2008	Feb. 2009	June 2009	Oct. 2009
G1	77 (30.3%)	74 (29.7%)	70 (28,2%)	70 (28.7%)	66 (27,2%)	63 (26.3%)	63 (26.3%)
G2	99 (39.0%)	95 (38,2%)	96 (38.7%)	95 (38.9%)	112 (46.1%)	113 (47.1%)	113 (47.1%)
G3	78 (30.7%)	80 (32.1%)	82 (33.1%)	79 (32.4%)	65 (26.7%)	64 (26.3%)	64 (26.7%)
Total	**254 (100%)**	**249 (100%)**	**248 (100%)**	**244 (100%)**	**243 (100%)**	**240 (100%)**	**240 (100%)**

2.4.1. Retention Rate

The Readiness to Learn project has an excellent retention rate, with only 14 withdrawals (5.5%) since the project began in 2007. Family moves outside the community account for the

[14] Households were categorized based on the combination of the mother's first official language spoken (FOLS) and the father's FOLS.

[15] Although children were registered for school in September 2008, changes in the composition of the treatment groups do not take effect immediately in the context of the impact analyses due to the attribution rules. These are detailed in Section 4.1.2.

main reason children were withdrawn from the project. Note that some children moved to other participating communities, and therefore continued to be followed. Thereafter, these children were considered participants in the receiving community. Table 2.3 breaks down the reasons for withdrawal from the project.

Table 2.3: Reasons for Withdrawing Child from the Readiness to Learn project

Reason for Withdrawal	Year 1	Year 2	Total
Move outside a participating community	4	2	**6**
Child changes to an Anglophone daycare /not enough English at the program daycare	3	0	**3**
Lack of family availability	1	2	**3**
Bothered by the questions in the baseline survey	1	0	**1**
Contact lost with the family	1	0	**1**
Total withdrawals	**10**	**4**	**14**

2.5. MEASURES

This section presents the plan for data collection from parents and children, and for the family literacy workshops. It continues with a description of the quantitative measurement tools used since the project began (summer 2007) until the end of the second year (October 2009).

2.5.1. Data Collection Plan

Data were collected from parents and children on a quarterly basis. Baseline data was collected from May to October 2007. The next collections took place in February, June and October 2008, as well as in February, June and October 2009. This last data collection was conducted post-program, that is, when all children were enrolled in school full-time.

From Parents

In total, eight surveys of parents were conducted since the start of the Readiness to Learn project. The baseline survey was carried out when the child enrolled for the project, that is, from May 1 to October 31, 2007. The survey lasted about an hour and was conducted by the community coordinator one-on-one with the person most knowledgeable (PMK) of the child. The first follow-up survey was administered only to parents who enrolled in the project before September 1, 2007. The purpose of this survey was to update the information regarding the type and frequency of literacy activities and activities outside the daycare done with the child in fall 2007[16]. There are relatively few respondents for this second survey because this update was not necessary for parents who answered the baseline survey in the fall. The seven surveys that followed were short, taking from 10 to 30 minutes to complete, and were carried out over the

[16] For the impact analyses, data from the first follow-up survey replaced data from the baseline survey where they existed. These two surveys are therefore considered a single survey within the framework of the analyses.

telephone by the coordinators or an SRDC member. The average rate of response for the follow-up surveys was 97.9% (see Table 2.4).

Table 2.4: Survey Response Rate for Parents

	Baseline Survey	***Follow-up Surveys***						
	May to Oct. 2007	*Nov. 2007**	*Feb. 2008*	*June 2008*	*Oct. 2008*	*Feb. 2009*	*June 2009*	*Oct. 2009*
Surveys administered	254	175	241	244	238	240	237	235
Absent	0	8	8	5	6	3	3	5
Withdrawals	0	5	5	5	10	11	14	14
Response rate	100.0%	95.6%	96.8%	98.0%	97.5%	98.8%	98.8%	97.9%

Note: * These data exclude participants enrolled after September 1, 2007.

For the Family Literacy Workshops

An assessment of the impact of the tested program on parental attitudes and behaviours was planned because of the family literacy workshop component targeting the parents of the Program Daycare group. Two surveys were developed in order to gather the information before and after the workshops were delivered. The pre-intervention survey was administered only to parents assigned to the Program Daycare group. It was conducted by telephone along with the first follow-up survey in November 2007. The pre-intervention survey measured expectations, opinions and certain behaviours of parents in the Program Daycare group at the beginning of the family literacy workshop intervention. A post-workshop survey was conducted by telephone three weeks after the last workshop in order to measure the evolution of these outcomes. This survey was also conducted by telephone.

A third anonymous survey was administered during last workshop to obtain the opinion of parents on the logistical aspects of the workshops, particularly with regard to the physical environment, the quality of discussion, and the quality of workshop delivery. All these data served to assess the quality of implementation for the family literacy workshop component, the results of which are detailed in the *Project Implementation Report* (Bérubé et al., 2014). The response rate was excellent for the pre- and post-workshop surveys, but relatively low for the logistics survey due to low attendance for this workshop (see Table 2.5).

Table 2.5: Survey Response Rates for the Family Literacy Workshops

	Pre-intervention	***Logistics***	***Post-intervention***
	Nov. 2007	*Workshop 10*	*March 2008*
Surveys administered	107	69	105
Absent	6	40	4
Response rate	94.7%	63.3%	96.3%

From Children

Beginning in October 2007, every parent survey was paired with a wave of direct child assessments. Children were evaluated every four months for the first two years of the Readiness to Learn project with an average response rate of 98.6% (see Table 2.6). In total, seven child assessments were conducted.

Table 2.6: Response Rates for Child Assessments

	Assessment Periods						
	Oct. 2007	*Feb. 2008*	*June 2008*	*Oct. 2008*	*Feb. 2009*	*June 2009*	*Oct. 2009*
Children assessed	245	247	243	240	241	237	237
Absent	4	2	6	4	2	3	3
Withdrawals	5	5	5	10	11	14	14
Response rate	98.4%	99,2%	97.6%	98.4%	99,2%	98.8%	98.8%

2.5.2. Child Measures

Over the course of the project, several tools were used to monitor child development along key dimensions of school readiness. Appendix A contains the schedule and tools for each assessment.

Early Years Evaluation — Direct Assessment (Willms, 2007)

The ÉPE–AD measures four domains of school readiness and a fifth domain designed specifically for the Readiness to Learn project:

(Domain A) Awareness of Self and the Environment;

(Domain B) Cognitive Skills;

(Domain C) Language and Communication;

(Domain D) Physical and Motor Development; and

(Domain E) Awareness and Engagement in Francophone Culture.

Each domain entails a certain number of questions presented in ascending order of difficulty. Domain A, Awareness of Self and the Environment, consists of 16 questions that determine the degree to which the child may recognize and identify the elements in his or her environment. For example, the child is asked to name colours, body parts, certain occupations and their role, as well as his or her date of birth.

Domain B, Cognitive Skills, consists of 17 questions that measure various logico-mathematical aspects. For example, the child is asked to count, form groups using various objects, compare different shapes and distinguish shape sizes. The child is also asked to name a few letters in the alphabet, identify their sound, and read eight words.

Domain C, Language and Communication, involves 14 items measuring the child's ability to communicate and understand. The evaluator asks the child to point to the picture representing a word that she says, repeat seven-syllable sentences, answer questions with complete sentences, use pictures to tell a story and show the meaning of four action words. This is the only domain administered to all children in French since it is part of the decision tree used to decide the language of assessment[17].

Domain D, Physical and Motor Development, consists of 16 items measuring the child's ability to accomplish certain activities that tap either fine or gross motor skills. Fine motor skills encompass several early writing elements, such as being able to trace letters, draw a straight line or colour inside the lines while holding a pencil correctly. Gross motor skills concern the child's ability to move his or her body, for example, by jumping over an object, balancing on one foot or hopping on one foot three times in a row. Note that a Canadian study based on NLSCY data shows that after age three, there is little difference among the children with normal development in this domain. These findings were confirmed by the assessment of February 2009, which is why this domain was not measured as part of the assessment conducted in June and October 2009.

Items are scored on a five-point scale for domains A, B and C and on a four-point scale for domain D. The first three domains require the use of test charts or objects that the child must point to or manipulate. While the test is being administered, a child may reach a "plateau" in a given domain when the questions become too difficult. In fact, assessment of a domain stops when a child scores "0" or "1" on three consecutive items, at which point the evaluator moves on to the next domain. The test takes about 45 minutes to an hour.

Until February 2009, children could be assessed in English or in French depending on their scores for domain E and domain C. Domain E, which was designed specifically for the Readiness to Learn project and measures the child's awareness and engagement in francophone culture, was used more to establish a friendly rapport with the child and determine the test language[18]. Its use was discontinued in February 2009 because it was found to be unreliable[19].

The ÉPE–AD presents good psychometric properties. The result of factorial analyses for this scale provided empirical support for the theoretical unidimensionality of every domain in the

[17] Appendix B describes the process for administering the ÉPE–AD as well as the decision tree for determining testing language.

[18] Domain E contains questions on the children's preferences in terms of books, television shows and songs, the language of these resources (English or French), as well as the languages spoken with their parents and friends.

[19] The reasons for this decision are given in the *Reference Report* (Legault et al., 2014).

French version of the test[20]. The internal consistency of each dimension ranged from acceptable to very good, with Cronbach alpha estimates varying from 0.60 to 0.94 for the French version of the test and from 0.79 to 0.95 for the English version (see Table 2.7).

Table 2.7: Cronbach Alpha for ÉPE–AD Domains by Assessment Period

	Cronbach Alpha [α (*n*)]						
	Oct. 2007	***Feb. 2008***	***June 2008***	***Oct. 2008***	***Feb. 2009***	***June 2009***	***Oct. 2009***
Domains administered in French							
A	**0.91** (*213*)	**0.91** (*215*)	**0.91** (*215*)	**0.90** (*220*)	**0.90** (*232*)	**0.89** (*237*)	**0.90** (*236*)
B	**0.85** (*213*)	**0.87** (215)	**0.86** (*215*)	**0.83** (*220*)	**0.86** (*233*)	**0.75** (*237*)	**0.78** (*237*)
C	**0.91** (*245*)	**0.93** (*247*)	**0.93** (*243*)	**0.93** (*240*)	**0.91** (*240*)	**0.63** (*237*)	**0.60** (*237*)
D	**0.89** (*212*)	**0.92** (*215*)	**0.94** (*215*)	**0.93** (*220*)	**0.93** (*232*)	n/a	n/a
Domains administered in English							
A	**0.87** (*30*)	**0.91** (*32*)	**0.90** (*28*)	**0.91** (*20*)	**0.90** (*6*)	n/a	n/a
B	**0.79** (*30*)	**0.87** (*32*)	**0.87** (*28*)	**0.90** (*20*)	**0.88** (*7*)	n/a	n/a
C	**0.89** (*43*)	**0.87** (*40*)	**0.90** (*35*)	**0.88** (*27*)	**0.95** (*8*)	n/a	n/a
D	**0.86** (*30*)	**0.94** (*31*)	**0.91** (*28*)	**0.95** (*20*)	**0.96** (*7*)	n/a	n/a

Épreuve de Dénomination de Gardner (Ska, 1995)

The *Épreuve de dénomination de Gardner* (1979) measures a child's expressive vocabulary. The tool used is the standardized French translation that has been validated with children from kindergarten to grade two of the Expressive One-Word Picture Vocabulary Test (EOWPVT). Among other things, the instructions and the order of difficulty for the items were adjusted for a francophone population. The test consists of asking the child to name a series of pictures. The test is stopped after six consecutive errors and takes between 10 and 15 minutes to administer. The value of this test is that it is highly sensitive to differences in French-language proficiency. The total score on the test is the number of pictures named correctly by the child in French. For informational purposes, some predetermined types of incorrect answers were noted, such as English words and regionalisms. This additional information did not affect the score obtained by the child.

[20] A factorial analysis could not be conducted for the English version of the test because too few individuals answered in this language (63 cases for domain C and 44 for the other domains). At least 160–170 cases per domain would have been desirable to ensure the statistical validity of calculations (Tabachnick & Fidell, 2006).

Échelle de Vocabulaire en Images Peabody — Révisé (ÉVIP–R; Dunn, Thériault-Whalen, & Dunn, 1993)

The ÉVIP–R is the validated French translation of the Peabody Picture Vocabulary Test – Revised (PPVT–R), which measures a child's receptive vocabulary. The test resembles a game: the child hears a word spoken out loud and must identify the correct picture from among four alternatives. The test stops once a child makes six errors in eight tries.

According to test rules, the point at which the child begins the test is determined based on his or her age and performance. For the purposes of the Readiness to Learn project, a common starting point for all children (i.e., the 15th question, corresponding to the starting point for children at age three and a half) was established regardless of their age or performance level. Since the goal of this project is not to compare the children's performance to a "norm",[21] the use of a common starting point very early in the test allowed us to capture the receptive vocabulary level of children from exogamous families.

The ÉVIP–R score used in the impact analyses is the total number of correct answers given by the child beginning with the 15th question. This method of calculating the ÉVIP–R score differs slightly from that recommended in the test manual (Dunn, et al., 1993, pp. 13–15). We observe a correlation greater than .99 between SRDC's method and that proposed in the manual. This means there is no loss of information. The SRDC method for calculating children's ÉVIP–R score has several advantages: in particular, it a) maximizes the variance and b) allows the inclusion of children who were unable to take the test because they failed the practice runs (assigned a score of zero).

Vocabulary Subscales (ÉPE–AD)

In winter 2009, SRDC reworked the structure of the ÉPE–AD scales, yielding two vocabulary scales that promised an increased sensitivity to slight differences in children's language skills. The first subscale measures expressive vocabulary, that is, the child's ability to say the word associated with a picture shown to him or her. This scale contains six items from domain A and two items from domain C. Internal consistency is very good for expressive vocabulary items, with Cronbach alphas ranging from 0.81 to 0.88 for the English and French versions.

The second subscale measured receptive vocabulary and consists of two items from domain A and three items from domain C. The child is shown a series of pictures and must identify the picture associated with the word spoken by the tester. Internal consistency for the items measuring receptive vocabulary in the French version is acceptable, with Cronbach alphas ranging from 0.62 to 0.71. However, internal consistency for the English version of the receptive vocabulary subscale is less acceptable, with several alphas of less than 0.50 (see Table 2.8).

[21] The current standards for the test's French version were established in 1990 and have not been updated. As a result, their validity is unknown.

Table 2.8: Cronbach Alpha for ÉPE–AD Vocabulary Subscales by Assessment Period

	Cronbach Alpha [α (*n*)]						
	Oct. 2007	***Feb. 2008***	***June 2008***	***Oct. 2008***	***Feb. 2009***	***June 2009***	***Oct. 2009***
Domains administered in French							
Expressive vocabulary (8)	**0.85** (213)	**0.88** (215)	**0.88** (215)	**0.88** (220)	**0.86** (*233*)	**0.85** (*237*)	**0.84** (*237*)
Receptive vocabulary (5)	**0.71** (213)	**0.69** (215)	**0.64** (215)	**0.52** (220)	**0.62** (*233*)	n/a	n/a
Domains administered in English							
Expressive vocabulary (8)	**0.83** (30)	**0.86** (32)	**0.87** (28)	**0.81** (20)	**0.87** (6)	n/a	n/a
Receptive vocabulary (5)	**0.37** (30)	**0.73** (32)	**0.40** (28)	**0.53** (20)	**0.48** (6)	n/a	n/a

Early Years Evaluation — Direct Assessment, Modified Version (SRDC, 2009)

In winter 2009, SRDC anticipated that ceiling effects were likely in upcoming fifth evaluation (due in June 2009) based on the strength of projects derived from early analyses of ÉPE–AD scores. A ceiling effect would render difficult our ability to follow children's development and by extension, establish differences among the three treatment groups. The tool's designer was therefore hired by HRSDC to create new, more difficult questions, thereby allowing for continued tracking until October 2009 of the children's developmental trajectory for domains A, B, and C. The designer compensated for the additional administration time associated with the new items by eliminating the test's easiest questions. The "extended version of the [ÉPE–AD]" (Willms, 2009) was ready in May 2009.

In parallel, SRDC reworked the questions that made up the extended version of the ÉPE–AD. The goal of this restructuring was to more sensitively measure the children's language skills while retaining the ability to follow their developmental trajectory based on three of the four ÉPE–AD domains (that is, Awareness of Self and the Environment [domain A]; Cognitive Skills [domain B]; and Language and Communication [domain C])[22]. The new tool, called "the modified ÉPE–AD", also gave rise to subscales measuring expressive and receptive vocabulary mentioned above. The modified ÉPE–AD includes:

- 15 questions in domain A concerning mainly expressive vocabulary and general knowledge;
- 12 questions in domain B measuring phonological awareness (that is, children's ability to play with the sounds that make up words) and numeracy concepts (that is, children's ability to play with numbers). Two other items in domain B were administered in a manner slightly different from the extended ÉPE–AD. Thus, the

[22] For more about the reworking of the ÉPE–AD and the theory upon which it is based, refer to the report presenting a review of direct measurement tools for children's developmental dimensions submitted to HRSDC in July 2009.

child's knowledge of the sound of all letters in the alphabet was measured using a booklet that presents uppercase letters in ascending order of difficulty for young Francophones. Knowledge of the name of every letter in the alphabet was measured using a booklet in which lowercase letters are presented in ascending order of difficulty for young Francophones;

- Finally, two questions in domain C were retained to continue the expressive vocabulary scale and two new, more difficult questions were added to measure children's phonological awareness.

The initial ÉPE–AD was used until February 2009, while the "modified ÉPE–AD" — the version of the tool reworked by SRDC — was used for the evaluations of June and October 2009. See Appendix C for a comparison of content for the three versions of the ÉPE–AD.

Within the framework of the Readiness to Learn project, it was important to monitor the children's developmental trajectory from October 2007 to October 2009. Establishing continuity in the dimensions measured by two tools is based on an examination of the pattern of correlations among the subscales measuring the same dimension though included in each of the tools. A strong correlation (that is, of 0.75 or higher) among subscales is a strong indication that they measure the same underlying construct. Indeed, an examination of the pattern of correlations in Table 2.9 confirms the presence of strong correlations between domains A and B of the initial ÉPE–AD (0.83 and 0.75 respectively, February 2009 assessment) and those of the modified ÉPE–AD (June 2009 assessment). The exception is the communication domain (domain C), where we observe a correlation of 0.61. This result is lower than was hoped; however, it is offset by strong correlations between domain C of the initial ÉPE–AD and the expressive vocabulary subscale created using the ÉPE–AD (0.84) as well as between domain C of the initial ÉPE–AD and Gardner's 'expressive vocabulary' scale (0.75). There is also a strong correlation between domain C of the initial ÉPE–AD and the scale of domain A for the modified ÉPE–AD. Together, these results suggest that use of the modified version results in a negligible loss of information. Consequently, we are able to follow the developmental trajectory of the three treatment groups from September 2007 to October 2009.

Table 2.9: Pearson Correlation Coefficients for ÉPE–AD Scales and Expressive and Receptive Vocabulary Scales

		Initial ÉPE–AD: February 2009				
	Scales	**A**	**B**	**C**	**Expressive vocabulary subscale (ÉPE–AD)**	**Receptive vocabulary subscale (ÉPE–AD)**
Modified ÉPE–AD: June 2009	**A**	**0.83** (*227*)	**0.66** (*232*)	**0.84** (*237*)	**0.81** (*221*)	**0.63** (*222*)
	B	**0.55** (*227*)	**0.75** (*232*)	**0.52** (*237*)	**0.53** (*221*)	**0.47** (*222*)
	C	**0.60** (*227*)	**0.63** (*232*)	**0.61** (*237*)	**0.61** (*221*)	**0.47** (*222*)
	Expressive vocabulary subscale (ÉPE–AD)	**0.79** (*227*)	**0.52** (*232*)	**0.82** (*237*)	**0.81** (*221*)	**0.58** (*222*)
	Expressive vocabulary (Gardner)	**0.75** (*227*)	**0.53** (*232*)	**0.75** (*237*)	**0.76** (*221*)	**0.50** (*222*)
	Receptive vocabulary (PPVT–R)	**0.66** (*227*)	**0.56** (*232*)	**0.69** (*237*)	**0.68** (*221*)	**0.43** (*222*)

Note: The modified ÉPE–AD does not include a receptive vocabulary subscale. These were eliminated by the tool's designer in the extended version of the ÉPE–AD since they are easy. An analysis of the Readiness to Learn project data confirms that these data were too easy and therefore no longer permitted a distinction among the children.

2.5.3. Parent Measures

Follow-up surveys of parents allow for determining the degree to which the child's environment changed in the four months prior to the survey. First, they measure the frequency of literacy activities, the language of these activities, and other factors that affect children's school readiness and the development of their francophone identity. Second, they systematically verify changes in the child's type of child care and the language used in the new type of child care so that analyses can take that information into account. Third, they document changes in family composition in order to determine whether the child's normal language environment at home is affected by such changes.

The scales used in the surveys of parents participating in the Readiness to Learn project are from studies of francophone populations in Canada, such as the NLSCY, the Québec Longitudinal Study of Child Development (QLSCD; Institut de la statistique du Québec, 2003) and the Survey on the Vitality of Official-Language Minorities (SVOLM; Corbeil, et al., 2007). The questions selected for the Readiness to Learn project are those relevant to the children's age bracket.

Note that this section details only the scales included in the impact analyses. It is therefore not surprising that this list of scales differs somewhat from those reported in the *Reference Report* (Legault et al., 2014). Several other constructs were measured but were not retained for the impact analyses. The decision on whether or not to include them is based on the preliminary analyses for which results are reported in Chapter 4.

Socio-Demographic Characteristics

The socio-demographic characteristics of the children and their parents stem from questions asked in the parental consent form (for example, the child's sex and date of birth) and the baseline survey. The follow-up surveys enabled us to make any necessary changes to this initial profile. This section identifies the socio-demographic variables used for the impact analyses.

Information such as the mother's and father's level of schooling, the mother's age at the child's birth and family income were gathered from the parents in the baseline survey. Family income was requested on two other occasions. An open question on income was initially asked in February 2009. When a number of parents (24.6%) refused to share this information, the same question was posed again in October 2009 to the parents who had refused to answer, but this time with income brackets as response options.

Several questions were intended to establish the linguistic profile of participating families. Among others, respondents were asked to identify their mother tongue, knowledge of official languages, the languages spoken with the child, and those spoken at home. The same questions were asked for the spouse. There are two common definitions for Canada's francophone population. According to Statistics Canada,[23] the first approach is to calculate the population that declares French as their mother tongue, that is, the first language learned and still understood. The second approach is to calculate the "first official language spoken" (FOLS), a score that factors in knowledge of both official languages, mother tongue and the language most often spoken at home (Forgues & Landry, 2006).

Every follow-up survey begins with a series of five questions addressed only to parents who experienced a change in family composition since the last survey completed. The impact analyses take into consideration changes in family composition in comparison with answers in the baseline survey (e.g., household size, single-parent families, older children, twin children, younger children).

The immigrant status of parents was established using a question in the October 2009 survey. This survey also updated the socio-demographic data for the respondent's spouse. Note that in the baseline survey, the information collected defined the characteristics of the spouse who was the child's biological parent, even if that parent had no contact with the child. To clarify the results, these questions were asked again in October 2009 in order to obtain information on the second adult who lives in the household with the child and who is actively involved in caring for him or her.

Language Habits

The baseline survey and follow-up surveys include several questions on language habits. The respondents choose the category most representative of their experience from among several options (for example, whether they speak English only, English and French, French more than English, etc.). These categories were combined to create a five-point scale representing a linguistic continuum where a score of 1 means the respondent speaks in "English only," a score of 3 means he/she speaks "French and English and/or another language," and a score of 5 means

[23] For a definition of the francophone population, see the Statistics Canada website at http://www.statcan.gc.ca/concepts/definitions/language-langue-eng.htm.

he/she speaks “French only.” The impact analyses used this linguistic continuum based on a five-point scale.

A first linguistic continuum was generated representing the language usually spoken by the mother to the child. A second linguistic continuum was calculated using the average scores for four questions on the languages spoken by the child with his or her mother, father, friends and siblings at home, as well as outside the home (Cronbach alpha of 0.95).

Two indices representing the language usually spoken at home were derived. A first index was derived by combining the mother’s FOLS with the father’s. A second index was created by combining the language usually spoken by the mother to the child with the language usually spoken by the father to the child. These two indices were used to determine the child’s language environment at home. Moreover, the FOLS was used to determine whether the child lives in an endogamous Francophone, endogamous Anglophone or exogamous household.

Family Processes

The baseline survey gathered information on family processes. The follow-up surveys also measured some of these processes. In total, five were measured using four scales: parenting style, family functioning, depression, and literacy activities.

The ***Positive Parenting*** scale encompasses five items measuring the frequency of positive contacts between the parent and child (for example, how many times the parent praises the child, laughs with him or her, or physically expresses affection). Items are rated on a five-point scale where 1 indicates “never” and 5 indicates “several times a day.” The scale’s internal consistency is acceptable at 0.66 (Cronbach alpha).

The ***Authoritarian Parenting*** scale consists of four items measuring the degree of supervision and discipline. Content of the items seeks to determine, among other information, whether or not the child respects the punishments given to him or her and whether these punishments vary depending on the parent’s mood. Items are rated on a five-point scale where 1 indicates “never” and 5 indicates “almost always.” Internal consistency is acceptable with a Cronbach alpha estimate of 0.61.

The ***Family Functioning*** scale, made up of eight items, assesses the quality of relationships within the family. Content of the items concerns emotional openness, expression of feelings and friendliness within the family. The choice of answers for the scale’s items range from 1, for “absolutely disagree”, to 4, for “fully agree.” The family functioning scale shows an excellent internal consistency with a Cronbach alpha estimate of 0.82.

The ***Depression*** scale measures, for example, the number of times that the respondent has felt depressed or unhappy, has cried or felt alone, or does not enjoy life. The scale’s eight items cover a four-point range, where 1 indicates that the respondent has “rarely or never” felt this way and 4 means that he or she feels this way “most of the time or all the time.” The depression scale has a very high internal consistency with a Cronbach alpha estimate of 0.80.

The ***Frequency of Literacy Activities*** scale consists of five items measuring the frequency of reading, writing and numeracy activities. More specifically, the items measure the frequency at which parents tell stories to their child without a book, sing songs with him or her, teach him or her to write or trace letters or numbers, show him or her how to read words, and encourage him or her to use numbers in his or her daily activities. These items are rated on a five-point scale,

where 1 corresponds to an activity that has never been done and 5 corresponds to an activity done several times a day.

Every question on literacy activities is followed by a question verifying the language used during that activity. Respondents choose the category most representative of their experience from among several options (for example, if they speak English only, English and French, French more than English, etc.). These categories were combined to create a five-point score representing a linguistic continuum *for this activity* where a score of 1 indicated the activity was done in "English only", a score of 3 indicated the activity was done in "French and English and/or another language", and a score of 5 indicated the activity was done in "French only." The average scores for each question are then grouped together to create a ***Language of Literacy Activities*** scale. Internal consistency is very high with Cronbach alpha estimates ranging from 0.91 to 0.95).

Questions leading to the literacy activities scale and language of literacy activities scale were asked at each survey, including the baseline survey. Internal consistency for these scales is acceptable enough at each survey period for confidence in the homogeneity of these scales (see Table 2.10).

Table 2.10: Cronbach Alpha for the Literacy Activities Scale and the Language of Literacy Activities Scale at each Survey Period

	***Cronbach Alpha** [α (n)]*						
	Oct. 2007	Feb. 2008	June 2008	Oct. 2008	Feb. 2009	June 2009	Oct. 2009
Literacy activities	0.59	0.54	0.54	0.60	0.64	0.67	0.59
Languages used during activities	0.95	0.94	0.94	0.94	0.91	0.91	0.93

Identity-Related Dimension

The baseline survey included several questions on perceptions concerning the ***Subjective Vitality of the Francophone Community***, ***Engagement in Francophone Culture*** and the ***Sense of belonging to the Francophone culture***. Of all these identity-related dimensions, only the sense of belonging to the francophone culture was deemed useful for the impact analyses. The purpose of the question was to determine the cultural group with which the parents identified (the Francophone group, Anglophone group, both, or another group).

2.5.4. Parents' Knowledge, Attitudes and Beliefs

The surveys before and after the workshops contained a series of questions intended to highlight the nature and scope of changes in the perceptions and behaviours of the parents in the Program Daycare group. For the purposes of assessing the family literacy workshop program, several measurements were created based on the questions used in family literacy studies, particularly those by the Centre for Interdisciplinary Research on Citizenship and Minorities (CIRCEM; LeTouzé, 2006) and the Coalition francophone pour l'alphabétisation et la formation de base en Ontario as part of their study on the impact of family literacy on families living in minority settings (Benoît, n.d.). Measurements were also created using the post-intervention questionnaire for the school transition program of the Conseil d'éducation du District scolaire 3

in Edmundston (Gauthier St-Onge, n.d.) and using the NLSCY (Statistics Canada, 2006). Finally, the retrospective approach used to create the questions was inspired by a retrospective survey created by Lamb and Tschillard (2005).

The choice of constructs for studying the impacts of the family literacy workshop program is based in part on the reasoned action model developed by Fishbein and Ajzen (for an introduction to the model, see Brigham, 1991). This model links beliefs and attitudes with the prediction of behaviours. According to the model, people's beliefs regarding behaviour influence both their attitudes toward this behaviour and their perceptions of control (a concept similar to sense of self-efficacy). In turn, attitudes and perceptions of control influence the intention to adopt the behaviour. Finally, intention determines the probability that the behaviour will be adopted. In the Readiness to Learn project, we measured participants' beliefs and attitudes with respect to the aspects targeted by the family literacy workshops as well as their self-efficacy. We also measured participants' knowledge regarding the main themes addressed in the workshops. Although not included in the reasoned action model, the level of knowledge also affects peoples' attitudes about an action. Finally, a modelling scale was added to capture the importance of parents' literacy activities on children's attitudes and behaviours. In fact, several studies have shown the positive effect of an environment rich in family literacy, such as the presence of written material or material for writing, as well as parental behaviours that value reading and writing, like reading or writing in front of children (Dionne, 2007; Sénéchal & LeFevre, 2002; Jordan, Snow, & Porche, 2000). The four constructs specific to the pre- and post-workshop surveys are detailed in the paragraphs below.

Four questions measure certain ***Beliefs and Attitudes*** of the parents, which are addressed in the family literacy workshops. These beliefs concern major program themes, particularly: the parent as his or her child's first educator, cooperation between educator and parents, the francophone cultural environment at home, and the appropriate age for reading to a child. The respondent indicates the degree to which he or she agrees with each of the statements by using a four-point scale where 1 indicates "absolutely disagree" and 4 indicates "absolutely agree." These questions were asked in the pre- and post-workshop surveys. Note that they do not form a scale, which means the information collected is descriptive in nature.

The ***Perceived Self-efficacy*** scale assesses parents' perception of their ability to carry out certain actions discussed in the family literacy workshops. Thus, parents are asked how confident they are about their ability to: manage their child's emotions and behaviours, speak to him or her in French, help him or her learn, and help him or her develop a francophone cultural identity. Each of these aspects is the object of two questions. The first is retrospective in nature, asking the parents how confident they felt *before* the workshops. The second measures the parents' state of confidence *after* the end of the workshops (that is, at the time of the survey). These items are rated on a five-point scale, where 1 indicates "little or no confidence" and 5 indicates "complete confidence." Internal consistency is good with a Cronbach alpha of 0.72 for the subscale measuring the retrospective perceived sense of self-efficacy, and of 0.62 for the subscale measuring perceived sense of self-efficacy after the workshops.

The self-reported ***Knowledge*** scale assesses the parents' level of knowledge in comparison with five aspects addressed in the family literacy workshops, including child development, how to enhance child development and learning at home, how to help a child prepare for school, the work of a daycare educator, and the French-language services or resources available in the

community. As with the scale on perceived self-efficacy, each of these aspects is the subject of two questions: one retrospective (How would you describe your level of knowledge on the subjects *before* the family literacy workshops?) and the other after the workshops (How would you describe your level of knowledge on the subjects *now*, *after* the family literacy workshops?)[24]. These items are rated on a five-point scale, where 1 indicates "no knowledge" and 5 indicates "excellent knowledge." Internal consistency is acceptable with a Cronbach alpha of 0.79 for the subscale measuring the perceived level of knowledge before the workshops and of 0.66 for the subscale measuring the perceived level of knowledge after the workshops.

The ***Modelling Behaviour*** scale consists of five questions based on the literacy activities model. It aims to measure the example modeled by the parent for his or her child with regard to the importance of reading and writing. Content of the items concerns the frequency at which parents read and write in front of their child, as well as the frequency at which parents make paper and pencils available for their child's use. The questions are rated on a four-point scale, where 1 indicates "never" and 4 indicates "several times a day." This scale was administered to parents in the Program Daycare group in the pre- and post-intervention survey administered 3 weeks following the end of the workshops (i.e., in March 2008 for the first cohort and in March 2009 for the second cohort). Internal consistency was good with a Cronbach alpha of 0.73 and of 0.68 observed for each respective administration.

2.5.5. Dosage

The number of hours of child care per week is an important variable to consider in measuring the program "dosage" that children receive at daycare. To gather this information, SRDC used the attendance record that educators already fill out as one of their routine tasks, including the children's time of arrival and departure. The number of hours the child spends at daycare is then compiled on a weekly basis and sent monthly to SRDC's Ottawa office. For the purposes of the first year impact analyses, the hours spent at daycare were analyzed by taking the average hours spent in daycare per week for the four months immediately prior to the children's assessment (see Section 5.1.3). The second year impact analyses use the average hours spent per week for the first eight months after the intervention (see Section 6.2).

In terms of the family literacy workshops, the attendance of parents, children and other adults was compiled for each workshop. Dosage for workshops is calculated using the total workshops taken by at least one adult.

2.5.6. Implementation Indices for the Daycare Component

The importance of tracking how interventions are implemented in the field is highlighted in a growing numbers of studies (Charlebois, Brendgen, Vitaro, Normandeau, Bourdreau, 2004; Conduct Problems Prevention Research Group, 1999; Dane & Schneider, 1998; Durlak & Dupre, 2008). On this point, Durlak and Dupre (2008) concluded on the basis of data from five meta-analyses that the effect of an intervention is two to three times greater when fidelity of implementation is taken into considerations in the impact analyses. Moreover, a more nuanced interpretation of impact analyses is possible when they are considered in light of the integrity and quality of the daycare program's implementation. Accordingly, six indices were incorporated

[24] For a comprehensive discussion of the benefits of using retrospective estimates in gauging the effect of training sessions, see Lamb and Tschillard (2005).

into the impact analyses, including structural fidelity, content fidelity, structural quality, reading quality, educational quality, and sensitivity of the educators. Each index is the result of the quantification of observation notes taken onsite in the program daycares and comparison daycares. These observation notes were transformed into subscales of the *Échelle d'évaluation de l'environnement préscolaire* – revised edition (ÉÉEP-R), the French version of the Early Childhood Environmental Rating Scale – Revised, for the purposes of the impact analyses[25].

One of the first elements used in the impact analyses was fidelity of program implementation. This involves ensuring that all the program elements were implemented at the daycares. An observation checklist was developed to check the presence of certain elements specific to the program being tested (like the presence of a weekly program that is visible to parents or of a routine chart that shows children how their day will unfold). The observers were also required to note the unfolding activities and the children's reaction during the various activities. These observation notes were used to develop two indices. The first, called ***structural fidelity***, concerns the presence of certain elements in the environment, like image–word posters or a routine chart. The second index, ***content fidelity***, concerns the integration of program elements into the programming of the child care environment, such as reading to the children at circle time or to carry out reasoning activities with them.

The observation notes also provided information about the quality of the children's various child care settings. Specifically, these observation notes allowed for filling out eight subscales of the ÉÉEP-R. The ÉÉEP-R is a validated tool for assessing the child care quality as defined by the early childhood education specialists of the National Association for the Education of Young Children (NAEYC). Interestingly, this tool has been used several times by francophone projects to assess the quality of child care (e.g., Japel, Tremblay, & Côté, 2005).

For the purposes of the impact analyses, four subscales of the ECERS–R were used to create a first index for the quality of child care settings. The first index, ***structural quality***, concerns the overall quality of the child care environment. It includes the ÉÉEP-R subscales (filled out based on observation notes) "Indoor space", "Health practices" and "Greeting/departing." The second index, ***quality of reading***, isolates the "Books and pictures" subscale, given the particular importance of reading to children's vocabulary acquisition. The third index, ***educational quality***, taps activities that encourage communication by children and enrich their vocabulary. This index comprises the ÉÉEP-R subscales "Informal use of language", "Encouraging children to communicate", and "Using language to develop reasoning skills." Finally, the fourth index takes into account the particularly important influence of ***educator sensitivity*** on the child's development. To this end, the observation notes enabled us to fill out the "Staff-child interactions" subscale (henceforth "educator sensitivity").

The indices are interpreted as follows: the two fidelity indices indicate the proportion of implemented program elements while the four quality indices reflect the degree to which the program's main elements are implemented based on a seven-point scale, where 1 corresponds to care that is well below the fundamental requirements of child care and 7 corresponds to personalized, high-quality care (Harms, Clifford, & Cryier, 1998). These six indices were used to

[25] A decision was made at the outset of the Readiness to Learn project to not directly fill out the ÉÉEP-R grid since this type of observation may be perceived as overly intrusive, potentially making it more challenging to obtain the full cooperation of program group and comparison daycares. It was therefore decided that observers would take notes on the various elements of the ÉÉEP-R for the targeted subcategories, which would be coded using the grid at a later time.

link, in the impact analyses, program fidelity and quality in the child care environment with the children's development and school readiness in the first year of program delivery. As regards the analyses for the second year of program delivery, the program's effects on the children were verified using two general indices of fidelity and quality. These general indices are the program's average fidelity indices and average quality indices.

These six fidelity and quality indices as well as the two general indices of fidelity and quality were used to validate the results of the main impact analyses for the daycare program (see Chapters 5 and 6). Note that this report addresses only the indices used in the impact analyses. A complete description of the modalities and tools used to study the integrity of daycare program implementation is found in the *Project Implementation Report* (Bérubé et al., 2014).

2.5.7. Implementation Indices for the Family Literacy Workshops

Observations notes taken during the family literacy workshops were a source of relevant data on program delivery as defined by the objectives of the Readiness to Learn project in general and those of the family literacy workshop program in particular. A semi-structured checklist was created for the characteristics to be observed, although the observations themselves consist of informal notes. The information gathered allows for an examination of fidelity and quality of program delivery across sites to determine whether variability in certain aspects of implementation may explain the presence or absence of change reported by parents.

Several documents served as the basis for creating the various sections of the observation checklist. Questions that capture more general aspects of workshop delivery are based on the observation framework proposed by Merriam (1988). Questions related to program content, organization and the workshop practitioner's style are based on the recommendations of the Fédération canadienne pour l'alphabétisation en français (FCAF, 2007). Questions concerning parents' participation and the workshop practitioner's style are based on those proposed in the "Guide for Using the Classroom Observation Tool" (Baylor College of Medicine, 2001). Finally, questions concerning clear presentation of the lesson's/workshop's goal are based on the guide entitled "Inside the Classroom: Observation and Analytic Protocol" (Horizon Research, 2000).

Program fidelity was determined via observation notes taken for each workshop. This indicator reflects the number of mandatory elements addressed in each workshop. ***Material coverage*** was found to be excellent with an average of 96% across the 11 workshop groups and a minimum observed coverage of 90% across workshop session for any given group. Next, program quality was assessed by means of various indicators including the ***practitioner's style of delivery*** and participant satisfaction with ***session length***. This information was collected in the anonymous survey administered during the last workshop. Program quality was also assessed using observation notes of participant ***positive and negative reactions*** during the workshop. Lastly a ***global quality index*** was computed consisting in the average scores obtained on the three indices of material coverage, practitioner's style of delivery and satisfaction with session length. The results of analyses investigating the impact of these five indices on parent outcomes are presented in Section 5.2.3.

2.5.8. The Impact of the Daycare Component on the Children's Linguistic Dimensions

The main targets of the daycare program are school readiness, the enrichment children's language skills, and the development of a francophone cultural identity. Accordingly, we expect the most pronounced program impacts on these aspects of child development. The stated hypotheses are:

1. Program Daycare group children will score higher in three of the four domains measuring school readiness, including the domains of language and communication, cognitive skills, and awareness of self and the environment, versus children in the comparison groups. No difference is expected between the Program Daycare group and the comparison groups in terms of physical and motor development since the program being tested does not target such skills.

   This hypothesis is intended as a direct test of the main research question. The study's internal validity is reinforced by the inclusion of the Comparison Daycare group, which allows an evaluation of the program relative to this alternative form of treatment. It is further buttressed by the addition of a second comparison group, which accounts for the influence of an informal child care setting on the development of French-language skills.

   Finally, this hypothesis takes into account other conditions in the community that may affect child development. For example, we assessed the community environment, which plays an important role in child development by providing resources (e.g., park, wading pool, bike paths) and French-language services (e.g., bookstore, library, swimming lessons, etc.).

2. Program Daycare group children will perform better on measures of language skill than comparison group children.

3. The tested program will have a more pronounced impact on the language skills of Program Daycare group children from exogamous households or Anglophone endogamous households.

4. The magnitude of the daycare program's impact on the children's school readiness dimensions will be influenced by their level of exposure to the program. Thus, greater exposure to the program being tested will be associated with better performance by Program Daycare group children in three of the four domains measuring school readiness, including the domains of language and communication, cognitive skills, and awareness of self and the environment, versus children in the comparison groups. No difference is expected between the Program Daycare group and comparison groups in terms of physical and motor development since the program being tested does not target such skills.

5. Program daycares will be characterized by higher program fidelity and program quality scores than those observed in comparison daycares.

6. The magnitude of the daycare program's impact on the school readiness and language skills dimensions will vary across daycares based on the fidelity and quality of program delivery.

2.5.9. The Impact of the Daycare Component on the Children's Identity and Cultural Dimensions

The tested program was conceived so as to modify the environmental contexts of children so as to promote the development of francophone cultural identity. The importance of measuring cultural identity is based on research findings emphasizing that children's exposure to French in several spheres of their life strengthens their sense of identity and belonging to the francophone community (Landry & Allard, 2000). However, these studies measure the concept of cultural identity at adolescence. This choice is appropriate according to Erik Erikson's theory of self (1994), which states that identity begins to form around age 12 and matures at about age 25. Thus, a direct measure of the cultural identity at this very young age is impossible.

Instead, their degree of ***exposure*** to French can be measured through their parents. Parents are in the best position to inform us about the various settings relevant to the Readiness to Learn project that contribute to the child's day-to-day language environment, such as the family, the daytime child care environment and access to French-language services and resources. The collected information will serve to adjust the statistical models estimated during impact analyses with the aim of better identifying the effects of the program being tested on the development of young children's language skills.

2.5.10. The Impact of the Family Literacy Workshop Component on Parental Attitudes and Behaviours

We anticipate the focus of the content covered by the family literacy workshop program will determine where the strongest effects will be observed. Accordingly, we expect strong effects for the following four dimensions:

- The parent as his or her child's first educator;
- The early reading, early writing and early numeracy activities likely to foster the child's school readiness;
- The importance of maximizing the child's exposure to French; and
- Parent–educator cooperation in the child's learning.

The stated hypotheses are as follows. Subsequent to the family literacy workshops and compared with the parents in the comparison groups, parents in the Program Daycare group will report:

1) A significantly more frequent occurrence of various literacy activities with their child;
2) A significantly more frequent occurrence of speaking French with their child during literacy activities.

Moreover, we anticipate that subsequent to the workshops parents in the Program Daycare group will report:

3) Significantly better knowledge of child development, ways to help them prepare for school, and French-language community services or resources;

4) Significantly better sense of self-efficacy, particularly in terms ways to help their child prepare for school;

5) Significantly increased frequency of modelling literacy activities;

6) A significant change in their beliefs, especially for elements targeted by the family literacy workshop component, notably: the parent as his or her child's first educator, parent–educator cooperation in the child's learning, the importance of the francophone cultural environment at home, and the appropriate age for reading to a child.

In the next chapter, we detail the analytic strategy adopted in the impact analyses for the program's first and second years.

3. Analytical Strategy

This chapter introduces the logic behind the analyses reported in Chapters 6 and 7. Here exposure to the tested program is conceived as a continuum spanning from low intensity to high intensity. The highest possible intensity of exposure would be received by a child who attends full-time a daycare that implements the program exactly as intended with the highest level of quality. In what follows, we elaborate on the way treatment intensity was defined and how this concept was used in estimating the impact of the tested program. Both treatment groups and more fine-grained indicators of treatment intensity were employed for this purpose.

3.1. CONCEPTUALIZING TREATMENT AND DOSAGE

The simplest possible model for capturing a treatment effect would classify participants into two groups: treated and untreated. Such a model is most valid when the distinction between treated and untreated is absolute, which is to say that the untreated group receives absolutely no treatment while all members of the treated group receive an equivalent treatment (e.g., an equivalent dosage or "exposure"). In a field study, such clean distinctions are relatively rare and the Readiness to Learn project is no different in this respect. The intensity of treatment received by participants varied along at least two dimensions: amount of exposure (dosage) treatment and treatment quality. We now discuss in more detail both sources of variability in treatment intensity.

3.1.1. Heterogeneous Treatment Exposure

When participants were enrolled in the project, they were assigned to one of three groups: the Program Daycare, the Comparison Daycare, or the Informal Care groups. Perhaps unsurprisingly, these assignments sometimes varied over time as parents altered their daycare choices. For instance, a relatively small percentage of parents (see Chapter 3) opted to change daycares between testing periods. As a consequence, a child who was enrolled at a program daycare for the first two testing periods could be enrolled in a comparison daycare or in informal care for the third and fourth.

Even in cases where participants did not migrate from one group to another, the amount of time spent in daycare could fluctuate. The two daycare groups did not differ in terms of their average time of exposure to the daycare environment for any of the seven four-month periods preceding the evaluations of the ÉPE–AD (see the results presented in Section 4.4). A practical consequence of this equivalence is that this variable does not present an obvious threat to the validity of our program-effect estimates. Nevertheless, it is interesting to ask whether the total exposure to a treatment condition matters and whether the program effects, if any, interact synergistically with total exposure. In other words, does an hour spent in one of the program daycares result in better outcomes on average than an equivalent amount of time spent in one of the comparison daycares? We addressed this issue by including the variability in treatment exposure in impact analyses.

3.1.2. Evaluating the Effect of Dosage/Exposure

Two strategies were adopted for managing the migration of participants over time. The first of these strategies is coarse but has the advantage of the simplicity associated with group comparisons (i.e., treated vs. untreated). In the second case, treatment is defined more precisely in terms of hours of exposure to a particular daycare setting (i.e., average hours per week spent in daycare). The latter strategy is slightly more complex, but has the advantage of representing the total amount of exposure to the two daycare treatment conditions. These two techniques for representing heterogeneity in treatment exposure are the object of two independent series of analyses. We now discuss these techniques in more detail.

The first series of analyses is based on the coarse definition of treatment exposure whereby dummy codes were used to represent group membership. For any given time period, a participant coded as belonging to the treatment group was deemed to be fully treated. Group membership was allowed to vary over time, but the integrity of the treatment groups was maximized by requiring that participants be exposed to their new mode of care for a minimum period of time before recognizing the change in the analyses. More precisely, the following two conditions were imposed. First, a child was considered to have changed groups only if the change occurred more than a month prior to the evaluation. Second, changes in daycare arrangements over the summer months were not considered valid unless they persisted for a few months into the school year. A participant who reported withdrawing from a program daycare and subsequently enrolling in a comparison daycare during the summer, for example, would be considered to be a member of the Program Daycare group for the fall evaluation. For the purposes of analyses, the group change would only take effect for the winter evaluation because by then, children would have been exposed to the program for a few months during which the habitual daycare program is in place. The two standardized vocabulary measures were exceptional in that they were each measured only once. Regardless of the outcome measure, we evaluated in all cases the potential bias that group changing could engender.

In the second series of analyses of the Year 1 data, treatment was defined using the hours of exposure to daycare in conjunction with the characteristics of the daycare setting (i.e., Program or Comparison Daycare group). Hours of exposure was defined as the average hours per week a child spent in daycare during the four month period preceding a given evaluation. The exception was the baseline period for which only two months (September and October 2007) were used. The inclusion of this variable in the analyses allowed us to define exposure to our daycare program in a more fine-grained manner. By crossing the hours of exposure variable with our grouping variable (i.e., specifying an interaction term), we were able to: a) estimate the average treatment effect associated with a given number of hours per week of exposure to daycare, and b) test whether the effect of amount of exposure to daycare varies as a function of program type. The latter test is simply an extension of our basic research hypothesis whereby given an equivalent amount of exposure, we expect an advantage for the program group. The added value of this more precise conceptualization of dosage is its greater power in detecting program effects and its superior level of detail, which could potentially reveal the amount of exposure necessary to show the desired effect.

The follow-up analysis of the Year 2 outcomes made use of the hours of exposure data slightly differently. By the end of the first year, most children in the study were enrolled in junior kindergarten either part-time or full-time (i.e., children living in the three Ontario communities).

For this reason, daycare attendance data that was collected during the second year of the project is difficult to interpret as a predictor of outcomes because it does not adequately reflect time spent in a quality child care environment. Children enrolled on a part-time or full-time basis in school are exposed to a very high-quality environment. Consequently, a child who spends a few hours per week in a program daycare environment may actually spend many hours outside the home experiencing a treatment of comparable intensity.

It is clear that 'hours of exposure to daycare' is not a meaningful variable in the second year of the project. It would technically be possible to perform an analysis of this type using only the community of Edmundston (i.e., no children in this community were enrolled in school), but the number of children in such an analysis would be too small to provide useful information. Instead, we treated the Year 2 evaluations as follow-up tests of the effect of hours of exposure to daycare in Year 1. The investment in providing daycare services in the second year of the project was nevertheless important in ensuring the maintenance of program effects for those children not enrolled in school. The performance of these children is contributing to the program effects reported in Chapters 5 and 6. In other words, the persistent effects of hours of treatment exposure were tested in this analysis.

To capture this idea, we computed a variable representing the average hours a week each child spent in either a program daycare or a comparison daycare during the first eight months of the program's implementation[26]. This variable was crossed with treatment group in a regression model predicting the Year 2 outcomes of children. Accordingly, Year 1 treatment group was crossed with Year 1 daycare dosage in an attempt to predict the Year 2 outcomes of children. If an hour spent in a program daycare in Year 1 results in better outcomes on average than an hour spent in a comparison daycare, then we would expect to observe a significantly more positive effect of dosage for the Program Daycare group than the Comparison Daycare group in the second year if this effect is either persistent or delayed.

3.1.3. Evaluating the Effect of Program Daycare Fidelity/Quality

The mechanism by which the intervention was supposed to affect the developmental outcomes of children was the quality of the program and the fidelity of its implementation. In this sense, treatment group membership is a proxy for quality and fidelity. The internal validity of the study depends crucially on the extent to which this is true. A necessary condition for observing a program effect on the targeted outcome variables is that the children who are enrolled in daycares where the intervention has been implemented (Program Daycare group) lived experiences in their environment that compare favourably to those of children in the comparison condition (Comparison Daycare group). Similarly, within each groups, the nature of the program being provided should be as similar as possible, which is to say consistent. In other words, the way daycares are grouped together should be coherent. These two conditions were verified by way of qualitative analyses presented in the *Project Implementation Report* (Bérubé et al., 2014) and in a series of preliminary quantitative analyses reported in Section 4.4.

The use of treatment groups for estimating the magnitude of treatment effects is a useful simplification by which potentially continuous dimensions (i.e., fidelity and quality) are reduced

[26] We excluded data from the four months preceding the fourth evaluation given that it captures variance due to summertime disruptions. The purest measure of the average treatment exposure effect is obtained by targeting the information that was collected during the school year.

to categories (i.e., Program Daycare group, Comparison Daycare group) for the purpose of making comparisons. This simplification is most appropriate when the members of each group are very similar to each other on the continuous dimensions. Of course, the nature of the program offered in a daycare always varies somewhat from other daycares, which means that using treatment groups results in a loss of information (i.e., the within-group variability).

We verified whether this loss of information was meaningful by conducting a series of analyses where the treatment group "middle-man" was eliminated by instead using continuous indicators of fidelity and quality as predictors of child outcomes. This technique ensures the fullest use of available information and an increased chance of detected true program effects (i.e., a more efficient research design).

We were also interested in determining whether any observed treatment group effects were due to the program itself or some other characteristic of program daycares. We tested this idea by estimating treatment group effects after controlling for fidelity and quality. If it is specifically the tested program that is responsible for the observed effects, then the adjusted treatment group effects resulting from this analysis should be statistically non-significant. The logic of this analysis is that of a mediation test (Cohen, Cohen, West, & Aiken, 2003, p. 457) whereby the effect of treatment group membership on outcomes is presumed to be exerted indirectly via quality/fidelity of service[27]. Any residual treatment group effects are necessarily due to some other factor (e.g., uncontrolled differences between groups at the start of the project, another aspect of the program or some other source of bias). The inclusion of these analyses in the present report increases its length substantially, but in exchange it improves our understanding of and confidence in the reported results.

In the analyses reported in Section 5.1.3, each type of indicator was operationally defined at a global level and at a more detailed level. Two detailed estimates of fidelity were computed based on adherence to elements that are specific to program structure and program content respectively. The global estimate of fidelity was computed by simply taking the average of these two indicators. With respect to quality per se, we computed indices representing the structural, educative, and educator sensitivity dimensions. Again, a global quality index was computed by taking the average of these three quality indicators. A fourth quality indicator captured the nature of literacy activities in the classroom. It was treated independently of the others due to its theoretical and empirical importance.

The fidelity and quality indices were inserted into the analyses in the following way. First, we ran a series of analyses using the fidelity and quality indices as an indicator of treatment exposure instead of group membership. The purpose of this analysis was to verify whether a) the results obtained using the simple definition of treatment would be replicated and b) the more sensitive analysis might reveal effects that are stronger, more persistent, or both. In a second series of analyses, we re-introduced treatment group membership as an indicator of treatment exposure while controlling for the fidelity or quality indices. The purpose of this second test was to verify whether the simpler way of defining treatment is redundant with fidelity/quality in explaining variance in children outcomes. If the quality indices adequately capture the route by which treatment group membership exerts its effect on children outcomes, then we would expect

[27] A proper mediation test typically involves a more elaborate series of tests (Baron & Kenny, 1986; Frazier, Tix, & Barron, 2004). One of the central conditions for demonstrating mediation is that the direct effect of variable X (in this case treatment) should be eliminated when controlling for the mediator M (in this case quality/fidelity).

the magnitude of treatment effect estimates to drop substantially when daycare quality and fidelity are controlled for in analyses[28]. If such a result were obtained, it would provide additional support for the contention that the estimated treatment effects reported here are not due to some methodological artefact, but instead indicate a true impact of the tested daycare program.

3.2. STATISTICAL APPROACH

In this section, we present the analytical strategy that was employed in modeling the impact of the intervention. We begin by introducing the notion of multi-level sampling and clustered observations (see Section 3.2.1). This is followed by a description of the model-specification strategies that were used to estimate the treatment effects in various analyses (see Sections 3.2.2, 3.2.3, and 3.2.4). A detailed account of regression model specifications is presented along with the results they are associated with in Chapters 5 and 6.

3.2.1. Hierarchical Linear Modeling (HLM)

Data were analyzed using linear regression models. A fundamental assumption of standard regression models is that each observation or data point involved in the analysis has been independently observed. This assumption is violated when sampling units (e.g., daycares or children) contribute multiple observations to a dataset, in which case observations within a sampling unit are said to be nested or clustered. In a longitudinal design, observations are clustered within participants (i.e., each participant contributes multiple observations) and sometimes additional units of analysis as well. The Readiness to Learn project in particular has data that is nested both within daycares and participants.

The problem that clustering presents for standard regression analyses is that observations within the cluster are more similar to each other than they are to other observations. For example, knowledge of a child's score on a vocabulary test during the baseline testing period is much more useful for predicting that child's future score on a test than that of some other child. If ignored, the presence of clustering and the correlated residuals this implies can lead to an underestimation of the magnitude of standard errors, and by extension an exaggeration of statistical significance (Hox, 2002; Moulton, 1990).

One strategy for generating appropriate standard errors is to estimate a regression model specifying random effects for the units that are responsible for the clustering. In the econometric literature, clustering is also commonly managed by way of the so-called cluster-robust standard error estimator (Williams, 2000). Cluster-robust standard errors have the advantage of being relatively stable under many sample size configurations in the dimensions of time and space (i.e., evaluation period and daycare/child, Kézdi (2004), but see Donald and Lang (2007) for its poor performance with very small numbers of groups, e.g., group size of four or less). This way of estimating standard errors is a generalization for complex samples of the Huber-White heterogeneity-consistent estimator (White, 1980). The Huber-White estimator is widely recommended because it is 'robust' to heterogeneity of unspecified form, which is to say that it

[28] An obvious exception is the case where the two variables in question are involved in a suppression effect (Cohen,et al., 2003, pp.77–78, 457–458). Suppression effects within the context of these analyses are discussed when they have a bearing on the interpretation of the results.

produces standard error estimates that allow tests of statistical significance to be interpreted with confidence even when assumptions related to the independence and distribution of scores are violated. This way of estimating standard errors can over-correct standard errors by making them too large in some situations, but we elected to be conservative by exchanging "efficiency" for "consistency" (Hayes & Cai, 2007). The specific details of how the cluster-robust estimator was used in the impact analyses presented in this report are discussed in Section 4.3.

3.2.2. The Difference in Difference Estimator

A common strategy for estimating program effects in the econometric literature is the difference in differences (DinD) estimator (Abadie, 2005; Bertrand, Duflos, & Mulliainathan, 2004). The DinD estimator is a panel data technique that is appropriate for non-experimental repeated-measures research designs with both a pre-test "baseline" measurement and a comparison group. Readers who are unfamiliar with the econometric literature may be more familiar with the idea of an interaction or a moderated effect, of which the DinD estimator is a special case. The term DinD estimator applies to an interaction term involving two dummy coded variables, one representing the pre- and post-treatment testing periods (Pretest vs. Posttest) and the other representing the two groups to be compared (Intervention vs. Comparison group).

This estimator can be generalized to more complex scenarios by specifying multiple interaction terms of this type within the same regression model (e.g., to compare the pre-test to multiple post-test measurements in the same regression model) or to compare multiple groups. The impact analyses reported in this document used multiple DinD estimators to represent comparisons of three treatment groups across multiple post-intervention evaluations. The three groups were included in the same analysis so as to maximize the stability of statistical tests, thereby improving the odds of detecting real program effects.

As its name implies, the DinD estimator has two basic components. The first of these is an estimate of change (Δ) from the pre-test evaluation to a post-test evaluation for each group (Δintervention group = Posttest - Pretest; Δcomparison group = Posttest - Pretest). Here the pre-test measurement is the standard against which all subsequent evaluations are compared. In other words, all post-test measurements are always compared with the pre-test measurement when estimating program effects. Such change scores (i.e., differences) are not sufficient for isolating the treatment effect, however, because the change being estimated could have occurred as the result of natural developmental processes for example (i.e., maturation).

This ambiguity is addressed by the second component of the DinD estimator, which involves taking the difference between the change scores for the intervention and comparison group (DinD=Δintervention - Δcomparison). Here the change estimate provided by the comparison group is used to adjust estimates of change for the intervention group. The validity of the DinD estimator is based on the premise that once you eliminate the counterfactual estimate of change provided by the control group, all that remains is change due to the treatment effect. As with most statistics, the assumptions that support this premise are likely to be violated in practice. This is especially true within the context of a non-experimental field study, of which the Readiness to Learn project is an example.

The validity of the DinD estimator depends on at least two assumptions. With respect to the change scores, we assume that the composition of the groups is constant for all evaluation periods being compared, which may not be the case if participants migrate from one group to

another over time. Such group changing is an issue in the Readiness to Learn project and it must therefore be controlled. For the comparison of change scores, we assume that in the absence of an intervention the groups being compared would have developed in an identical way. There is of course no way to verify this assumption directly, but initial differences between groups on characteristics that are related to the dependent measure suggest the possibility of non-parallel slopes. The group differences that were identified in the *Reference Report* (Legault et al., 2014) are an example of potential sources of this type of bias.

We minimized these threats to the validity of the DinD estimator by including relevant covariates in our analyses which served to adjust for group differences in initial state and in developmental trajectory in the absence of treatment. Variables were included selectively in regression models based on whether doing so would yield a material difference in the reported pattern of results. The details of this strategy are discussed in Section 3.3.3.

3.2.3. Statistical Control in the Context of a Longitudinal Design

The DinD estimator effectively neutralizes the static effect of observed and unobserved differences among the treatment groups at the baseline period[29]. By construction, such differences are eliminated by way of the initial differencing of Posttest and Pretest. Nevertheless, as discussed in the preceding section, the DinD estimator can be biased when group composition changes over time and when Pre-test differences in the developmental trajectories of the groups being compared are suspected. The strategies employed for managing both types of bias are discussed in turn.

Group Composition Changes

As discussed earlier, changes in group composition can bias the DinD estimate. Such bias can occur when some characteristic that is related to an outcome varies over time for a group on average. For example, gender is known to be related to several outcomes in the developmental literature. If the proportion of girls in a group suddenly increases prior to a post-test measurement, the difference calculations required for the calculation of the DinD estimator will be biased.

This threat to statistical validity was controlled in two ways. First, changes in group composition were monitored over time. For example, we verified whether group changing behaviour was associated with our baseline covariates or with treatment condition. No such association was observed (results available upon request). We further verified whether there was a statistically significant association between treatment group and attrition (see Section 4.2.2). Again, no such association was observed. In sum, our preliminary analyses failed to indicate important problems with group composition changes. Nevertheless, we employed a second strategy whereby we included as covariates all non-redundant baseline variables that were associated significantly with an outcome. The purpose of including such covariates was to maximize the validity of the DinD estimator[30]. We assume that the set of covariates included in

[29] The static effect of a baseline characteristic refers to an association with outcomes that is stable in time. An example would be the fact that the average effect of gender is of comparable magnitude across all evaluation periods for the first year. If the effect of gender where dynamic, the strength of its association with an outcome would vary significantly depending on the evaluation period considered.

[30] In this case, the DinD estimator represents the estimated average treatment effect of our intervention conditional on the covariates that are included in the model. We assume that treatment condition does not moderate the effect of the covariates.

our model has adequately compensated for any bias arising from the selection of participants to treatment groups.

Parallel Developmental Slopes

Even when group membership is constant over time, initial differences between groups can be associated with non-parallel developmental trends. As discussed above, non-parallel developmental trends (in the absence of treatment) are a threat to the validity of the DinD estimator. If the cause of such non-parallelism is observed (i.e., data on an appropriate covariate has been collected), it is possible to perform statistical adjustments that will correct for such bias (Abadie, 2005). An example of such an adjustment is to include as a covariate the interaction between time and the relevant variable. Variables that might be considered for use in performing such an adjustment include exogenous baseline covariates which are: a) related to the outcome measure, and b) distributed differently across the groups to be compared. The practical limits imposed by our sample size meant that we had to be selective in choosing covariates for use in such an adjustment to avoid over-fitting the data. Thus, we only included Time by Covariate interaction terms if they were statistically significant predictors of an outcome (any outcome) for our sample (see Section 4.4).

Validity of Estimated Dosage and Quality Effects

In the preceding discussion, treatment exposure or "dosage" and quality were presented as alternatives to simple treatment group membership for grading participants along the continuum from untreated to treated. Effect estimates that are based on these more precise definitions may be more efficient, but they are subject to the same limitations as those based on simple treatment group membership. Specifically, we refer here to the assumptions upon which the validity of our treatment effect estimates are based: that the treatment conditions, in the absence of any new investment on the part of various levels of government, would have been comparable.

The focus of the preliminary analysis was on establishing the comparability of the treatment groups. If we are to compare children with varying levels of dosage within and across these groups (i.e., estimate the effect of hours spent in daycare), then we must similarly assume that these children are otherwise comparable. For instance, we assume that children who average 40 hours a week in daycare are equivalent across all other relevant characteristics to children who average only 10 hours a week in daycare. This strong assumption is unlikely to be met given that a variety of socio-demographic variables are potentially related to both the amount of time spent in daycare as well as the outcome variables. We also assume that the program- and comparison-group daycares would have been equivalent on estimates of fidelity and quality in the absence of intervention. This assumption was not verified empirically (i.e., no true pretest measures were taken of these dimensions) and is unlikely to be (perfectly) met given the small number of daycares and the impossibility of matching on these variables.

We used two strategies for addressing these potential sources of bias in the estimated effect of dosage and quality. First, the longitudinal design of the present study allows the use of the DinD estimator, which neutralizes the static effect of baseline characteristics (both observed and unobserved). Second, the final specification that served to estimate the treatment effects included all non-redundant covariates that were related to at least one of the outcome measures (see Section 4.4). If we assume that all the relevant variables associated with varying levels of daycare exposure (or fidelity/quality) have been adequately accounted for using this strategy

(i.e., that the model has been correctly specified), then the resulting conditional estimates of the dosage effects are unbiased. The same strategies and assumptions were enlisted in estimating the effect of program fidelity and quality[31].

3.2.4. Family Literacy Workshops: A Special Case

The analysis of the Family Literacy Workshop data has to be set aside as a special case within the broader context of the analyses that figure in this report. Only two parental scales were present in each follow-up survey: frequency of literacy activities and the language of literacy activities. These scales were analyzed in a similar manner to those of the ÉPE–AD (i.e., via the DinD estimator). For the remaining outcomes (i.e., the Knowledge, Self-efficacy, and Modeling scales), data were available only for parents whose children were enrolled in one of the program daycares. Henceforth this group of scales will be referred to as the Parental Workshop scales. For the analysis of the Parental Workshop scales, a slightly different strategy was adopted which will now be described.

Parental Workshop Scales: Choice of Sample

The first distinguishing feature of the analysis of these parental subscales is that they are based on the full sample of participants (all six communities) who were enrolled in a program daycare at the time the workshops were offered (N = 105). Unlike the analysis of child outcomes, the communities of Edmonton and St-Jean were not excluded from this analysis. This decision was justified on three grounds. First, unlike the daycare intervention, the Family Literacy Workshops were correctly implemented in all communities (see the *Project Implementation Report*, Bérubé et al., 2014). Second, since the analysis involves only parents whose children were enrolled in one of the program daycares, the availability of a daycare control group in all communities is not an issue. Finally, the exclusion of the communities of St-Jean and Edmonton would reduce the sample to a level where a fair test of the workshop impacts was not possible. The characteristics of the full sample of program group families are not reported here (see Table 2.5 for response rates), but they are described in detail in the *Reference Report* (Legault et al., 2014).

Parental Workshop Scales: Analytical Strategy

The second feature that distinguishes the analysis of the Parental Workshop Scales is the absence of a planned comparison group. These measures were only administered to parents whose children were enrolled in one of the daycares that participated in the intervention, and data are not available on some measures for families who participated in fewer than three workshops (n = 15). A balanced design would have a complete set of pre-test and post-test measurements on all variables for all participants, but this was not the case here. For reasons of efficiency, an unbalanced design was employed requiring a more selective data collection method.

Specifically, the families classified as ‘non-participants’ were only asked to provide post-test estimates of their attitudes, which means that retrospective estimates (Lamb & Tschillard, 2005; Rockwell & Kohn, 1989) are not available for this group. A complete set of post-test and retrospective pre-test data were available for the Knowledge and Self-efficacy scales for those

[31] All covariates are at the level of ‘participant’ rather than daycare except for the variable Community. The number of daycares was insufficient to allow the inclusion of a broad set of characteristics at this level in the analysis.

families classified as "participants." True pre-test and post-test estimates were collected for the Modeling scale of all families regardless of participation. The analytical approach used in the associated analyses was adapted to the design's complications.

Given the structure of the dataset, at least two ways of estimating the impact of the workshops are available. The first method would estimate the impact of the workshops as a function of the difference between the pre-test and the post-test measures. The second way of estimating the impact of the workshops involves comparing those parents who were classified as participants (n = 90) to those that did not (n = 15). Each type of comparison on its own lacks validity, but the global pattern produced by a series of tests can be informative.

The change scores (difference between pre-test and post-test measures only for the participants) are of dubious validity not because they are retrospective (for a discussion of the advantages of retrospective estimates over traditional pre-tests, see Lamb & Tschillard 2003; 2005), but rather because of the lack of a comparison group to control for maturation and history effects. This validity issue is further compounded by the fact that participants are aware of being treated and so could produce estimates of their pre-test and post-test state that confirm their expectation that attending the workshops should have positive effect[32]. Similarly, the comparison involving participants and non-participants is invalidated by an obvious potential for selection bias, which cannot be controlled either statistically or via matching because of the prohibitively small number of participants in the non-participant group.

Individually, the two treatment effect estimates are biased, but together they can produce informative results without completely ruling out all sources of bias. For example, we contend that the following pattern of results is consistent with a true treatment effect: (a) the treatment effect based on the change score for participants is statistically significant and positive, (b) the retrospective pre-test estimate for the participants is equivalent to the 'post-test' estimate provided by the non-participants, and (c) the participants report post-workshop estimates that are superior to those of the non-participants on average. We assume that because the non-participants did not participate in the program their "post-test" estimate is equally valid as an estimate of their pre-test state. In other words, we make the assumption that their true state, on average, did not change during the period over which the workshops were offered. If the predicted pattern of results is obtained, it suggests that the workshops had a positive effect. If only result (b) is obtained, it suggests that the effect of the treatment is null (or negative). If the results (a) and (c) are obtained without result (b), it strongly suggests that the estimates are biased and the interpretation of positive program effects should be appropriately qualified. In other words, there is a specific pattern of results that provides strong support for a true positive impact of the Family Literacy Workshops, and many other patterns that fail to provide such support. If our "risky" prediction is confirmed, it lends credibility to our interpretation of the effects as valid.

[32] This kind of bias could take many forms and it is not specific to retrospective pre-tests. Parents could exaggerate the program effect by underestimating their pre-test state and by overestimating their post-test state, or by producing estimates that are contaminated in both ways. There is some evidence that an underestimate of the pre-test state is typical of retrospective measures (Taylor, Russ-Eft, & Taylor, 2009). Traditional prospective pre-test measures are biased in other ways (e.g., response-shift bias; Howard, 1980), which in some cases results in an overestimate of the pre-test state (Moore & Tananis, 2009).

The limitations imposed by the available data mean that an estimator such as DinD can be ruled out because of the lack of a pre-test measure for both groups. Instead, the three hypotheses formulated above can be tested using a series of independent tests, one for each hypothesis. Covariates were not used in the test of group differences by reason of the prohibitively small sample size of the non-participant group (n = 15). Further, covariates were not used in the estimation of the change scores for participants because: a) the static effect of participant characteristics is neutralized ("differenced out") in the process of this estimation; and b) changes in group composition are not an issue.

Estimates of program fidelity, program quality and dosage (proportion of workshops attended) were available for the sample of parents who participated in the workshops (n = 90). The impact of these factors was estimated by way of a DinD estimator as in the other analyses reported here. The resulting estimates represent the workshop impacts in terms of the intensity of the treatment received. Again, since group composition changes were not an issue for this analysis, we relied on the fact that the DinD estimator is unbiased by both the observed and unobserved baseline characteristics.

4. Preliminary Analyses

Prior to performing the impact analyses, a variety of preliminary quality checks and analyses were conducted. The purpose of this chapter is to provide an overview of the preliminary steps that were taken in preparing the data for analysis. We describe the quality control processes that were put into place for the data collection in Section 4.1 and the results of the missing-values analysis and data imputation in Section 4.2. From there, analyses immediately relevant to conducting the impact analyses are described. The process that determined decisions related to error-term specification and the selection of covariates are described in Sections 4.3 and 4.4 respectively. Finally, the issue of external validity is addressed in Section 4.5 where the characteristics of the present sample are compared with the characteristics of a sample that was designed to be representative of the target population (Survey on the Vitality of Official-Language Minorities, henceforth SVOLM).

4.1. QUALITY CONTROL PROCESSES

Ensuring the quality of the data collected comprises multiple steps ranging from instrument selection or conception, to data collection, up to data entry. The selection or design ***of measuring tools*** (e.g., survey, interview grid, field observation grid, etc.) is a first step in producing a "clean" dataset. Whenever possible, the Readiness to Learn project team selected pre-existing scales that have been tested and validated. When such measures were unavailable, the Readiness to Learn project research team developed new instruments (e.g., scales, observation grids or interview grids) by applying solid psychometric principles. For instance, care was taken to ensure that the importance of various sources of measurement error was minimized. Questions were written in a clear, precise and simple language. Among other considerations, the format of measurement instruments and questions were conceived so as to avoid placing an excessive burden on the recall memory of participants. All measurement tools were pilot-tested prior to their use in the field. The battery of pilot-testing control procedures included asking a group of team members to validate: a) the wording of questions for their clarity (in English and French); b) the consistency of questions across languages in terms of both form and content; c) the logical flow of the items; and d) the appropriateness of response options.

For the child outcome measure (i.e., the ÉPE–AD) specifically, a pilot test was conducted in the Spring of 2007 with a sample of children. The purpose of the pilot test was to address three important concerns: (1) to determine if the measure could be used with children as young as those targeted by the Readiness to Learn project (which were younger by several months than the original sample that was used to fine-tune the ÉPE–AD when it was initially constructed by the test developer; (2) to determine if the measure could be used with bilingual children (which is a sub-population that was not explicitly represented in the sample employed by the test developer); and (3) to gain field experience with administering the measure under different conditions and in different locales (e.g., home, child care, etc.). In sum, SRDC took proactive steps for minimizing various sources of measurement error that can arise in these circumstances. These steps were considered essential given the relative novelty of the measurement tool and the special circumstances of its use within the context of the Readiness to Learn project.

High standards of data collection quality by research personnel were ensured by developing ***tool-specific technical material and instruction manuals***. These materials were distributed to ***all relevant personnel, who were also trained in the practical administration of the measurement instruments***. Relevant personnel included the interviewers, evaluators, and field observers. The training of all personnel was updated prior to each testing period. For instance, evaluators received ongoing training which was scheduled prior to every evaluation. These training sessions served to refresh evaluators on the administration of the ÉPE–AD and to introduce them to new child outcome measures as these were added to the battery of instruments. For their part, the interviewers who administered the parental surveys were briefed on the nature and purpose of the survey questions prior to each data collection wave. The ongoing training of the research staff ensured consistency across time in use of the measures as well as consistency across data collection staff. All questions and concerns on the part of the data-collection staff were addressed during these training or/briefing sessions. The Readiness to Learn project research team closely monitored the progress of the data collection team, providing support whenever needed by supplying techniques, strategies and assistance to the data collection staff. In some cases, the data collection plan was adjusted in response to the performance of the measures in the field.

A ***first control of the data quality*** was put in place at the outset of the Readiness to Learn project. The raw data were collected in two steps. Community coordinators served as the initial hub, receiving the data collected in their respective communities. Once the information was collected, it was sent to the SRDC Ottawa office. Accordingly, the community coordinators were responsible for a first verification of the quality of the data, which ensured that timely feedback was given to the data collection staff in the field. The alternative would have been to wait until the data were received by the staff at SRDC, which would have delayed feedback by several days due to the time requires to transfer the information. The community coordinators ensured that missing data was minimal in the parent surveys and evaluations by verifying whether they were completed properly. Where errors or omissions were found, interviewers or evaluators were asked to retrieve the missing information by contacting the parent or by completing the child's evaluation. Community coordinators were instrumental in ensuring the high response rates and retention rates observed in the Readiness to Learn project.

A ***second control of the data quality*** was enacted at the Ottawa office to verify the quality of the data entered and the psychometric properties of the measurement scales. Data were entered into an electronic database and then submitted to a rigorous verification system to ensure accuracy. In a first step, a random check of 10% of the data was conducted to verify accuracy of data entry. Next, descriptive analyses were conducted to verify if item frequencies fell within the expected range. An extensive series of crosschecks were conducted based on the electronic databases to ensure that responses were consistent within respondents both for individual surveys and across time. Inconsistent or implausible values were verified with the paper copy or the data collector. Finally, statistical methods were used to confirm the quality of the scales. The internal consistency of scales and the validity of the measured dimensions were verified respectively using Cronbach alpha and factor analysis. The construct validity of the measured variables constructs was verified by examining whether the direction of the observed inter-correlations among them was consistent with expectations. Lastly, the data were routinely screened for univariate and multivariate outliers prior to analysis. All collected survey data were subjected to imputation of missing values, according to accepted procedures (Cohen, et al., 2003; pp., 431–451; Tabachnick & Fidell, 2006, pp. 62–71).

A ***third control of the data quality*** consisted of the use of a mixed methods approach. The Readiness to Learn project used a diversity of tools, both quantitative and qualitative, and multiple sources of data chosen in function of the research objectives. Each of these sources of data served to triangulate research findings further bolstering our confidence in conclusions. In addition, the complementarity of the collective findings leads to a more complete and nuanced understanding of the phenomenon under study.

Lastly, ***consistency of scale composition*** across time was essential for preserving the internal consistency of the impact analyses. As such, verification was done to confirm that scales repeated over time comprised the same items and identical response choices. Where the item composition of scales varied over time (e.g., as in the case of children's outcomes), standardization of scores within each version of the ÉPE–AD was undertaken to render the measurements comparable across time, and therefore amenable to the planned analyses of program impacts (see Section 5.1.2)[33].

4.2. MISSING VALUES ANALYSIS

Missing values in a database can threaten the validity of an analysis. This threat arises from two principal sources. First, cases with missing values on a variable are typically excluded from an analysis. The exclusion of cases with missing values can be problematic when they differ in important ways from those that remain in the analysis. If this is the case, the composition of the sample changes, which has implications for the external validity of the results. For example, if missing values occur disproportionately for girls, then the results that are obtained may not generalize to this population. Second, the internal validity of the estimates of the program effect depends on the stability of group composition in time. If missing values disrupt group composition, then this may bias estimates of the program effect. In what follows, we present an analysis of missing values that evaluates the potential for both kinds of bias.

4.2.1. General Breakdown of the Missing Values

Missing values in the database can be divided into two main types: planned and unplanned. Planned missing values only affected the ÉPE–AD scales, both French and English versions, for the 2nd, 3rd, and 5th evaluation periods. These missing values were the result of the application of start rules for the administration of certain subscales. No children met the condition for applying the start rule for the 1st evaluation; they were all under the age of four.

Unplanned missing values occurred for both the measures administered to children and those administered to adults. The unplanned missing were of two types: a) the result of a failure to collect any information from the participant at a given testing period (complete missing) and b) the result of a process whereby a participant failed to answer a particular question or was not administered a particular scale (partial missing)[34]. We now describe in turn the frequency of occurrence of these two types of missing values.

[33] The term standardization is used here to refer to the statistical transformation of the raw scores (x) of a variable into z-scores, where z = (x- average score) / standard deviation. All standardized variables have a mean of zero and a standard deviation of 1.

[34] Partial missing can arise due to various causes such as an error on the part of the test administrator or, in the case of ÉPE–AD evaluations, use of different languages for some subscales which cannot subsequently be combined for analysis.

Planned Missing Values

All planned missing values resulted in partial missing values. An entire evaluation or scale was never intentionally skipped. Easy items at the beginning of ÉPE–AD scales were intentionally skipped for some children due to the application of the start rule prescribed by the test developer. The purpose of the start rule is to increase the efficiency of the evaluation process and to minimize measurement error due to the boredom or fatigue of children. The assumption that supports the use of a start rule is that the child would have obtained the maximum score on the omitted items had they been administered. If this assumption holds, then the putative missing values are replaced with the maximum possible score without any loss of information. However, it is only likely to be met with a performance measure when the items of a scale are presented in their order of difficulty and when a strong predictor of child performance is used to determine when to apply the start rule.

For the second and third administrations of the ÉPE–AD (February 2008, June 2008 respectively), suitable candidates for the application of the start rules were identified by their age as per ÉPE–AD guidelines. Children older than four years were to skip the first six items of the Communication and Self-Awareness domains and the first eight items of the Cognitive Ability and Physical Ability domains. A rule was in place whereby children older than four who did not achieve a score of either 3 or 4 on the two first questions had to return to the beginning of the scale. In practice, this so-called “go back rule” was not always respected. Even in cases where the evaluator correctly decided against applying the go back rule, the child would not necessarily have received the maximum score on skipped items (i.e., 4) had they been administered. In sum, the application of the start rule created missing values for items at the beginning of certain scales. For the second and third administrations of the ÉPE–AD, the number of children with missing values due to the start rule was 6 and 42 respectively.

Thus, it was necessary to estimate what value the missing values would have taken had the items been administered, but the method used to accomplish this task remained to be determined. As noted above, the test developer recommended that start-rule missing values be replaced by the maximum possible score the child could have obtained. This plan was abandoned based on two arguments. First, the items of the ÉPE–AD subscales were manifestly not presented in order of ascending difficulty. Preliminary analyses conducted by SRDC suggested that some easier items were presented later than more difficult items. This observation is corroborated by the fact that more recent versions of the ÉPE–AD have a reworked item order. The second argument is based on the fact that the age-based cut-off was not a sufficiently reliable predictor of the outcome measures to predict with certainty that the child would have obtained the maximum value on the start rule items. This, combined with errors in the application of the go back rule, meant that the maximum item score did not represent an adequate approximation of a child’s true score. This being the case, the direct imputation of planned missing values on the ÉPE–AD with the maximum item score was abandoned in favour of a strategy that would more adequately capture variability in performance at the individual level. This strategy is discussed in more detail in the Section 4.2.3, Data Imputation Strategy.

For the fifth evaluation period (and final administration of the initial version of the ÉPE–AD, February 2009), a start rule was again applied this time based on the performance of children on the preceding evaluation. It was thought that capitalizing on data from previous evaluations would improve the precision of the start rule, minimizing any loss of information. The start rule

was applied only if the child received a score of 10 or better on the first four items of the Communication scales for the previous evaluation and a score of 21 or better on the first seven items of the Self-Awareness scale; Its application resulted in children skipping the first three items of the Communication scale and the first six items of the Self-Awareness scale in their subsequent evaluation to take place during the fifth evaluation period. An additional constraint was that the language of administration had to be the same in the fourth and fifth evaluations[35]. This condition was based on the fact that no solid evidence exists that the French and English version of the ÉPE–AD are inter-changeable[36]. In the absence of such evidence, we considered it more prudent to avoid mixing data from the two languages of evaluation for any purpose. A list of children who could potentially skip the early items of the ÉPE–AD scales was prepared in advance of data collection for 5th evaluation, which was comprised of 188 of the 240 children who were still participating in the project at that time and for whom language of administration remained constant.

The major source of missing values on the ÉPE–AD at the level of individual items was the implementation of the start rules, but a handful ($n < 5$) of unplanned missing values was also observed across all items and time periods of the ÉPE–AD. Given the trivially small number of such missing values and the fact that they occurred at the item level, we did not subject these to a more detailed analysis and they were treated like start-rule missing for the purpose of imputation.

Unplanned Missing Values

Unplanned missing values affected both the parental surveys and the evaluation of children. These missing values could result either from a missing survey or evaluation. In addition, survey data were sometimes incomplete which means that valid data were available for some questions but not for others, either because of interviewer error or the participant's refusal to answer[37]. Of the two types of missing values, the latter is potentially more problematic because the mechanism that generated them could be a function of the question itself[38]. We describe in what follows the sources of missing values described above and their implications for the choice of data imputation/replacement strategy.

A small percentage of participants were classified as having complete missing data for each testing period. Complete missing data were caused by either a failure to administer the measures or by participant attrition (see the method in Chapter 6). Missing data due to attrition is problematic in that it limits options with respect to data imputation in a longitudinal design. For instance, in a longitudinal design it is possible to pose the same question on repeated occasions to improve the odds of collecting the desired information for all participants. To take an example from the current project, questions on family revenue where posed three times over the course of

[35] For 12 cases, this condition was not respected. The scores that the participants would have obtained on these items were estimated using imputation in a manner similar to that used for the second and third evaluations. See in Section 5.2.3, Data Imputation Strategy.

[36] Unreported preliminary analyses demonstrated that children who switched languages from one evaluation to another produced change scores that were significantly different from the other children in the sample. Further, a content analysis of the French and English versions of the ÉPE–AD indicated that the French translation was not well adapted. In the absence of positive evidence to the contrary, we concluded that the two version of the ÉPE–AD should not be combined.

[37] This type of partial missing value was not observed for the French version of the ÉPE–AD scales. All analyses reported in this document are based on the French version of the ÉPE–AD, therefore the issue of partial missing values is not considered further for these outcome measures.

[38] The textbook example that is typically used to illustrate is the case where questions are posed regarding family income.

the two first years of the study[39]. For participants who remained with the project up until the eighth survey, it is possible to fully exploit such redundancy by cross-imputing missing values for such repeated questions when they occur[40]. Given these particularities, isolated cases of "complete" missing data were handled differently from cases of attrition in the analysis and management of missing data. More details are provided in the Section 5.2.3, Data Imputation Strategy.

An additional source of missing values in the database was surveys with incomplete data. As noted above, this type of missing data can pose problems for data analysis. The main issue concerns whether some property of the question posed systematically increases the likelihood of a given participant not answering the question. For example, it is well known that respondents from the extreme ends of the income distribution are more likely to resist answering questions related to income. Clearly, in this example, the process that is generating the missing values is systematic. If the mechanism generating the missing values is systematic (not random), then exclusion of cases with missing values is likely to meaningfully alter the characteristics of the sample.

An example of such a mechanism is the rule used for deciding the language of administration of the ÉPE–AD. This mechanism causes missing values for the French versions of all scales except Communication for children with the weakest French-language skills. Consequently, the reduced sample is stronger in French than the original, which has the effect of compressing the range of this dimension, thereby attenuating potential program impacts on variables related to mastery of French. Other mechanisms responsible for creating missing values can create similar types of bias. The analyses presented in Section 4.2.2 concern themselves with describing the pattern of missing values and taking steps to limit the consequences for the validity of the study.

4.2.2. Pattern of the Missing Data

Missing values are like any other outcome in that it is possible to model the process that generated them. The conclusions arising from this modeling exercise determine what steps are taken to preserve the internal validity and external validity of the study. If the process that generated the missing values is random, then validity remains unthreatened. If, however, the process is non-random, then steps must be taken to avoid introducing bias by excluding cases that have been self-selected. We now describe the modeling strategy that was used to elucidate the pattern of missing data in the project database.

According to Little and Rubin (1987), unplanned missing values may be classified into three types: a) missing completely at random (MCAR), b) missing at random (MAR) and c) not missing at random (non-ignorable MNAR). The most felicitous scenario is a situation where a small number of missing values is randomly distributed throughout a dataset (i.e., MCAR). In contrast, the worst case scenario would involve a large number of non-randomly distributed missing data. In the first case, the problem of missing values can be accommodated by a strategy of listwise deletion without biasing the results of an analysis (i.e., deletion of cases with missing

[39] This question was posed at the baseline survey, and during the seventh and eighth follow-up parent surveys (June and October 2009 respectively).

[40] We must additionally assume that the variable being measured is unlikely to change systematically over time or be affected by the treatment. We consider that family income is a prime example of a variable that can fluctuate over time but which is unlikely to show a systematic relationship with either time or the assignment of participants to treatment groups.

values). Treating missing values that are non-randomly distributed in this way can introduce bias in the estimates of treatment effects. This bias can either be avoided by a) excluding the offending variable(s) from the analysis, b) employing a data imputation strategy, or c) it can be accommodated by describing the nature of the bias and interpreting the results in this light (e.g., missing values were more prevalent among population X, therefore results based on the remaining cases may not apply to this population). In this report, all three strategies are used where appropriate.

The first step in this decision process is to assess the prevalence of missing values in the database (i.e., the percentage of missing values across all variables in the database) and the pattern of these missing values. The prevalence of missing values is evaluated by way of basic descriptive analyses which require no additional explanation. In contrast, the method used to determine the pattern of missing data requires more explication.

The most basic question concerns whether the missing values are predictable or not. As noted earlier, unpredictable or "random" missing values are classified as MCAR. A necessary condition for demonstrating MCAR is to show that the relationship between the occurrence of missing values and the observed values of variables in a database is statistically null. This property can be verified by conducting independent tests of association for each variable in a database. In the case of continuous variables, an alternative is available in the form of Little's MCAR test. If the MCAR test or another test of association indicates that the missing values are systematic, then the working hypothesis of an MCAR distribution must be rejected in favour of a less restrictive assumption: the distribution of missing values is actually MAR.

For the pattern of missing values to be considered MAR, two conditions must be met: (a) the missing values are non-randomly distributed and (b) the value that the missing value should take should be predictable. The first condition is satisfied when variable(s) in the database allow the occurrence of missing values to be predicted; the latter is satisfied when strong predictors of the measured variable with the missing values are available in the database. If the first condition is met and the latter is not, then the distribution of missing data is assumed to be MNAR. Options in this situation are limited to describing the pattern of missing values in detail so that limits to the validity of the study are understood.

The issues that are at stake in the assessment of missing values have now been reviewed. In the following section, we report the results of the missing-values analysis conducted in preparation for the Readiness to Learn project impact analyses. We report first the results of a quantitative analysis of participant attrition. Then, the result of an analysis of complete missing values due to a failure to evaluate a child or receive a survey is presented. Finally, we present an analysis of missing values caused by the language of administration of the ÉPE–AD. In each analysis, we describe the missing values at two levels: first as a function of their distribution across treatment groups and then as a function of their relationship to other covariates in the database.

Missing Values Due to Attrition

The incidents of attrition are described in Section 2.4.1 of Chapter 2 (Method). A total of 14 families withdrew from the study by the end of the second year. By the end of the first year, 10 families had left the study, which represents an attrition rate of 3.9% relative to the original sample of 254 that was recruited from the four communities retained for the impact analyses. By

the second year, the attrition rate increased to 5.5% of the four-community sample. This percentage is low relative to other evaluation studies (e.g., Rogers, Fernandez, Thurber, & Smitley, 2004) with a correspondingly negligible potential for bias.

The infrequency of participant attrition prohibited a formal test of whether participant retention was significantly related to treatment group membership. However, a descriptive analysis of the frequencies does not hint at a systematic relationship. In total, five families from the Program Daycare group, two families from the Comparison Daycare group, and four families from the Informal Care group left the study. The remaining three participants left the study prior to the baseline evaluation and were therefore excluded from the analysis by treatment groups. A similar analysis failed to reveal a systematic relationship with the linguistic composition of the home. The 14 cases of attrition were spread relatively evenly across the categories: endogamous Francophone (5), endogamous Anglophone (2), exogamous (5) and bilingual (1). The gender split was similarly even, eight boys and six girls.

For variables that were measured at the rank level of measurement or higher, we computed Kendall's tau (τ) rank-based correlation coefficients to identify the characteristics associated with attrition (for a list of covariates, see Section 5.4). This analysis revealed statistically significant but weak relationships between attrition and the following variables: baseline Communication score (N = 251, τ = -.125), language of literacy activities (N = 254, τ = -.116), the language spoken by the mother to the child (N = 251, τ = -.139), and the language of care for the child from 0 to 12 months (N = 244, τ = -.140). We report the descriptive statistics associated with these variables in Table 4.1 for the attrition cases and in Table 4.4 for the cases that were retained.

The pattern of results in both tables indicates that families and children who scores lower on use of the French language were more likely to leave the project. The low attrition rate and even distribution across treatment group means that, practically speaking, the threat to the internal and external validity of the study is minimal.

Table 4.1: Descriptive Statistics for Attriton Cases on Variables that Correlate with Attrition

Measured Construct (range)	Mean	SD	n
Baseline Communication Score (raw score)	10.00	8.69	11
Language of Literacy Activities (1–5)	3.08	1.45	14
Language Spoken to the Child (Mother) (1–3)	2.14	0.77	14
Language of Care (0–12 months) (1–3)	2.21	0.80	14

Missing Values Due to Complete Missing (Evaluation or Survey)

The overall response rates for both the surveys and the ÉPE–AD are reported in Section2.5.1 of Chapter 3 (Method). Below, we discuss the pattern of missing values for the evaluation and then, we discuss the pattern of missing data for the surveys. The information presented is based on an analysis for the data from the first year of the project. The sample excludes those children who left the project by the end of the first year.

Children's evaluations: The overall response rate was quite high for all evaluations, which means that instances of complete missing were too infrequent to permit a test of how such missing values are related to treatment group membership (cell size < 5). We note based on a descriptive analysis that the missing values were evenly distributed across the treatment groups regardless of which evaluation is considered, with three or fewer missing evaluations per group (unreported data). Table 4.2 presents the distribution of missing values collapsed over the first four evaluation periods.

A variety of behavioural child development outcomes, socio-demographic, linguistic, socio-linguistic, and parental variables were used in an attempt to predict the pattern of missing evaluations. The analysis based on behavioural child development indicators (imputed as described in Section 4.2.3) revealed that children were more likely to have skipped at least one evaluation during the first year if they scored low, during the first evaluation, on the Communication (N = 244, τ = -.17), the Expressive vocabulary (N = 228, τ = -.14), or the Receptive vocabulary scales (N = 229, τ = -.14). Further, an analysis using variables from the baseline parental surveys indicated that children from single-parent families (N = 244, τ = -.14), those with lower reported family income (N = 244, τ = -.19), with lower maternal education (N = 244, τ = -.12) and those who reported less identification with the francophone community (N = 243, τ = -.13) were more likely to skip at least one evaluation. Thus, there is evidence to suggest that the pattern generating these missing data (i.e., missing evaluations) is systematic. This issue was easily corrected because the same instruments were administered multiple times to the children, providing very strong predictors of the unobserved values. In sum, these missing data were distributed as MAR, which justifies replacement of the missing values by way of imputation.

Table 4.2: Number of Missing Evaluations over the First Year of the Project as a Function of Treatment Group Membership

Missing Evaluations	Program Daycare	Comparison Daycare	Informal Care
None	70	89	72
1 missing	2	6	3
2 missing	0	1	1
% missing	2.8%	7.8%	5.6%

Note: Missing evaluations over the first year of the study as a function of baseline treatment group membership. For any given time period, the % missing per group is < 1%.

Parental surveys: Complete missing data is only an issue for the follow-up parental surveys. As reported in Section 2.4.1, all parental surveys for the baseline period were received. For the follow-up surveys, the number of missing surveys was quite small for each time period (as it was with the evaluations), which again limited the type of analyses that could be conducted in describing the distribution of missing values (e.g., cell-sizes under 5 when a grouping variable is used).

As with the evaluations, we limit ourselves to a simple description of the missing values as a function of treatment group. Taking baseline treatment group membership as a point of reference, we observe eight missing surveys for the Informal care group, four missing surveys for the Program Daycare group, and three missing surveys for the Comparison Daycare group

over the first year of the project. Because the primary purpose of the surveys was to track the evolution of variables that are reasonably stable in time (e.g., family composition), it was possible to infer by simple linear interpolation a reasonable estimate of the value the unobserved score would have taken. This procedure could not be applied for all variables, which means that some of the complete missing cases were classified as having partial missing values following such this stage of imputation (see below).

Missing Values Due to Language of Administration (Evaluation Only)

An additional source of missing data for the French version of the ÉPE–AD was the language of administration of the test. All scales except Communication, which had to be administered in French to all children (and then English for some), were affected. This type of missing value was obviously non-random because it is a direct function of a decision rule based on the child's competency in using French. The only issue that remains to be determined is whether the missing data is systematically associated with treatment group.

The total number of children who completed the remaining ÉPE–AD scales in English was 28, 31, 27, 20, and 44 for evaluations 1 through 5. Thereafter, all evaluations were conducted in French. Chi-square tests indicated that the choice of language of administration did not vary significantly as a function of treatment group membership for any evaluation period at the .05 alpha level (the full set of results is available upon request).

Previously, SRDC decided that the French and English versions of the ÉPE–AD could not be combined for the purpose of analysis. Due to the small number of observations in English, we elected to analyze only the scores from the French version of the test. The external validity of the experiment may be affect by this decision given that it yields a sample whose range of French-language competency is restricted (i.e., due to truncation of the lower end of the distribution). The obtained pattern of results for the affected scales may therefore only be replicated with a sample that similarly under-represents the lower end of the ability spectrum. The threat to internal validity is minimal given that such missing values are evenly distributed across the treatment groups. Nevertheless, the missing values were imputed whenever possible to minimize the potential for bias (see Section 4.2.3).

Missing Values Due to Partial Missing (Survey Only)

Some survey questions were posed multiple times over the course of the study. This redundancy allowed many missing values to be directly and accurately estimated, resulting in an important reduction in the number of missing values in the database. The strategy is described in greater detail in Section 5.2.3. Despite these steps, many of the 21 variables that were selected for the impact analyses still had missing values (for a complete list, see Section 4.4). We describe the pattern of these missing data in more detail in this section.

The missing data rate for only a single variable exceeded 5%: 14.3% (N = 244) for Vitality (i.e., the respondent's perceived vitality of francophone culture in the community)[41]. No other variable had more than 2% missing. Baseline Communication score was included along with the survey variables in a formal test of the pattern of missing data. In this analysis, treatment group and gender were specified as categorical variables and the remaining variables were treated as

[41] The arbitrary value 5% is a conventional cut-off for flagging variables as a problem (Tabachnick & Fidell, 2007, p. 63).

continuous. Little's MCAR test indicated that the randomly distributed missing values (i.e., the null hypothesis) should be rejected at the .05 alpha level, [χ^2 (288) = 337.31, p = .02].

T-tests comparing cases with missing values against those without missing values indicate that this result was driven primarily by Vitality. Independent t-tests indicated that cases with missing values on Vitality differ from the rest of the sample by way of: a) a lower Income, 7.21 vs. 5.92, [t(42) = 2.4], b) lower Maternal Education, 3.21 vs. 2.77, [t(42) = 2.7], c) less French spoken by the child, 15.96 vs. 13.12, [t(42) = 2.3], and d) lower baseline Communication scores, 19.25 vs. 15.01, [t(26) = 2.0]. Given these findings, we decided to drop Vitality from the analysis because it could only serve to adjust for pre-existing differences between the Comparison Daycare and Informal Care groups (i.e., the Program Daycare group is not affected). Therefore, dropping the Vitality variable represents a trivial loss of information while greatly simplifying the impact analyses. When Vitality was excluded, Little's MCAR test failed to reject the hypothesis of randomly distributed missing values. This was true for the entire sample, [χ^2(150) = 151.41, p = .45]; and when the hypothesis was tested for the Program Daycare group only, [χ^2(74) = 75.99, p = .55]. We conclude that the remaining missing values among the covariates are randomly distributed (MCAR) and may therefore be accommodated in the analyses by way of list-wise deletion without significantly affecting the external or internal validity of the study.

4.2.3. Data Imputation Strategy

The strategy that was used to handle missing values varied according to which variable was treated. In all cases, we capitalized on the fact that our study employed a repeated measures design. Having measured the same people on the same (or highly similar) variables repeatedly over the course of the study allowed us to estimate with a high degree of precision the value that would have been obtained had the missing values been observed. We describe the imputation strategy that was employed for the evaluation measures (i.e., the child outcomes) and the survey variables (i.e., covariates and parental outcomes) in separate sections.

Imputation of Child Outcomes

We adopted an imputation strategy for the ÉPE–AD scales that capitalized on the longitudinal nature of the study. The original ÉPE–AD was administered over the first 5 evaluations in either French or English. Thereafter, the revised version of the ÉPE–AD was administered in French only and using a different set of items. No start rule was used with the revised ÉPE–AD, which eliminated the need for imputation of missing data for individual items. The problem of missing values applies only to the first five evaluations.

The imputation procedure was divided into two steps. The first step involved imputing item-level missing values due to the use of the start rule. The second step involved imputing scale-level missing values using a similar procedure. The imputation procedure was based only on scores collected from evaluations that were conducted in French. Both steps of this procedure are now described[42].

[42] Note that the imputation procedure employed here underestimates the amount of within-child variability that truly exists in the population because the algorithm does not add noise to the estimations (Cohen, et al., 2003). In the case of item-level imputation, this criticism is not as relevant given that it is an improvement upon the original imputation procedure specified by the test developer which would have imputed a score of 4 to all start-rule missing values, resulting in an even more severe underestimation of variance. No such plan existed for imputation at the level of scales and therefore the procedure we adopted

Item-level Imputation

As a first step in the imputation procedure, we replaced all the item-level missing values due to the start rule — which exist only for the second and third evaluations (and the fifth evaluation, as described below). The values of the second evaluation were imputed first and then the procedure was repeated for the third evaluation missing. All such missing values were successfully estimated using this procedure (the number and items affected are reported earlier in this chapter, see Section 4.2.1), which will now be described.

The algorithm used for imputing missing items scores made use of within-participant information only from the French version of the ÉPE–AD[43]. First, an "average" gain score was estimated for each child-item combination. These average gain scores represent the average changes in score observed for the child between two consecutive evaluations[44].

A limitation of this method is that the imputation procedure could only be applied to children who completed at least two evaluations in French over the course of the first five evaluations. The non-missing observed values were then used as anchor points for estimation procedure. For instance, missing values on the second evaluation were imputed by adding the average gain score for a given item to the score that was observed on that item during the first evaluation. If information on the first evaluation was missing, then the procedure would estimate "backwards" from the observed score on the third evaluation by subtracting the average gain score.

The start-rule missing values from the fifth evaluations were imputed by a simpler procedure: imputation of the maximum score for each item. As mentioned previously, application of the start rule in this instance was based on the previous performance of the child and therefore the assumption that the child would have obtained the maximum score on the items was much more reasonable[45].

Following imputation of the item-level missing values, scores for the six ÉPE–AD scales were computed for each evaluation period. These scale-level scores were then submitted to an additional imputation procedure intended to replace unobserved values, whether these were due to missing evaluations or use of the English version of the instrument on one or more occasions.

---

could be criticized on the grounds that the cure in this case is worse than the disease. It is true that the precision of the regression coefficient estimates could be slightly overestimated. The potential for bias is greatest where the number of missing values is most important (see Table 4.3). In any case, the Communication scale is unaffected and the potential for bias in the estimate of precision is mitigated by the fact that within-participant estimates are very accurate.

[43] The use of within-participant information ensures that the unique attributes of the participant are adequately represented by the imputation procedure. The procedure therefore does not exaggerate the congruence between a given child's scores and the rest of the group.

[44] Average change scores were computed as the average expected difference between two consecutive evaluation periods. If the child was tested twice during the first and fourth evaluations and obtained scores of 4 and 14 respectively, the average change scores would be the observed difference between evaluations (14-4 = 10) divided by the number of intervals or "steps" between the evaluations, which in this case is three (Average change score of 10/3 = 3.33). All differences between evaluations contributed to this calculation so long as they were within three "steps" of each other. Further, no imputation would take place if the two scores observed for a child on a given scale were for the first and fifth evaluation (i.e., 4 steps).

[45] The exception was 11 cases where the start rule was not implemented correctly in the field for the Self-Awareness scale. In these cases, imputation of the maximum score for each item was no longer justified by reason the language of administration of the ÉPE–AD having changed from English for the fourth evaluation to French for the fifth evaluation. It was not possible to perform an imputation of these scores based on previous responses to the items in question as was done with the start-rule missing from the second and third evaluations. The start-rule missing values for these cases were therefore imputed based on regression estimates derived from their performance on the Self-Awareness scale on the sixth evaluation.

Scale-Level Imputation

The same imputation algorithm that was used for the item-level imputation was applied in the imputation of the scale-level scores. As before, imputation was only possible when a child completed the ÉPE–AD in French at least twice over the course of the first five evaluations. In a first step, an average gain score was estimated for each child. In a second step, this average gain score was combined with the observed values for each child on the French ÉPE–AD to fill in the gaps in the observations wherever possible[46].

A number of missing values remained for the ÉPE–AD scales (other than Communication) because a sub-sample of children completed most of their evaluations in English. This resulted in a loss of 26 children for the associated impact analyses, effectively increasing the average proficiency of the sample in question. The number of observed, imputed missing, and un-imputed missing is reported by evaluation and treatment group in Table 4.3.

Table 4.3: Imputed French ÉPE–AD Values as a Function of Evaluation and Treatment Group

	ÉPE-AD Administrations				
	1st	**2nd**	**3rd**	**4th**	**5th**
Communication Scale	# imputed / n	# imputed / n	# imputed / n	# imputed / n	# imputed / n
Program Daycare group	2/72	0/72	0/70	0/70	0/65
Comparison Daycare group	3/96	1/94	1/95	2/95	2/112
Informal Care group	2/76	0/78	3/79	2/79	1/63
Total	7/244	1/244	4/244	4/244	3/240
Total (% imputed)	2.9%	0.01%	1.6%	1.6%	0.02%
Other ÉPE-AD Scales	# imputed / n (# un-imputed missing)	# imputed / n (# un-imputed missing)	# imputed / n (# un-imputed missing)	# imputed / n (# un-imputed missing)	# imputed / n (# un-imputed missing)
Program Daycare group	10/72 (3)	6/72 (5)	3/70(4)	1/70 (4)	0/65 (0)
Comparison Daycare group	9/96 (6)	6/94 (5)	5/95 (4)	2/95 (4)	1/112 (1)
Informal Care group	1/68 (7)	0/78 (10)	4/79 (11)	2/79 (11)	2/63(5)
Total	20/244 (16)	12/244 (20)	12/244 (19)	5/244 (19)	3/240 (6)
Total (% imputed / % non-imputed missing)	8.1% / 6.6%	4.9% / 8.1%	4.9% / 7.8%	2.0% / 7.8%	1.2% / 2.5%

Note: Results for the "Other ÉPE–AD Scales" are based on the Self-Awareness scale. The exact numbers can vary slightly for some of the other scales, but the difference is negligible (1-2 in the frequency counts). Missing values remaining after imputation are indicated within rounded brackets. List-wise exclusion of cases with missing values results in the elimination of 26 children from the analyses involving scales other than the Communication scale.

Imputation of Survey Data

When it made sense to do so, missing values on these variables were imputed directly based on scores of the variable from other waves of data collection. When this simple imputation procedure was not possible, the missing data were ignored by reason of their MCAR distribution.

[46] We restricted the procedure so that imputation would not be performed if the only two observations were from the 1st and 5th evaluation.

It is important to note that the imputation described below was performed prior to testing the random distribution of the missing values in Section 4.2.2.

For instance, the variable Income was measured three times over the first two years of the study. The most useful of these measurements was that obtained prior to the seventh evaluation period, which distinguished among 10 levels of income (29 missing, 11.9% of sample). The Income variable for the baseline survey distinguished among only six levels of income and suffered from a ceiling effect problem. The two variables were highly correlated (r = .78), which meant that the baseline income scores could be imputed based on the measurement taken prior to the seventh evaluation using a simple regression technique.

For the baseline period, we considered that the most useful covariate was the continuum of French spoken by the child (19 missing, 7.8% of sample). This variable was measured in multiple follow-up surveys, but the measurements taken following participant exposure to the program may have been affected, rendering this type of imputation inappropriate. Instead, simple imputation by regression was performed using a highly correlated variable (r = .81) that was not involved in any of the analyses reported here, but which was measured on a similar scale (i.e., from totally English to totally French): maternal language of the child[47].

The two main dependent measures that were drawn from the surveys were the frequency and language of literacy activities respectively. Missing values for both variables were imputed using the same procedure described in the section for the child outcomes. This decision was taken in recognition of the fact that these variables could vary in time. All missing values due to missing surveys were successfully replaced by this procedure (for response rates, see Table 3.4).

4.3. ERROR TERM SPECIFICATION

As noted earlier, the design of the present study involved sampling observations from two units or levels of analysis. The first of these is daycare and the second of these is participant or perhaps more accurately "family." According to the formalism developed in Chapter 4, observations may be said to be "clustered" within these units of analysis. The main challenge that this presents to a regression analysis is that estimates of standard errors may be too small (Moulton, 1990). If this is the case, then inferential statistical tests will be too sensitive to differences among our treatment groups. The result is an inflation of the Type-I error rate, which in this case is the probability of incorrectly concluding that there is a real difference between the groups that is likely to be replicated by subsequent studies.

This project employed a repeated-measures design. When the same participants are measured repeatedly, positive serial correlation is induced in the observation through time which can bias standard error estimates downward. The treatment effects reported in this paper are based on the DinD estimator, which by construction, neutralizes this kind of bias. In fact, the standard error

[47] The very high correlation between the two variables means that the potential for downward bias in the estimation of standard errors is minimal despite the use of simple imputation (Cohen, et al., 2003, p. 446). The fact that the variable used to perform the imputation is not involved in any of the reported analyses means that there is no danger of introducing multi-collinearity into the analyses.

associated with DinD is essentially equivalent to treating participants as a fixed effect[48]. Even after this source of bias is neutralized, the residuals generated by the regression model may still be correlated (Donald & Lang, 2007). The performance of each child, for example, could be influenced by the daycare context the child experiences. Because multiple children from each daycare participate in the Readiness to Learn project, this context can induce correlation in the residuals of the regression models. If the structure of this correlation is incorrectly specified, then the associated standard error estimates are biased.

We adopted the following strategy in modeling the structure of our residual error. First, we identified a restricted number of plausible clustering specifications. We considered that correlations associated with both daycare and participant could potentially be present in the residuals. We therefore attempted specifications where standard errors were calculated by grouping residuals on the basis of daycare and participant[49]. Of the two possibilities, clustering on daycare produced the largest standard errors. In other words, clustering on daycare yielded relatively conservative estimates of the precision of our program effect estimates. We preferred to understate rather than overstate the precision of our estimates, which is why the bulk of the analyses feature standard errors computed at the level of daycare. This decision provided a total of 20 degrees of freedom for the t-tests involving the DinD estimates[50].

For the analysis of Family Literacy Workshop scales, clustering standard errors based on daycare was impractical due to the small number of daycares implicated in this analysis (n = 6). We compared two alternatives: a) clustering on individuals and b) robust standard errors without clustering. The latter method produced the most conservative tests of statistical significance with the largest standard error estimates. We report results based on the more conservative estimates.

4.4. LIST OF CONTROL VARIABLES

The internal validity of any comparison between treatment conditions depends on the demonstration that the three groups are similar on all dimensions except the one that is manipulated by the tested program. In non-experimental research (and even experimental research), internal validity is maximized through some version of control by way of covariates. The idea here is to measure all relevant variables that are likely to be related to the outcome variable. When such measures are available, the groups to be compared are rendered as similar as possible by way of either participant matching or statistical control (e.g., Behrman, Cheng, & Todd, 2004). Participant matching was impossible because of the limits imposed by sample size. Instead, we adopted a strategy whereby selective statistical control of covariates was undertaken

[48] This equivalence holds when the data are balanced, which is to say that the same participants are observed at each time period and there is no missing data. The inferences that result from such standard errors apply to the group being studied rather than to a population of individuals.

[49] We conducted analyses where standard errors were calculated by grouping residuals according to daycare-time period. The purpose of calculating standard errors at this level was to neutralize the negative correlation in the residuals that was observed across time within individuals for many analyses. In some instances, this strategy was more conservative than grouping residuals by daycare only (i.e., the main impact analyses involving group membership only). In other instances, the resulting standard errors were more liberal (i.e., the impact analyses where hours spent in daycare were incorporated into the model specifications). For the sake of consistency, we decided to compute standard errors at the level of daycare for all analyses based on children outcomes.

[50] The exact number of degrees of freedom can vary depending on the analysis.

when these were found to be related to treatment condition[51]. As a result, the estimated program effects that are reported in Chapters 5 and 6 should be interpreted as the effect of the program conditional on the covariates that were included in the model.

This strategy implies a two-step process whereby potential control variables are identified first and then subsequently inserted into impact analyses. An exhaustive analysis of group differences based on the variables measured during the baseline survey is presented in the *Reference Report* (Legault et al., 2014). However, this analysis was based on the entire sample who agreed to participate in the project. In contrast, the analyses reported here are based on the subsample of participants from the four communities that were selected for analysis of the effects on child-development outcomes (the analysis of parental outcomes was based on the same six communities as in the *Reference Report*, Legault et al., 2014). The possibility exists that for the present sample of participants a different set of covariates is required to ensure the internal validity of the study.

This possibility was investigated by re-analyzing the baseline survey data with the sub-sample of participants that is the object of the current analysis. It was not expected that the result would be drastically different from those reported in the *Reference Report* (Legault et al., 2014). Nevertheless, we considered it prudent not to take this consistency for granted. Moreover, with the extension of the database along the time dimension, the possibility presents itself of testing group differences at each data collection wave. Doing so recognizes the fact that some variables change with time (e.g., family composition variables such as Household Size) including the composition of the treatment groups (e.g., via attrition and group changes). This variability over time implies the possibility of group differences emerging over time that were not apparent at the baseline survey. The strategy that has just been outlined implies a large number of statistical tests, one for each combination of variable and time period (k variables x t Time period). To report the results of this preliminary analysis in detail would add little value to the report while significantly increasing its length. For the sake of efficiency and clarity of presentation, we report in what follows the results of these tests in a synthetic manner.

Group Differences: Years 1 and 2

The following strategy was followed for identifying variables that are potentially confounded with treatment group. Confounded variables are characteristics on which the treatment groups differ. They are a problem because they provide competing explanations for observed program effects when the goal of research is to rule them out.

We therefore tested associations between treatment group and covariates in search for confounded variables. We repeated these tests for each time period for reasons explained in the previous section. As appropriate, we made use of Chi-square tests and Analysis of Variance (ANOVA) to test the group differences. For the purpose of the impact analyses, all variables except Community, Household Linguistic Type (e.g., exogamous French), Single-Parent Household, and Gender were considered continuous. The scoring of all other variables was coded so as to reflect an ordered progression from low to high values on a given dimension and

[51] By selecting our covariates on the basis of their observed relationships with the outcome variables, we assume that the observed estimates are valid (i.e., unbiased by sampling error or measurement error). Statistical control (and even case matching) can only adjust for differences that are observed in the measured variables. Real differences that are undetected (e.g., by imperfectly measuring the offending variable) cannot be controlled in this way.

treated as if they were measured at the interval level of measurement. We report in Table 4.4 a summary of these analyses for variables that a) met the criteria for inclusion as covariates and b) were not redundant with variables already included in the analyses.

Table 4.4: Descriptives, Observed Treatment Group Differences on Observed Characteristics and their Association with Child Outcomes

	Baseline Statistics		*Treatment Group Confound?* [a]			*Relevance* [b]
Covariate Type	**Mean**	**SD**	**G1 vs. G2**	**G1 vs. G3**	**G2 vs. G3**	**Related to DV?**
Continuous Variables						
Sociodemographic						
Household size	4.00	.92	NS	G1 < G3	NS	y
# of older children	.68	.80	NS	NS	NS	y
# of younger children	.38	.50	NS	G1 < G3	NS	y
Child Age (months)	38.36	3.89	NS	NS	NS	y
Income (1–10 scale)	7.02	2.63	NS	NS	NS	y
Education (Mother, 1–4)	3.15	.81	NS	NS	NS	y
Age of First Birth (Mother, 1–5 scale)	2.37	.93	G1>G2	G1>G3	NS	y
Linguistic & Socio-linguistic (Higher values denote better "Francophone" outcomes)						
Language Continuum (Child, 5–20)	15.47	5.79	NS	G1<G3	G2<G3	y
Language Spoken to Child (Mother, 1–5 scale)	4.11	1.37	NS	G1<G3	G2<G3	y
Language of Literacy Activities (1–5 scale)	4.03	1.17	NS	G1<G3	G2<G3	y
Language of Care (0–12 months) (1–3 scale)	2.61	.70	NS	G1<G3	NS	y
Cultural Engagement (1–3 scale)	2.47	.73	NS	G1<G3	NS	y
Vitality (4–24 scale)[c]	16.41	5.64	NS	NS	G2<G3	y
Parental Scales						
Family Functioning (8–32 scale)	29.89	3.02	NS	NS	NS	y
Depression (Mother, 8–32 scale)	10.48	3.39	NS	NS	NS	y
Authoritarian Parenting (4–16 scale)	12.98	2.10	NS	NS	NS	y
Positive Parenting (5–25 scale)	23.30	1.79	NS	NS	NS	y

	Baseline Statistics		*Treatment Group Confound?* [a]			*Relevance* [b]
Covariate Type	**Mean**	**SD**	**G1 vs. G2**	**G1 vs. G3**	**G2 vs. G3**	**Related to DV?**
Categorical Variables						
Single-Parent Household[c]	9.4% of sample		NA	NA	NA	n
Immigrant Status	4.1% of sample		NS	NS	NS	y
Gender	47.5% boys		NS	NS	NS	y
Household Type (exogamous, etc..)	56.6% Endog F, 2.5% Endog E [d], 23% Exog, 16.8% Bilingual		NS [d]	G3 > Endog F[d] G3 < bilingual[d]	G3 > Endog F[d] G3 < bilingual[d]	y

Note: Reported descriptive statistics are for the sample of four communities. Family composition variables are based on the baseline evaluation (and may vary somewhat over time). Association between Treatment Group and Continuous variables is determined by Welch's F-test and Ryan-Einot-Gabriel-Welsch post-hoc tests (Robust to unequal N, heterogeneous variance). We only list those variables that either a) were associated significantly with treatment group or b) were significantly related to one of the dependent variables. DV = Dependent Variables; NS = not statistically significant at the .05 alpha level; NA = Information not available due to insufficient cell-size (<5), Endog F = Endogamous Francophone, Endog E = Endogamous Anglophone, Exog = Exogamous.

[a] G1 denotes the Program Daycare group, G2 the Comparison Daycare group, G3 denotes the Informal Care group. Group differences are reported if statistically significant for any evaluation period across the two years of the study (first through seventh).

[b] Covariate significantly related to at least one DV at the .05 alpha level on the basis of regression/ANOVA for at least one evaluation during the first year of the program.

[c] Vitality was excluded from the impact analyses due to a problem with missing values. The dichotomous variable Single-Parent household was excluded due to excessive skewness. Neither variable had a substantive influence over the results.

[d] Exogamous Anglophone participants excluded due to small cell sizes associated with this category.

In Table 4.4, we report only whether an association was observed between a given covariate and treatment group membership for any of the seven evaluation periods that are the current object of study. In addition, we considered whether the importance of these variables to child development changed significantly over the course of either the first year or the second year of the study (not reported in the table). To test this idea, we estimated a series of regression models for each covariate listed in Table 4.4 based on the specification described in Model 1 of Section 6.1.2. This initial specification was adapted so that the Time by Covariate interaction was used in the specification instead of the interaction between treatment group and time. Using this specification, we re-estimated the model for each covariate and each ÉPE–AD outcome measure. In the end, only one of the tested interactions was retained for the impact analyses because of its significant interaction with time, specifically Language of Care (0–12 months), for Communication: Wald F (3, 18) = 3.07, p = .05[52]. When this variable was inserted into regression models for impact analyses as presented in Chapters 6 and 7, it was invariably as a main effect and in interaction with Time. Its purpose was to correct for any differences in slope between groups that would have been observed between treatment groups due to this characteristic (Language of Care (0–12 months)) in the absence of treatment. This issue was introduced in Section 3.2.3.

[52] Another language variable that interacted significantly with 'Time' was the language continuum spoken by the child. To minimize redundancy in the regression model, this variable was not used in the main analyses reported in Sections 5.1.2 and 6.1. This decision has no bearing on the substantive conclusions to be drawn from the reported analyses. This was confirmed by a series of unreported analyses that included the interaction between Time and this variable as a covariate.

Other Potential Confounds

In addition to associations between the covariates and treatment group, we tested whether the Program Daycare group (G1) and the Comparison Daycare group (G2) differed significantly with respect to the number of hours per week spent in daycare. The impact analyses verified whether the predictive power of this information varied as a function of treatment group. The validity of these tests depended on the assumption that the distribution of this variable be equivalent for the two groups. Welch's F-test indicated that this was the case for the 4-month period preceding the:

- First evaluation, G1: 30.80 (SD = 10.76) vs. G2: 29.35 (SD = 11.50), F(1,158.07) = .70;
- Second evaluation, G1: 27.32 (SD = 8.60) vs. G2: 26.88 (SD = 9.36), F(1,158.61) = .10;
- Third evaluation, G1: 30.01 (SD = 8.92) vs. G2: 28.61 (SD = 10.29), F(1,158.21) = .87; and
- Fourth evaluation, G1: 19.98 (SD = 9.80) vs. 18.31 (SD = 9.54), F(1,142.25) = 1.16.

These null results were confirmed by non-parametric Mann-Whitney U tests. We conclude that the distribution of these dosage variables is statistically equivalent and that tests of dosage interactions with treatment group are not biased by this potential confound.

Another factor that can affect the baseline differences between the program and comparison daycares is individual differences among the educators. A total of 25 women served as primary early childhood educators within participating daycares for the baseline period in the four communities that were part of the analyses (8 in the Program Daycare group, 17 in the Comparison Daycare group). The total years of experience reported by these women did not vary significantly across daycare groups, Program Daycare group Mean = 6.00 (SE = 2.71) versus Comparison Daycare group Mean=7.19 (SD = 8.22) by Welch's t-test, [t(20.19) = .523] or the Mann-Whitney U rank sum test, sum= 50.50, [Z = -1.02]. The minimum cell size was insufficient to test whether the remaining characteristics varied as a function of daycare group ($n < 5$).

Finally, by the end of the first year of the program (October 2008), approximately 52% of the sample was enrolled in school on a part-time or full-time basis. If enrolment was uneven across the treatment groups, this would represent a challenge to the internal validity of the study that would require statistical control. In October of 2008 (fourth evaluation), there were no systematic differences in school enrolment according to a chi-square test, [χ^2 (2) = 2.11]. By February of 2009 (fifth evaluation), school enrolment did vary significantly as a function of treatment group, [χ^2 (2) = 7.08, p =.03]. The effect was driven by a lower level of school enrolment in the Informal Care Comparison group (38%) than either the Comparison Daycare group (57%) or Program Daycare group (60%). This relationship cannot be explained by a disproportionate representation of children in informal care in Edmundston where age-four kindergarten was unavailable. A plausible explanation is that parents who do not have to pay for daycare outside the home have less incentive to send their child to school for junior kindergarten. In any case, school enrolment was entered as a covariate to control statistically for any bias

induced by these observed differences. The potential bias would only have affected estimates of the program effect relative to the Informal Care group.

4.5. REPRESENTATIVITY OF THE SAMPLE: READINESS TO LEARN PROJECT VS. SVOLM (SURVEY OF THE VITALITY OF OFFICIAL-LANGUAGE MINORITIES)

The issue of the external validity of the Readiness to Learn project sample was addressed previously in Chapter 6 of the *Reference Report* (Legault et al., 2014), where a series of detailed comparisons were performed involving the Readiness to Learn project sample and the Survey on the Vitality of Official-Language Minorities (SVOLM) population. Stated simply, the objective of that chapter was to answer the following question: If the daycare program and family literacy workshops were extended to the entire francophone minority population in the communities participating in the project, would the effects observed be similar to those obtained in the Readiness to Learn project? This question is taken up again here with the four communities that were retained for the impact analyses on the child development. If the analyses comparing the Readiness to Learn project and SVOLM are to serve their intended purpose, it is critical that the impact analyses are based on a comparable sample of participants.

The sampling strategies employed in the Readiness to Learn project and the SVOLM limit the comparability of the two samples in a number of ways. As explained in the *Reference Report* (Legault et al., 2014), the Readiness to Learn project sampling procedure is more restrictive than that of the SVOLM. Important demographic differences between the two samples could therefore exist. In order to ensure that a sufficiently large sample was extracted from the SVOLM database, data from children aged 3–5 were considered in the comparative analyses. This contrasts with the average age of the project sample when the baseline survey was administered, which was approximately three. Finally, Readiness to Learn project participants were selected non-probabilistically based on daycare attendance, which means that the geographic distribution of the sample is localized within specific communities. In contrast, the distribution of the SVOLM sample is more geographically diffuse by virtue of this study's use of random sampling. The interpretation of any observed differences between the samples must be qualified in light of the methodological differences.

The analyses reported in the *Reference Report* (Legault et al., 2014) were based on a substantially larger Readiness to Learn project sample, which included children from the communities of Edmonton (AB) and St-Jean (NB). In these analyses certain differences were observed between the Readiness to Learn project and SVOLM samples on socio-demographic and linguistic characteristics. It remains to be seen whether the same pattern of results is obtained when the communities of Edmonton and St-Jean are excluded from the Readiness to Learn project sample, as they are in the impact analyses reported in Chapters 6 and 7. This question was addressed here using the same analytical strategy that was employed in the *Reference Report* (Legault et al., 2014)[53]. The exception is that the analyses to follow are conducted exclusively at the global level. No analyses by community are presented as they would be redundant with those reported in the *Reference Report* (Legault et al., 2014). Whenever possible, we used available

[53] Interested readers are referred to that document for technical details.

data to re-estimate the SVOLM statistics based on the reduced four-community sample so as to maximize the validity of the comparison with the Readiness to Learn project sample employed in the impact analyses[54]. Where this strategy was unfeasible for practical reasons, comparisons to the SVOLM sample gleaned from all six geographical areas are reported and acknowledged in the body of the text.

4.5.1. Immigration Status and Linguistic Profile

The sampling procedures for the Readiness to Learn project and the SVOLM were carefully detailed and contrasted in the *Reference Report* (Legault et al., 2014). The following quote summarizes the conclusions of this analysis:

> *According to Forgues and Landry (2006), a francophone population (such as the one used in the Readiness to Learn project) that is defined using the "ayant droit" criterion would result in a much more restrictive pool whereas a francophone population (such as the one used in the SVOLM) that is defined using several criteria (e.g., mother tongue, knowledge of official languages and languages spoken at home) would result in a greater number of eligible individuals.*

Two predictions were advanced in light of this analysis: a) the SVOLM should comprise a greater proportion of immigrants than the Readiness to Learn project and b) relatively fewer children should list French as their mother tongue in the SVOLM. The first question could not be directly addressed in the *Reference Report* (Legault et al., 2014) because parents were not asked about their status as immigrants. This lacuna was filled in the eighth follow-up survey and the results of an analysis based on this information are presented here. Comparisons based on the mother tongue of children and parents are also reported[55]. The pattern of results reported here for the analysis by mother tongue is equivalent to that communicated in the *Reference Report* (Legault et al., 2014).

Canadian Born Respondent

The immigration status of respondents for the Readiness to Learn project and SVOLM samples (four communities) are reported in Table 4.5. The first row reports the frequency of respondents who report being born in Canada and the second those born outside Canada. As expected, inspection of the response distribution for the two surveys clearly indicates that the Readiness to Learn project comprises a greater proportion of respondents who were born in Canada. Approximately 92% of the Readiness to Learn project sample was born in Canada, while this was true of just 76.4% of the SVOLM sample. This apparent difference was confirmed by a statistically significant chi-square test [X^2 (2, N = 1 025) = 29.14, $p < .001$].

[54] The SRDC presently has access to frequency data by community, which allowed appropriate estimates to be calculated for the global sample comprising the four communities. However, this was not possible for certain variables where the analysis by community resulted in sample sizes too small to be extracted from Statistics Canada. The six community SVOLM sample nevertheless represents an interesting comparison group for the purpose of establishing the external validity of the four community Readiness to Learn project sample.

[55] The FOLS (first official language spoken) was not used here as a basis for comparing the samples due to the fact that this data was collected for the SVOLM in such a way that all comparisons are invalidated (see the *Reference Report*, Legault et al., 2014).

Table 4.5: Comparison between the Readiness to Learn project and the SVOLM

Immigrant status	Readiness to Learn project	SVOLM	Significant differences between groups
	N (%)	N (%)	*Chi square*
Born in Canada	216 (91.9)	598 (75.7)	Yes***
Born outside of Canada	19 (8.1)	192 (23.6)	

Note: Readiness to Learn project sample based on families who responded to the eighth survey only. Significance levels: *** < 0.1%; ** < 1%; * < 5%.

First language learned and still understood: children

The mother tongue of the Readiness to Learn project children is taken from the consent form completed by the parents. For the SVOLM, the child's mother tongue is deduced from the following question (Statistics Canada, 2006, p. 35): "What is the language that [child's name] first learned at home in childhood and still understands?"

Table 4.6 shows that the Readiness to Learn project sample is indeed more Francophone on balance than the SVOLM. A more important percentage of children in the Readiness to Learn project are reported as having French only as their mother tongue (first row of the table). The percentage of children whose mother tongue was English only or English and another language was greater in the SVOLM sample (third row of the table). The representation of bilinguals in the two samples was virtually identical (second row of the table).

A Chi-square test confirmed that the distribution of the Readiness to Learn project children (four communities) across the different categories of mother tongue is not representative of the francophone minority population in the six geographic areas based on SVOLM data [X^2 (2, N = 1 015) = 83.16, $p < .001$]. The results are not surprising given the differences in sampling strategy previously noted for the two research projects.

Table 4.6: Comparison between the Readiness to Learn project and the SVOLM: Children Grouped by Mother Tongue

Mother tongue	Readiness to Learn project	SVOLM	Significant differences between groups
	N (%)	N (%)	*Chi square*
French only	184 (72.4)	306 (40.2)	Yes***
English and French equally OR French and another language	23 (9.1)	89 (11.7)	
English only OR English and another language OR other language(s)	47 (18.5)	366 (48.1)	

Note: Significance levels: *** < 0.1%; ** < 1%; * < 5%.

First language learned and still understood: mothers

Table 4.7 indicates that the majority of mothers from the Readiness to Learn project (69.7%) and the SVOLM (58.3%) samples report French as their sole mother tongue (first row in the table), though the proportion is slightly higher in the Readiness to Learn project sample. Moreover, a lower proportion of Readiness to Learn project mothers report "English only OR English and another language OR other language(s)" category (23.6% in the third row). A Chi-square test suggests that the distribution of Readiness to Learn project mothers (four communities) across the different categories of mother tongue is not representative of the francophone minority population in the six geographic areas based on SVOLM data [X^2 (2, N = 774) = 10.49, $p < .01$].

Table 4.7: Comparison between Readiness to Learn project and SVOLM: Mothers Grouped by Mother Tongue

Mother tongue	Readiness to Learn project	SVOLM	Significant differences between groups
	N (%)	N (%)	*Chi square*
French only	177 (69.7)	460 (58.3)	Yes**
English and French equally OR French and another language	17 (6.7)	75 (9.5)	
English only OR English and another language OR other language(s)	60 (23.6)	254 (32.2)	

Note: Significance levels: *** < 0.1%; ** < 1%; * < 5%.

First language learned and still understood: fathers

Table 4.8 shows the language profile of the Readiness to Learn project and SVOLM fathers based on their mother tongue. The pattern appears at first glance to be similar to that observed with the mothers. The main difference here is the important number of fathers in the SVOLM sample that listed "English only OR English and another language OR other language(s)" as their mother tongue. The proportion of SVOLM fathers at the two extremes of the distribution in Table 4.8 is virtually identical (46.7% versus 47.7%). In contrast, the fathers in the Readiness to Learn project sample are more heavily represented in the 'French only' category (60.8% in the first row) than in the "English only OR English and another language OR other language(s)" category (32.2% in the third row). This latter pattern resembles that observed in both samples for the mother tongue of the mother.

A Chi-square test confirms that the distribution of the Readiness to Learn project fathers (four communities) across the different categories of mother tongue is not representative of the francophone minority population in the six geographic areas based on SVOLM data [X^2 (2, N = 1 031) = 18.16, $p < .001$].

Table 4.8: Comparison between the Readiness to Learn project and the SVOLM: Fathers Grouped by Mother Tongue

Mother tongue	Readiness to Learn project	SVOLM	Significant differences between groups
	N (%)	N (%)	*Chi square*
French only	149 (60.8)	367 (46.7)	Yes***
English and French equally OR French and another language	17 (6.9)	44 (5.6)	
English only OR English and another language OR other language(s)	79 (32.2)	375 (47.7)	

Note: Significance levels: *** < 0.1%; ** < 1%; * < 5%.

4.5.2. Sociodemographic Characteristics

In the *Reference Report* (Legault et al., 2014) the Readiness to Learn project sample was compared with the SVOLM on the basis of the gender of the child, family composition (family size, family structure), and socio-economic status (parental education, family income). The report concluded that the two samples were equivalent in terms of the variables child gender, family structure, family size, and family income. In contrast, the distribution of responses was found to vary across samples for parental education (mother, father) and the number of siblings. In all cases, the general pattern of results was replicated in the analysis based on the four Readiness to Learn project communities presented below.

Total family income

Table 4.9 shows that Readiness to Learn project and SVOLM families are similarly distributed across the income categories considered here. In both cases, the modal and median category for the two samples is $60,000 or more per year. A Chi-square test confirms that the distribution of the Readiness to Learn project parents (four communities) within the different income classifications is statistically equivalent to that observed with the francophone minority population in the six geographic areas based on SVOLM data [X^2 (5, N = 1 031) = 5.42, $p > .05$]. The results suggest that most children in both samples benefit from a good quality/quantity of material resources for their development.

Table 4.9: Comparison between Readiness to Learn project and SVOLM: Families by Income Classification

Income classification	Readiness to Learn project	SVOLM	Significant differences between groups
	N (%)	N (%)	*Chi square*
$10,000 or less	16 (6.6)	54 (3.5)	No
$20,000 to $29,999	13 (5.3)	23 (3.5)	
$30,000 to $39,999	15 (6.2)	64 (8.5)	
$40,000 to $49,999	13 (5.4)	57 (4.3)	
$50,000 to $59,999	33 (13.6)	95 (13.7)	
$60,000 or more	152 (62.8)	496 (66.3)	

Note: Significance levels: *** < 0.1%; ** < 1%; * < 5%.

Mothers' level of educational

Table 4.10 reveals three key points. First, the Readiness to Learn project mothers have a higher average level of education than mothers in the SVOLM sample. In fact, close to 80% of the Readiness to Learn project mothers have a college diploma (DEC) or university degree, while only about 70% of SVOLM mothers have an equivalent level of education. This difference was driven primarily by a smaller number of mothers with at least a college diploma or certificate in the SVOLM compared with the Readiness to Learn project (second row). Second, there are as many mothers with a college diploma (39.0%) as there are with a university degree (39.0%) in the Readiness to Learn project. Third, there are more mothers in the SVOLM who attended university (42.7%) than in the Readiness to Learn project (39.0%), but the difference is negligible. A Chi-square test confirmed that the distribution of the Readiness to Learn project mothers across the different levels of education is not representative of the francophone minority population in the four geographic areas based on SVOLM data [X^2 (2, N = 795) = 16.8, $p < .01$].

Table 4.10: Comparison of Readiness to Learn project and SVOLM Mothers' Level of Education

Level of education	Readiness to Learn project	SVOLM	Significant differences between groups
	N (%)	N (%)	*Chi square*
Secondary school diploma or less OR a few post-secondary courses	56 (22.0)	172 (31.8)	Yes***
College diploma/certificate (e.g. trade school)	99 (39.0)	138 (25.5)	
University degree (bachelor's; master's; PhD)	99 (39.0)	231 (42.7)	

Note: Significance levels: *** < 0.1%; ** < 1%; * < 5%.

Fathers' level of education

Table 4.11 indicates that a comparison between the Readiness to Learn project and SVOLM fathers' educational level resembles the comparison between mothers. As was the case for mothers (see Table 4.10), there are more fathers who attended university in the SVOLM (35.9%) than in the Readiness to Learn project (31.2%). However, Readiness to Learn project fathers are generally more educated than SVOLM fathers. Just over two-thirds of them have a college diploma (DEC) or university degree, while about 60% of SVOLM fathers have an equivalent level of education. Finally, a Chi-square test suggests that the distribution of the Readiness to Learn project fathers across the different educational levels is not representative of the francophone minority population in the four geographic areas based on SVOLM data [X^2 (2, N = 788) = 13.5, $p < .01$].

Table 4.11: Comparison of Readiness to Learn project and SVOLM Fathers' Level of Education

Level of education	Readiness to Learn project	SVOLM	Significant differences between groups
	N (%)	N (%)	*Chi square*
Secondary school diploma or less OR a few postsecondary courses	79 (32.0)	216 (39.9)	Yes**
College diploma/certificate (e.g. trade school)	91 (36.8)	131 (24.2)	
University degree (bachelor's; master's; PhD)	77 (31.2)	194 (35.9)	

Note: Significance levels: *** < 0.1%; ** < 1%; * < 5%.

Family Size

According to Table 4.12, the modal and median family size is four for both samples (four communities). In both surveys, these families represent approximately half the sample. The remaining families are distributed fairly evenly between families of three or less and families of five or more. A Chi-square test suggests that there is no significant difference between the Readiness to Learn project and the SVOLM [X^2 (2, N = 792) = 0.31, $p > .05$] in the distribution of the family size.

Table 4.12: Comparison of Family Size[1] in the Readiness to Learn project and the SVOLM

Number of people	Readiness to Learn project	SVOLM	Significant differences between groups
	N (%)	N (%)	*Chi square*
3 people or less	68 (26.8)	143 (26.6)	No
4 people	136 (53.5)	280 (52.0)	
5 people or more	50 (19.7)	115 (21.4)	

Note: Significance levels: *** < 0.1%; ** < 1%; * < 5%.

[1] The number of people in a family includes only the number of parents and the number of children.

Siblings

Table 4.13 indicates that the modal and median number of children per respondent (family) is two for the Readiness to Learn project and the SVOLM. However, there are slightly more families with exactly two children in the Readiness to Learn project (57.9%) than in the SVOLM (50.4%). Also, there are more families with three or more children in the SVOLM (30.1%) than in the Readiness to Learn project (20.1%). On the other hand, the number of families with an only child, approximately 20%, is about the same in both surveys. A Chi-square test shows that the distribution of the number of children per respondent is significantly different in the two samples [X^2 (2, N = 1 040) = 9.52, $p < .01$].

Table 4.13: Comparison between the Readiness to Learn project and the SVOLM: Number of Children per Respondent

Number of children	Readiness to Learn project	SVOLM	Significant differences between groups
	N (%)	N (%)	*Chi square*
1 child	56 (22.0)	154 (19.5)	Yes**
2 children	147 (57.9)	396 (50.4)	
3 children or more	51 (20.1)	236 (30.1)	

Note: Significance levels: *** < 0.1%; ** < 1%; * < 5%.

Family structure

We had to redefine the Readiness to Learn project families so that they were categorized as either single-parent or two-parent in order to compare the family structure of the Readiness to Learn project and SVOLM samples (see Table 4.14). The latter category comprises intact as well as blended families where two parents (or one parent and his/her spouse) live with the child. The single-parent category includes families where only a single parent lives in the home.

It should be noted that the child's mother/father could be either a biological parent or an adoptive parent. We would also like to mention that same-sex couples were excluded from the analysis, along with any children raised by someone other than their (biological or adoptive) mother and father. As illustrated by the Chi-square test, the distribution of the Readiness to Learn project children in single- or two-parent homes is representative of the francophone minority population in the six geographic areas based on SVOLM data [X^2 (1, N = 1 043) = 0.79, $p > .05$].

Table 4.14: Comparison between the Readiness to Learn project and the SVOLM: Number of Single- and Two-parent Families

Family structure	Readiness to Learn project	SVOLM	Significant differences between groups
	N (%)	N (%)	*Chi square*
Single-parent	23 (9.1)	87 (11)	No
Two-parent	231 (90.9)	702 (89)	

Note: Significance levels: *** < 0.1%; ** < 1%; * < 5%.

4.6. SUMMARY

In this chapter, various methodological issues affecting the internal and external validity of the analyses to be presented in Chapters 5 and 6 have been addressed. Technical issues related to quality control procedures, missing values, standard error estimation, confounded variables, and to external validity have been addressed. We summarize the main points of each section in turn.

An extensive list of data checking and control procedures was enumerated in Section 4.1. These procedures were designed to minimize problems (e.g., measurement error) arising from the data collection process and to identify and correct the problems that did make it into the electronic databases. These procedures and a general adherence to the mixed-methods (or "converging operations") approach to research ensure that the results arising from impact analyses are valid.

The analysis of missing values reported in Section 4.2 indicates that overall the magnitude of the missing-value problem is quite small and evenly distributed across the treatment groups. The attrition rate was very low over the first two years of the project and the response rate was quite high. Analyses indicated that families who scored lower on French-language indicators were more likely to withdraw from the project. Statistical tests indicated that the remaining missing values associated with the survey data were randomly generated. Non-random missing data were

an issue with all ÉPE–AD scales except the Communication domain. These data were caused by the fact that some children completed the ÉPE–AD in English rather than French. The imputation strategy adopted here allowed some of the missing data to be imputed as MAR (i.e., non-random missing that can be imputed reliably on the basis of other variables in the database), but not all (MNAR, i.e., non-random missing that could not be imputed).

Consequently, impact analyses based on these scales may only generalize to a population of minority Francophones whose children are likely to meet the criteria for completing the ÉPE–AD in French (i.e., children who score relatively high on the Communication scale in French, Domain C, and Awareness and Engagement in Francophone Culture, Domain E, or more precisely the population of children who completed the ÉPE–AD in French two or more times over the first five evaluations.

The analytical strategy adopted here (i.e., panel analysis) allowed some flexibility in determining how the standard errors should be estimated (Section 4.3). The logic behind the decisions taken here was described summarily. The general strategy adopted was a conservative one that ensures the precision of reported results is not overstated. Consequently, tests of the program's impact are relatively inefficient, but compensate with their robustness to violations of the assumption of heterogeneity of variance/dispersion. In practice, implementing this strategy meant clustering standard errors at the level of daycares.

An extensive list of covariates was identified in Section 4.4 whose purpose was to a) adjust statistically for changes in group composition over time and b) adjust for differences in developmental trajectory that would have existed between the groups in the absence of treatment. The insertion of these covariates into the analyses enhanced the statistical validity of the results.

Finally, the issue of the external validity of the study was addressed. Previous work has shown the full Readiness to Learn project sample to be more Francophone than expected based on a comparison with the SVOLM sample. This finding was replicated here with the four-community Readiness to Learn project sample. It was argued previously in the *Reference Report* (Legault et al., 2014) that this apparent bias is probably characteristic of the population of children who attend daycare in French. To the extent that this is true, any results obtained with the present sample should generalize to the population of francophone children who attend daycare. This limitation is natural given that this is the population of children that a daycare intervention can reach.

If, contrary to the argument presented above, the differences in mother tongue that were observed between the Readiness to Learn project and the SVOLM are real, which is to say that the Readiness to Learn project sample is in fact non-representative of its target population, then the results of the impact analyses reported in Chapters 5 and 6 of this report may actually under-estimate the importance of the true program effect with a population that is more diverse linguistically. The potential of language variables for accentuating or attenuating the tested program's effect was investigated by re-estimating program effects for children who use the French language more and again for those who use it less (see Sections 5.2.3 and 6.2.3). The information provided by these analyses can inform decisions regarding whether the intervention might be more effective if targeted to specific sub-populations, such as those with the most exposure to languages other than French in their home environments.

The fact that the Readiness to Learn project and the SVOLM are generally similar when non-linguistic characteristics are examined supports the argument that the Readiness to Learn project sample is representative of Francophones in a minority context. The only difference worth noting is that the parents in the Readiness to Learn project sample are slightly more likely to achieve a level of education superior to high school.

In sum, a number of safeguards were put into place to ensure that the estimated program effects presented in Chapters 5 and 6 are as valid as possible both internally and externally. Additional checks are discussed as they become relevant to the interpretation of the results in latter sections of this report.

5. Impact Analyses: First Year

The two central components of the intervention were the new daycare program and the family literacy workshops. The main goal of the first component was to directly influence the outcomes of the children, while that of the second was to directly influence the attitudes and behaviours of the parents. Accordingly, the success of this intervention was evaluated by analyzing the results of outcome measures obtained from both children and their parents. The results from the first year of the project are presented in the present chapter and those from the second year are presented in Chapter 6. We split up the analyses in this way because there were insufficient degrees of freedom to consider all evaluations within the same regression model and the entry of most children into junior kindergarten at the start of the second year of the program provided a natural conceptual breaking point[56].

In the Year 1 analyses, the effects of the tested daycare program on the developmental outcomes of children were tested by: (a) comparing the three treatment groups, (b) evaluating whether the effect of time spent in daycare (in hours) was affected by the tested program, (c) testing the direct effect of program quality and fidelity (for children attending daycare), (d) testing whether quality and fidelity of the program adequately account for the observed treatment group differences, and (e) re-estimating the differences between treatment groups specifically for children whose exposure to French is high and again for those children whose exposure to French is low. The purpose of these last analyses was to identify sub-populations for whom the tested daycare program is particularly effective.

The impact of the family literacy workshops on the attitudes and behaviour of parents was tested by comparing Program Daycare group families based on their workshop attendance and the quality/fidelity of the instruction they received. All families in the sample were implicated in analyses of the frequency and language of literacy activities where it was possible to compare the three treatment groups.

These two distinct research questions, focused respectively on the children and on the parents, are addressed respectively in Sections 5.1 and 5.2. A summary of these findings is presented in Section 5.3.

5.1. CHILD LEVEL IMPACTS

5.1.1. Variables Retained for Analyses

Among the numerous variables that were measured over the course of the study, the following were retained for the analyses of the Year 1 impacts. These variables can be divided into two broad categories: outcome measures and predictors. We now turn to a brief discussion of both types of variables.

[56] As described in Chapters 4 & 5, we chose to cluster standard errors at the level of daycares, which is a decision that yields approximately 20 degrees of freedom for significance tests. The inclusion of all repeated measures in the same analyses would drop the degrees of freedom below 10 for some significance tests, which means that a fair test of the program would not be possible.

Outcome Measures

The focus of this impact report is on school readiness. For this reason, we selected the following outcomes from the ÉPE–AD (French version only): the Communication, Self-Awareness, Cognitive Ability, and Physical Ability scales. Complete data were available only for the Communication scale[57]. More fine-grained vocabulary subscales were created by harvesting relevant items from the four principal outcome domains. In constructing these scales, we made the well-documented distinction between Receptive Vocabulary and Expressive Vocabulary. In sum, a total of six ÉPE–AD outcomes were the object of the following analyses.

Standardization of Outcomes

Prior to analyses, the outcomes were rescaled so that they were standardized within time period. As a result, each outcome variable had a mean of zero and a standard deviation of one for the complete sample of participants. A practical implication of this transformation is that scores at each time period can be interpreted in terms of a common scale. The decision to standardize the scores was motivated by the fact that the nature of the raw scale scores changed qualitatively in the second year of the study. Specifically, the composition of the scales was altered (i.e., items were dropped) with corresponding changes to the total number of items. The strong correlation between the original and revised scales supports our contention that very little information was lost as a result of these changes. Nevertheless, the differences in scaling (e.g., maximum score) in our measures across time presented a technical obstacle to statistical analyses in that the meaning of the absolute values of scores varies over time, invalidating any comparisons across time period. This minor technical challenge was overcome via the standardization procedure discussed above which ensures that the outcome scores have the same meaning over time.

Interpretation of Standardized Scores

The standardized scores employed in the impact analyses are interpreted as follows. Each participant's standardized score represents the difference between that participant's score and the average score of the sample. Critically, this difference is expressed in standard deviation units. For example, a score of 1.11 means that the participants obtained a score that was 1.11 standard deviations above the average score obtained by participants in the Readiness to Learn project sample for the evaluation in question. The DinD estimates of program effects are interpreted similarly as differences between groups in standard deviation units. Cohen (1988) provides conventional benchmarks for interpreting the magnitude of effects in a standardized scale. A standardized difference between groups of .20 is considered small, .50 is considered medium, and .80 is considered large. These benchmarks are intended only as a rough guide for judging the importance of an effect.

[57] For the remaining three scales, data were available only for participants who completed the ÉPE–AD in French at least twice over the course of the first five evaluations. For obvious reasons, this means that analyses conducted on the Self-Awareness, Cognitive Ability, Physical Ability, Receptive Vocabulary and Expressive Vocabulary scores constitute effective sub-sample analyses based on a subgroup that excludes those children who are weakest in French language skills. This exclusion was not significantly related to experimental group, which means that the internal validity of the program effects estimates is not threatened. Practically, this issue affects external validity by limiting the generalizability of results to children who were able to meet the criteria for passing the test in French.

Weighing the Costs of Standardization

The cost of standardization is that we lose the ability to a) compare the raw scores of our sample to a normative population and b) evaluate the developmental trajectory of our sample over time[58]. We argue that this cost is trivial because a) no population norms exist for the version of the ÉPE–AD that is employed by the Readiness to Learn project and b) the children in the study have completed the test several times which has no doubt induced massive practice effects (i.e., better scores with repeated performance of a task). These practice effects are necessarily confounded with any attempt at estimating the development of our sample over time. In other words, the intrinsic meaning of the absolute value of observed raw scores is limited and changes in scores over time (i.e., the developmental trajectory of the sample) could not be disentangled in any case from practice effects. What matters is the ability to meaningfully compare the *developmental trajectories* of our three treatment groups over time. The "difference in differences" estimate adequately represents these comparisons and it applies just as well to standardized scores as to raw scores[59].

Substantive Predictors and Covariates

A number of variables were entered into regression equations modeling the development of children's school readiness outcomes. The broadest distinction that can be made here is that between substantive predictors of children outcomes and covariates. The substantive predictors are the object of specific research interests, while the simple covariates serve only to improve the internal validity of the more substantive tests. In the interest of economy and clarity of presentation, the focus of our reporting is on the substantive predictors in our models.

Substantive predictors include time (testing period), the experimental group to which the child has been assigned, the average hours a week the child spends in daycare, and the quality of these daycare services. This list will be extended to include the linguistic characteristics of participants (i.e., household type, language of literacy activities at baseline, and the languages spoken by the child in the home at baseline) in Section 6.1.3 where moderators and mediators of the experimental effect are investigated. These variables are described in the Method (Chapter 3) and the list of covariates to be included in analyses is also described in details in the Section 5.4. A total of eight variables were inserted into analyses to explicitly control for group differences, while 11 additional variables were included due to their statistically significant relationship with outcome variables.

5.1.2. Results of the Group Comparisons

The results of the Year 1 impact analyses are reported in this section. Estimates of the treatment effect are based on the relative developmental trajectories of the three treatment groups over the course of the first four evaluations. The first evaluation is considered a pre-test evaluation while the subsequent three evaluations are considered post-test evaluations. As noted earlier, the treatment effect is estimated by way of the DinD estimator which captures the

[58] Standardization of scores results in an average score of zero. Standardization within each time period results in an average score of zero at each time period. If the average score is zero at each time period, then the average is constant across time, and the effect of 'Time' in a regression model will necessarily be null. When the overall developmental gains for the sample of participants are not of direct interest, this represents a trivial loss of information.

[59] Preliminary analyses indicated that all child outcome measures were sensitive to change over time. Further, none of the measures were at ceiling by the 5th Evaluation.

difference between groups in their respective developmental trajectories. A total of six outcome measures were the object of analyses which were based on data collected on the French version of the ÉPE–AD: Communication, Self-Awareness, Cognitive Ability, and Physical Ability scales and the improvised Receptive Vocabulary and Expressive Vocabulary subscales developed by SRDC. We predicted positive impacts of the program on all measures except the Physical Ability scale, which taps an ability that is not directly targeted by the program. We now describe how these outcomes measures were treated in the impact analyses.

Overview of Analyses & Specification Details

The purpose of this section is to convey the logic behind the impact analyses. First, we explain the reasoning behind the structure that was adopted in the presentation of the results. Then, we describe the details of our regression-model specifications.

Structure of This Section

The results for this series of analyses are presented in two parts. The first part reports a detailed account of analyses of French Communication scores across time based on the full sample of participants. For the second part, a more synthetic analysis is presented of the entire set of six ÉPE–AD scales, including the Communication scores. The results reported in the latter section are based on the sub-sample of participants who completed the ÉPE–AD in French at least twice during the first five evaluations. Recall that children only continued their evaluations in French if they scored sufficiently well on both a short scale assessing the linguistic home environment of children and the ÉPE–AD Communication scale (see appendix B for details related to the decision rule). The result is missing data for the Self-Awareness, Cognitive Ability, and Physical Ability scales because of the impossibility of combining scores from the French and English versions of the ÉPE–AD.

Special status was given to analyses based on the full sample of participants as it represents a strong test of our program effects. The sample size for these first analyses is relatively large and unbiased. The second series of analyses is based on a sample that is both smaller and biased towards those children who possess relatively strong French-language skills. This bias was unavoidable given the nature of the missing values that caused the exclusion of cases[60]. We decided to perform a second set of analyses of the Communication scores with this sub-sample so that a common sample of participants is used to generate estimates for all scales. Comparison of this second set of analyses of the Communication scores against the first set of analyses gives a general idea of the magnitude of bias engendered by the exclusion of children who completed the ÉPE–AD in English. In fact, a comparison of the two sets of results indicates that the affect of this bias on program effect estimates is negligible.

Model Specification Details

An extensive set of covariates was employed in estimating the program impacts. Rather than enter all covariates at once, we adopted a strategy whereby groups of covariates were entered into the regression model in a series of incremental steps. An advantage of this approach is that it

[60] A sound imputation strategy would require a sufficiently large sample of children that should have completed the test in English, but were administered the test in French anyway. If this has been the case, accurate estimates of the scores that the children would have obtained on the French version could be obtained.

allows changes in the magnitude of effects to be tracked across different specifications. We could, for example, isolate the group of variables that resulted in the elimination of the treatment effect. Another practical advantage of this strategy is that it allows the diagnosis at each step of certain technical problems with regression estimates like suppression (i.e., large changes in the magnitude or direction of an effect) and multicollinearity (i.e., inflation of standard errors). The specifications that were employed at each step are now discussed. We indicate each specification model by a numerical value placed within rounded brackets (see below).

The initial model (1) for evaluating program impacts included dummy variables representing Time period, dummy variables representing treatment group membership, and a term representing the interactions between these dummy-variable indicators (i.e., the DinD estimators). In accordance with the logic of the DinD estimator, the baseline or pre-test time period was placed in "reference." In other words, scores on all subsequent time periods were compared against the baseline value. For the treatment effect dummy variables, we chose to place the Program Daycare group in reference. Placing the Program Daycare group in reference facilitated the interpretation of the DinD estimates generated by the model by causing the developmental trajectories of both comparison groups to be compared against the Program Group. In other words, we obtained estimates of our program impacts relative to the Comparison Daycare group and the Informal Care group in a single regression analysis. *It is important to note that the decision to place the Program Daycare group in reference means that* ***negative values of the DinD estimator for all group comparisons represents a positive program effect*** *(i.e., an advantage for the treated group).* For the purposes of this base model, we controlled for the effect of Community by including a series of dummy variables.

In a second step (2), covariates of a linguistic nature were included in the model. The list of variables includes: a) the continuum of languages spoken by the child at baseline, b) the language of literacy activities at baseline, c) the linguistic composition of the household (e.g., Endogamous Francophone, exogamous), d) the language of care when the child was aged 0 to 12 months, and e) the interaction between this language of care variable and time. The purpose of including these variables was to adjust for the statistically significant linguistic differences that were observed predominantly between the Program Daycare group and the Informal Care group. The interaction between time and language of care was included because it was significantly related to most of the dependent measures. The other covariates were not. Its inclusion in the model improves the validity of the DinD estimator by correcting for differences in slope that may have existed at the baseline evaluation. This issue was discussed at greater length in Section 4.3.3.

In a third step (3), covariates of a demographic and socio-economic nature were inserted into the model. The child's gender was entered into the model at this step along with his or her age in months when the ÉPE–AD was administered. Socio-economic control variables included the mother's level of educational attainment, the mother's age when the child was born, and the family's revenue class.

In a fourth step (4), covariates related to family composition were included in the model. These covariates are dynamic in that the value taken by these variables could change over time. In contrast, the other covariates are time-invariant or "fixed." This list of covariates comprised the following variables: Household size (number of people living in the home), the number of younger children living in the home, and the number of older kids living in the home. The

dummy variable representing whether or not the child lives in a single-parent home was excluded from the final set of reported results because it was highly skewed (exceeding a 90:10 split, Tabachnick & Fidell, 2006, p. 73) and highly correlated with the other family composition variables.

In the fifth and final step (5), a number of parental scales were entered as covariates. Some of these scales captured dimensions relating to parenting style, such as the Authoritative Parenting scale and the Positive Parenting scale. An additional scale was included in the model that was designed to estimate Family Functioning.

The strategy of incrementally including covariates was applied in all of the analyses conducted. However, the results that were obtained at each step are only reported here for the Communication scale analyses based on the entire sample of participants. Where the results obtained at earlier steps affect the substantive interpretation of the results, this is discussed in the body of the text.

IMPACT RESULTS: Standardized Communication Score (in French) — Full Sample

The impact of the tested program on French-language communication skills is evaluated in this section. Only estimates of the program impacts relative to the Comparison Daycare and the Informal Care groups are reported, which take the form of DinD estimates. These estimations should be interpreted as estimates of the program impacts conditional on the covariates that are included in the model (covariate effects are unreported). A summary of the results of these analyses are reported in Table 5.1.

Several points are worth noting about the results presented in Table 5.1[61]. In the first place, we note that the differences between the Program Daycare and comparison groups for the baseline period are not statistically significant for this outcome measure at the .05 alpha level[62]. There is very little evidence then to suggest that reliable differences between groups existed prior to the intervention. Initial differences on the outcome measure may therefore be discarded as a threat to the internal validity of our treatment effect.

In terms of the treatment effect, the first thing to note is that the global test of the Group by Time interaction is not statistically significant at any stage of the analyses (see the row labelled "Omnibus Wald F-test")[63]. This F-test evaluates the pooled effect of all DinD estimates in the

[61] The number of participants involved in the analyses decreases slightly as more covariates are inserted in the model. This reduction in sample size is attributable to the presence of a handful of missing values spread across these variables. In principle, the sample size discrepancy does not pose problems for the interpretation of the results because the number of missing values is small and their pattern is not systematic.

[62] The result of this test seems to contradict that reported in the *Reference Report* (Legault et al., 2014), where a significant difference was indeed observed, but the contradiction is only apparent. We remind the reader that the sample of children involved in the present analysis is smaller after the exclusion of two communities (Edmonton, St-Jean) and ten first-year attrition cases. This change in sample composition is responsible for the difference across reports in treatment-group similarity for the baseline period.

[63] A common technique for minimizing the number of inferential tests performed is to ignore simple contrasts when the global test of significance is null. The idea here is to reduce the number of statistical tests and by extension the probability of making a Type-I error. If this approach were to be applied here, we would have to conclude that the program had no impact on Communication scores. However, we have not applied this approach because a) the DinD estimate associated with the 1st Post-test was identified as being of special interest (i.e., an a priori contrast) and b) presenting the results in a more detailed form allows for a coherent pattern of results to be extracted, which can then be integrated into the broader picture of results presented in this report. In short, we chose to let the data speak.

regression model. That this general test fell short of significance suggests parallel developmental trajectories overall for the three groups over time. This finding indicates that the impact of the program is either absent, weak, or transitory. The latter possibility seems most likely given that implementation of the program was most consistent leading up to the first Post-test (see *Project Implementation Report*, Bérubé et al., 2014). We therefore predicted that the size of the program effect should be greatest for the first post-test period and proceeded to test this idea by examining individual DinD estimates[64].

In doing so, we discovered that the Program Daycare group invariably shows larger gains in communication ability over time than either of the two comparison groups. Relative to the Informal Care group, this advantage does not approach statistical significance for any post-test period. As anticipated, the largest DinD estimates were observed for the first post-test (four months after the baseline evaluation). Only the comparison of program and comparison daycares yielded a statistically significant effect.

The associated coefficient indicates that the Program Daycare group gained approximately .30 of a standard deviation more in Communication scores on average than the Comparison Daycare group. The size of this effect falls between conventional benchmarks for a small and a medium effect, .20 and .50 (Cohen, 1988). Subsequent post-test estimates indicate a fifty per cent reduction in effect size and a lack of statistical significance. In sum, the initial gains sparked by exposure to the tested program were modest and did not persist past the first post-test.

[64] Every test of statistical significance is associated with a non-zero probability of incorrectly concluding that the effect being evaluated is real rather than a statistical anomaly. This type of error is called a Type-I error. Conventionally, the probability of a Type-I error is controlled by fixing the nominal level of statistical significance at 5% ($p < .05$). For each statistical test that is performed, the probability of having made at least one Type-I error increases additively such that when two tests are performed the probability of making at least one Type-I error is 10 percent instead of 5. For the sake of comprehensiveness, we report the results of all DinD tests. Rather than implement an overly conservative adjustment to the nominal Type-I error rate, we chose to report the observed range of statistical significance (two-tailed), applying an unadjusted nominal .05 Type-I error in evaluating statistical significance. When participants from the second cohort are inserted into the analyses, the statistical power of the tests will allow for a more conservative approach. In interpreting the results presented here, readers are free to apply their own corrections for Type-I error rate (e.g., Bonferroni). Alternatively, we suggest that a useful strategy for guarding against Type-I errors is to weight results most heavily in interpretation when they conform to targeted predictions and when they are consistent with a broader pattern of results.

Table 5.1: Year 1 Difference in Difference (DinD) Program Effects for Standardized ÉPE–AD Communication Scores — French Version (Full Sample)

	Incremental Inclusion of Covariates – Errors Clustered by Daycare									
	(1)		**(2)**		**(3)**		**(4)**		**(5)**	
Program Effects	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Program Daycare vs. Comparison Daycare at Baseline (G1[a] vs. G2)	.241	.147	-.184 [m]	.112	.195 [m]	.110	.223 [m]	.110	.183 [m]	.117
DinD 1st Post-test	-.338[*]	.144	-.309 [m]	.149	-.313[*]	.147	-.312[*]	.146	-.314[*]	.151
DinD 2nd Post-test	-.194	.149	-.192	.167	-.177	.168	-.176	.171	-.189	.175
DinD 3rd Post-test	-.158	.149	-.139	.160	-.115	.156	-.117	.158	-.135	.153
Program Daycare vs. Informal Care at Baseline (G1 vs. G3)	-.045	.105	.205	.150	-.169	.081	-.108	.081	-.164	.088
DinD 1st Post-test	-.221	.144	-.217	.143	-.200	.133	-.204	.132	-.223	.137
DinD 2nd Post-test	-.083	.174	-.051	.180	-.032	.175	-.033	.177	-.060	.188
DinD 3rd Post-test	-.105	.195	-.081	.188	-.058	.174	-.070	.172	-.115	.186
Omnibus Wald F-tests:	F (6, 15) = 1.44		F (6, 15) = 1.26		F (6, 15) = 1.46		F (6, 15) = 1.60		F (6, 15) = 1.59	
N participants:	243		238		237		237		230	

Note: The outcome has been standardized, which means that the DinD effects can be interpreted in terms of standard deviation units. ***Program Daycare group (G1) is in reference, which means that negative values of DinD indicates a positive program effect***. Standard errors were estimated via the SPSS implementation of the heterogeneity consistent "robust" White estimator (White, 1980). Further, standard errors were clustered by daycare to adjust for correlation in the residuals over time. Omnibus Wald F-test evaluates the combined effect of all DinD estimates.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed, 20 degrees of freedom).

[a] G1, G2, and G3 denote Program Daycare group, Comparison Daycare group, and Informal Care group respectively.

Model specification details are provided in Section 6.1.2.

IMPACT RESULTS: Standardized ÉPE–AD Scales (in French) – Sub-Sample

The impact of the preschool program on the remaining five ÉPE–AD outcomes is evaluated in this section. The Communication scale is re-analyzed as well with the same sub-sample of participants, which excludes the 26 children who did not complete the ÉPE–AD in French at least twice during the first year. For reasons of economy, we report only the estimates of the program effect for the regression model that included the full set of covariates. The omitted set of results has no bearing on the substantive interpretation of the results and can be made available upon request.

The estimates generated by these regression analyses are reported in Table 5.2 for the standardized Communication, Self-Awareness, Cognitive Ability, and Physical Ability scales. In addition, the estimates associated with the improvised Receptive Vocabulary and Expressive Vocabulary subscales developed by SRDC are reported in the same Table. As indicated earlier, a second set of analyses were conducted on the Communication scores with this sub-sample of participants in order to obtain a set of results for all ÉPE–AD scales based on a common sample of participants. Given the focus of the tested preschool program, we anticipate the strongest program effects for the scales that tap most directly into the ability of children to communicate and reason in French. This list of scales includes the Communication, Self-Awareness, Vocabulary, and to a lesser extent the Cognitive Ability scales. In contrast, we do not anticipate program effects for the Physical Ability scale given that it is comprised of tasks that tap abilities which are not directly related to the goals of the tested program. The results for each outcome measure are now discussed.

Communication Scale (Standardized)

Re-analysis of Communication scores using a restricted sample yielded results that are consistent with those reported earlier. Note that the baseline test failed to indicate a statistically significant difference between our Program Daycare group and either comparison groups. We conclude that prior to the intervention the groups were equivalent on this measure.

The story for the program effects was the same as before with an estimated effect size of .30 standard deviations for the first post-test's DinD estimate. Yet again, the magnitude of this effect dropped off substantially in subsequent time periods, at which point it was no longer statistically significant. Finally, comparisons involving the Informal Care group failed to indicate the presence of a statistically significant effect.

Some minor differences in the original and re-estimated program effect estimates are worthy of note. For one, the size of the program effects tapered off more markedly over time in this second set of analyses so that it was virtually null by the third post-test. Another difference was that the joint test of all DinD effects (i.e., Omnibus F-test) in the present set of analyses was marginally statistically significant at $p < .10$. The conclusions of these analyses need not however be qualified: modest but significant program effects were observed on Communication scores for the initial post-test period (four months following the baseline evaluation) that do not persist long enough to be detected eight months from the baseline evaluation (second Post-test); even after statistical adjustment, the Program Daycare group at no point differed in its developmental trajectory from the Informal Care group.

Self-Awareness Scale (Standardized)

Analyses of the Self-Awareness scores yielded a picture that is the mirror image of that obtained with the Communication scale. All groups were equivalent for the Baseline evaluation and a statistically significant advantage in the order of .30 standard deviations is observed for the Program Daycare group relative to the Comparison Daycare group for the first post-test evaluation only. The smaller standard errors reported for these analyses indicate that the associated estimates are more precise than those associated with the Communication scale analyses. It is perhaps for this reason that the pooled test of the DinD effects for all time periods is statistically significant at the .05 level[65]. Preliminary analyses reported in the Method (Chapter 3) indicate that this scale is strongly associated with vocabulary measures. We therefore attribute this effect to gains in French-language communication abilities.

Cognitive Ability Scale (Standardized)

The preceding results are echoed again in the analyses of Cognitive Ability scores. Baseline differences between the Program Daycare group and the two comparison groups are not statistically significant. Yet again, this result is qualified by a statistically significant interaction between Treatment Group and Time period (Omnibus F-test), which is driven by the statistically significant DinD effect arising from the comparison of the two daycare groups. No other DinD estimate generated by these analyses was statistically significant.

Physical Ability Scale (Standardized)

There is little evidence to suggest that francophone children in a minority context are disadvantaged relative to their peers with respect to their physical development, and consequently the tested preschool program was not designed to influence such skills. We therefore did not expect to observe a significantly program effect on this outcome and our expectations were confirmed.

Receptive Vocabulary Scale (Standardized)

Analyses of children's performance on the Receptive Vocabulary subscale indicated that overall the trajectories of the three treatment groups were non-parallel, as evidenced by the statistically significant Omnibus F-test of the joint effect of all DinD variables. However, this initial positive result was not due to statistically signification DinD program effects relative to either the Comparison Daycare or Informal Care groups. Contrary to expectations, analyses of the Receptive Vocabulary subscale failed to reveal a statistically significant program effect for the initial post-test period or any subsequent period. A lack of statistical power could not have been the cause given that no hint of an effect could be discerned in the direction and size of the DinD estimates; the sign of the effects indicated a positive program effect in some instances and a negative impact in others. Fluctuations of this type strongly suggest statistically equivalent developmental trajectories.

[65] Again, a common approach to the analyses of interaction effects holds that the joint test (Omnibus F-test) should act as a sort of gatekeeper. According to this view, we would only estimate and test the DinD effects if the global effect is statistically significant. If we had adopted this approach, the Self-Awareness scale would be the first to show a statistically significant program effect.

Expressive Vocabulary Scale (Standardized)

The abilities tapped by Receptive and Expressive Vocabulary measures are recognized as distinct. Previous research has shown that each measure typically has a unique relationship with various outcome measures (e.g., Wise, Sevcik, Morris, Lovett, & Wolf, 2007). Therefore, it would not be unexpected to observe a program effect with the Expressive Vocabulary measure in the absence of a corresponding effect with Receptive Vocabulary. This is true whether the object of analyses is one of the ad hoc scales or the well-validated vocabulary scales (i.e., ÉVIP–R, EOWPVT) that are presented in Year 2 impact analyses.

Indeed, analyses of Expressive Vocabulary scores revealed a relatively strong program effect relative to the Comparison Daycare group for the initial post-test evaluation. The effect was in the order of .50 of a standard deviation, which corresponds to the benchmark value for a medium-sized effect. This finding is qualified by the observation of non-equivalence between the groups for the baseline period, where the children in program daycares show a .30 standard deviation disadvantage. Recall that the DinD estimator cancels out such initial differences in calculating program impacts under the assumption of parallel developmental slopes (see Section 6.3 for an elaboration on this point). Invariably, we observed statistically equivalent development in the Program Daycare and Informal Care groups.

Table 5.2: Year 1 Difference in Difference (DinD) Program Effects for Standardized Subscale Scores of the ÉPE–AD (French Test Takers Only)

	Analyses by Subscale – Errors Clustered by Daycare											
	Communication		**Self-Awareness**		**Cognitive Ability**		**Physical Ability**		**Receptive Vocabulary**		**Expressive Vocabulary**	
Program Effects (Model 5)	DinD	Robust SE	DinD	Robust SE	DinD	Robust SE	DinD	Robust SE	DinD	Robust SE	DinD	Robust SE
Program Daycare vs. Comparison Daycare at Baseline (G1[a] vs. G2)	.182	.116	.160	.145	.187	.145	.110	.133	-.098	.180	.319*	.151
DinD 1st Post-test	-.318*	.149	-.337†	.084	-.279†	.126	-.055	.187	.061	.373	-.543†	.144
DinD 2nd Post-test	-.179	.178	-.257	.149	-.069	.158	-.276 m	.154	-.330	.298	-.230	.200
DinD 3rd Post-test	-.094	.148	-.084	.163	-.071	.166	-.217	.164	-.140	.182	-.147	.154
Program Daycare vs. Informal Care at Baseline (G1 vs. G3)	-.135	.094	-.081	.140	.101	.149	-.129	.137	-.429 m	.210	-.040	.125
DinD 1st Post-test	-.175	.133	-.062	.112	-.214 m	.115	.019	.229	.111	.361	-.154	.168
DinD 2nd Post-test	.048	.178	-.181	.142	-.132	.143	-.059	.190	.220	.227	.049	.220
DinD 3rd Post-test	.015	.180	.025	.110	-.222	.159	.089	.248	-.017	.153	.096	.139
Omnibus Wald F-tests:	F (6, 15) = 2.53 m		F (6, 15) = 2.89*		F (6, 15) = 2.93*		F (6, 15) = .758		F (6, 15) = 2.94*		F (6, 15) = 3.67*	
N participants:	217		215		216		215		216		215	

Note: The outcomes have been standardized, which means that the DinD effects can be interpreted in terms of standard deviation units. Analyses based on the sub-sample of participants who completed the ÉPE–AD in French at least twice during the first year of the study. Thus, the sample is biased towards those children who are stronger in French. ***The Program Daycare group (G1) is in reference therefore negative values of DinD denote an advantage for the Program Daycare group***. Standard errors were estimated via the SPSS implementation of the heterogeneity consistent 'robust' White estimator (White, 1980), and were clustered at the level of daycare to adjust for correlation in the residuals over time.

m denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed, 20 degrees of freedom).

[a] G1, G2, and G3 denote Program Daycare group, Comparison Daycare group, and Informal Care group respectively.

Model specification details are provided in Section 6.1.2.

Summary

The impact of the tested program on a series of school-readiness outcomes was reported in this section. The experimental design called for analyses to be centered on two critical contrasts. The first of these pitted the development of the Program Daycare group against that of the Comparison Daycare group, while the second compared the Program Daycare group to the Informal Care group.

The question addressed by the first contrast is whether enrolment in a newly created preschool daycare program results in better outcomes than enrolment in a daycare whose program is typical of what is normally available at a government-approved facility. For this series of contrasts, we observed an advantage of the new preschool daycare program on the expected outcome variables (i.e., with the Communication, Self-Awareness, and Cognitive Ability) and we failed to observe such an effect with the Physical Ability scale where such an effect was not expected. This pattern of results lends credibility to the reported findings because it demonstrates both convergent and discriminant validity[66]. Analyses of the Vocabulary scales suggest that results are being driven primarily by items that tap Expressive Vocabulary in French. Strong indications were obtained of the program's impact on the readiness of children to attend school in French.

The question addressed by the second contrast is whether enrolment in a newly created preschool daycare program results in better outcomes than a less formal child-care setting. The results of the preceding analyses provide little evidence that this is the case. The developmental trajectories of the two experimental groups appeared to be equivalent for all time periods. This equivalence was apparent both before and after controlling for a wide assortment of variables.

The most interesting contrast for assessing the impact of the new preschool daycare program is the one pitting the development of the Program Daycare group against that of the Comparison Daycare group. Examination of the pattern of results indicates that the advantage of the new preschool daycare program over existing comparison daycare programs is transitory for the outcomes considered here. The program effect is strongest for the first post-test where it is also statistically significant (four months from baseline), but it is weaker for latter post-test periods and not statistically significant.

5.1.3. Moderators and Mediators of the Intervention Effect

In the preceding section, the results of the main impact analyses were presented, which evaluated the impact of treatment group membership on four school readiness outcomes and two ad hoc vocabulary measures. We enriched this pool of information by conducting additional analyses including variables that are supposed to moderate or "interact with" the program effect. In other words, we considered predictors of the size of the tested programs effect in a series of regression analyses. This approach takes the initial research question, "did the tested program have a positive effect for the current sample of children?", and alters it slightly so that it

[66] The term convergent validity is used here to refer to the case where the expectation of an association between two variables is confirmed. In contrast, the term discriminant validity refers to the case where the expectation that an association will be absent (or negligible) is confirmed. In this study, neither experimenters nor participants were blind to treatment. In this situation, the validity of the results is strengthened if the program effects are specific to target outcomes (i.e., when both convergent and discriminant validity is present). This claim is based on the assumption that various sources of bias would affect all variables in similar ways.

becomes: “for whom did the tested program have the largest effect?” This more flexible approach can reveal important program impacts that were not readily apparent in the main impact analyses. If we successfully predict this aspect of results, it contributes to the body of evidence supporting our contention that the tested program is having a real impact on the development of children.

This line of enquiry was explored using three types of moderator variables: program exposure or dosage, fidelity/quality of implementation of the tested daycare program, and the linguistic characteristics of the sample. The first type of variable tests whether the amount of time spent in daycare makes a difference of the observed program impact. The second type of variable was actually treated as a “***mediator***” in the analyses, which is a related but distinct concept[67]. Finally, the analyses based on linguistic characteristics are germane to the issue of external validity. If the new preschool daycare program evaluated here is to be implemented in other contexts, it is important to know for which groups it will be most effective.

Fine-grained Definitions of Dosage: Average Hours Spent in Daycare

The hypothesis behind the main impact analyses reported in the preceding section was that enrolment in a new preschool Program Daycare would lead to superior development on school-readiness outcomes. This formulation of the research hypothesis implies sufficient exposure for the intervention to have its effect. A natural extension of this hypothesis is to suppose that the relationship between the amount of exposure to daycare and the outcome variables will also depend on the type of daycare program to which a child is exposed. Specifically, we expect that exposure to the new preschool daycare program will result in better developmental outcomes than exposure to programs delivered in the comparison daycares. Stated another way, an hour of exposure to the new preschool daycare program should be worth more in terms of developmental outcomes than the same amount of exposure to existing comparison daycare programs. We now turn our attention to a series of analyses designed to determine whether this is indeed the case.

Specification Details

The effect of dosage was evaluated by modifying the specification that was used in the main impact analyses for dosage by including fixed effects for dosage. We operationally defined dosage as the average number of hours a week the child spent in daycare during the four-month period immediately preceding the evaluation[68]. Children who did not spend any time in daycare were assigned the value of zero on this variable, which was mean-centred prior to its insertion into the model so as to avoid degrading the results of analyses unnecessarily with non-essential

[67] Moderation and mediation are related concepts. The idea behind moderation is that one variable is taken to control the effect of another either by making its effect more powerful or less powerful. Mediation refers to the case where a given variable, in this case treatment group, exerts its effect on an outcome indirectly by way of another variable, in this case program fidelity/quality.

[68] The other option that was considered was defining dosage as some function of the cumulative number of hours of exposure to the daycare setting over the course of the study. This second option presents some complications in that inserting such a variable into a regression analyses would require a transformation (e.g., logarithmic) to account for its skewed distribution and potentially non-linear relationship with the outcome measures (i.e., such a measure exaggerates the tendency towards decreasing returns with increased exposure). Moreover, preliminary correlational analyses indicated that a) the two definitions of exposure considered here were correlated in excess of .76 for the first four evaluation periods (Spearman Rho). Given the near-equivalency of the two definitions, we chose to use the simpler conceptualization which in any case captures dosage over a reasonably large interval (four months).

multicollinearity (Cohen, et al., 2003, pp. 261-266). We chose to centre based on time-specific means rather than the grand mean to accommodate seasonal variations in daycare exposure.

The five regression model specifications described in Section 5.1.2 were modified to include a) the direct effect of dosage, b) the interaction between dosage and time, and c) the triple interaction between dosage, time, and treatment group membership. The effect of interest is the triple interaction. We are specifically interested in estimating the effect of dosage on the developmental gains of children for each daycare group and taking the difference between the two estimates. In technical terms, estimation of this effect involves first estimating the DinD dosage effects (Time x Dosage) separately for the two daycare groups. This initial estimation is then pushed a step further by taking the difference between the DinD estimates. This secondary calculation yields what is called the Difference in Difference in Differences (DinDinD) estimate (for an example of the use of DinDinD estimators in program evaluation, see Monheit & Steinburg Schone, 2004). As with the DinD estimates reported in the preceding sections, one DinDinD estimate is reported for the Program vs. Comparison Daycare contrast for each post-test period. A global test of the joint effect of the DinDinD estimates is also reported: the F-test tests the three-way interaction among Time, Treatment Group, and Dosage. We were specifically interested in how the two daycare groups faired and therefore, we specifically examined their contribution to the pooled interaction effect. In other words, the regression models were applied to the complete sample of participants, but comparisons involving the Informal Care group were ignored[69].

The distribution of the dosage variable was statistically equivalent for the two daycare groups both within and across time. This is a necessary condition for the DinDinD estimates to be valid. If the distributions were not similar across groups a significant effect could result, but this effect would be difficult to interpret (hence the decision not to interpret effects involving the Informal Care group). Specifically, the effect could be driven by this initial difference in the distribution of scores rather than a cause that is of substantive interest. Consequently, the associated DinDinD estimates cannot be interpreted sensibly. Finally, we note that the statistical power of the three-way interaction tests (e.g., the DinDinD estimates) is less than that of the two-way tests (i.e., DinD estimates) reported earlier due to the expended degrees of freedom and the fact that the dosage variable is measured with error.

Impact of Dosage on Standardized Communication Scores (Full Sample)

The DinDinD estimates of the dosage effects are reported in Table 5.3. For the baseline period, the estimated effect of dosage for the entire sample was negative, but weak and not statistically significant, [b = -.006, SE = .007] (main effect for dosage not reported in Table 5.3). For this evaluation at least, the amount of exposure to the new preschool daycare program did not appear to be related to performance on the Communication scale.

Turning our attention to the results presented in the table, we observe that the DinDinD estimates of the dosage effect are consistently more positive for the Program Daycare group than the Comparison Daycare group, which are negative in sign. The differences apparent in this

[69] The complete sample was used for this analysis to increase the efficiency of the significance tests. The reported estimates are nevertheless specific to the G1 vs. G2 comparison. Comparisons involving the Informal Care group (G3) were not readily interpretable because the distribution of the dosage variable for this group was drastically different from that in the other two groups. The estimates arising from these comparisons were therefore not reported.

comparison are only statistically significant for the first post-test evaluation, though this effect is robust to the inclusion of covariates. The pattern of results is the same here as when dosage was defined earlier more simply as treatment group membership.

The DinDinD estimates provided in Table 5.3 require a broader context to be properly understood. We provide this context in what follows by presenting dosage-effect estimates at various levels of analyses. The estimates in question derive from Model 5, which includes the full list of covariates.

The size of the dosage effect across time and treatment group is of special interest. The DinD estimates of the dosage effect for the Program Daycare group were consistently positive across all post-test periods. At no point though did the DinD effect of dosage approach statistical significance for the Program Daycare group. The DinD estimates were [b = .013, SE = .010; b = .009, SE = .009; b = .022, SE = .013] for the first, second, and third post-test evaluations respectively (Time x Dosage results not reported in Table 5.3). From these estimates we can conclude that varying degrees of exposure to the new preschool daycare program has little effect on Communication score outcomes.

What appears to be driving the DinDinD effect reported in Table 5.3 is a negative effect of dosage for the Comparison Daycare group relative to the Program Daycare group. As noted earlier, this effect is specific to the DinDinD estimate for the first post-test period.

To take a concrete example, the DinD estimate of the dosage effect for the first post-test period indicates that every 10 additional hours a week spent in a program daycare is associated with an expected .13 standard deviation increase in Communication score (.13 = 10 hours x .013). This effect is not statistically significant, but the equivalent expected score for the Comparison Daycare group is significantly lower and of negative sign (i.e., -0.014). Importantly, the model estimates suggest that as the amount of exposure to daycare increases, so does the magnitude of the effect. At 10 hours of exposure above the sample average, the expected Communication score of a child enrolled in a comparison daycare is .27 standard deviations lower (.27 = 10 hours x .027) than that of the program group daycare. At 20 hours of exposure above the average, this gap increases to .54 standard deviations (.54 = 20 hours x .027), a medium-sized effect. In sum, the magnitude of the program's effect on development increases as a function of the number of hours spent on average in the daycare setting. For the highest levels of exposure, the magnitude of the effect estimated by the model is substantial. The distribution of the residuals did not suggest the presence of a non-linear relationship, which is consistent with the literature indicating that dosage effects are linear (National Institute of Child Health and Human Development Early Child Care Research Network, 2003).

Table 5.3: Year 1 Program Effects Based on DinDinD Estimator (Time x Group x Dosage) and the ÉPE–AD Communication Scores — French Version (Full Sample)

	Incremental Inclusion of Covariates – Errors Clustered by Daycare									
	(1)		**(2)**		**(3)**		**(4)**		**(5)**	
Program Effects	b	Robust SE	b	Robust SE	b	Robust SE	b	Robust SE	b	Robust SE
G1 vs. G2: Difference in magnitude of the Dosage effect at Baseline	-.002	.017	.001	.017	.002	.012	.003	.012	.002	.011
DinDinD 1st Post-test	-.031†	.010	-.033*	.012	-.026 m	.013	-.029*	.012	-.027*	.013
DinDinD 2nd Post-test	-.005	.015	-.003	.014	-.003	.013	-.003	.012	-.002	.012
DinDinD 3rd Post-test	-.013	.018	-.011	.019	-.017	.018	-.018	.018	-.019	.017
Wald F-tests for simple 3-way Interaction:	F (3, 18) = 4.45*		F (3, 18) = 3.91*		F (3, 18) = 1.74		F (3, 18) = 1.91		F (3, 18) = 2.36	
N participants:	243		240		239		239		232	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units. *Program Daycare group (G1) is in reference, which means that negative values of DinDinD denote an advantage for the Program Daycare group*. Standard errors were estimated via the SPSS implementation of the heterogeneity consistent 'robust' White estimator (White, 1980). Further, standard errors were clustered by daycare to adjust for the correlation in the residuals over time.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed, based on 20 degrees of freedom). Model specification details are provided in Section 5.1.3.

[a] Exposure here is defined as the average hours spent weekly in daycare during the 4 months preceding evaluation. [b] G1 here indicates enrolment to an Program Group daycare, G2 indicates enrolment with a Comparison Daycare group. G3 children (part of the Informal Care group) were involved in the estimation of the model, but the associated comparisons are not reported. The F-test for the 3-way interaction is based on the G1 and G2 groups only.

Impact of Dosage on Standardized ÉPE–AD scales (Sub-Sample)

The same analytical strategy as in the preceding section was used to evaluate the importance of dosage as a moderator of the program effect on all six ÉPE–AD scales. These analyses are based on the sub-sample of participants who met the criteria for completing the ÉPE–AD in French a sufficient number of times. Again, we report only the results of regression analyses with the full set of covariates. The DinDinD estimates of differences between the daycare groups in the effect of dosage are reported in Table 5.4.

Inspection of the DinDinD estimates reveals that the effect that was observed with the Communication scale is replicated with this sub-sample of participants. For the remaining outcomes, we verified whether the sign of the coefficients was consistent with a positive program effect. This was the case for all the scales except Physical Ability and Receptive Vocabulary. The Expressive Vocabulary subscale in particular yielded a program effect estimate of comparable magnitude to that obtained with the Communication scale, with an estimated gain of .026 standard deviations per hour of daycare exposure. This effect was not statistically significant however. As in preceding analyses, the results for the second and third post-test evaluations are not statistically significant. The effect involving the Communication scale is not discussed further due to the fact that the detailed analyses were presented in the previous section. Further, the remaining DinDinD estimates are not described further due to the fact they are not statistically significant.

We pause for a moment to consider the effect of daycare exposure independently of group membership. For all scales, the direct effect of dosage at baseline was weak and not statistically significant. The estimate of this effect for the Self-Awareness scales was [b = -.014, SE = .008], for the Cognitive Ability scale, [b = -.013, SE = .011], for the Physical Ability scale, [b < .001, SE = .016], for the Receptive Vocabulary subscale, [b = .003, SE = .006] and finally, for the Expressive Vocabulary subscale [b = -.014, SE = .008] (main effect results not reported in Table 5.4). We re-estimated the regression model this time dropping the three-way interaction term involving Group, Time, and Dosage. The purpose of these secondary analyses was to verify whether the importance of daycare exposure varies over time. For the Self-Awareness scale, this was not the case as the F-test of the two-way interaction between Time and Dosage was not statistically significant, [Wald F (3, 18) = .797]. Similarly, the Time by Dosage interaction was not statistically significant for the Cognitive Ability scale, [Wald F (3, 18) = 1.06], or the Physical Ability scale, [Wald F (3, 18) = .733]. The effect of dosage over time was found to be similarly invariant for the Receptive [Wald F (3, 18) = .147] and the Expressive [Wald F (3, 18) = 1.46] Vocabulary scales (Time x Dosage results not reported in Table 5.4). In sum, these secondary analyses offer little hint that the average number of hours spent in daycare over the preceding four months was a useful predictor of either school readiness outcomes or vocabulary scores. The only exceptions include the DinDinD effect that was observed for the Communication scale and to a lesser extent the Expressive Vocabulary subscale (i.e., the associated non-significant trend).

Table 5.4: Year 1 Program Effects Based on the DinDinD Estimator (DinDinD) of Program Effects (Time x Group x Dosage) for Standardized Subscale Scores of the ÉPE–AD (French Test Takers Only)

	Analyses by Subscale – Errors Clustered by Daycare											
	Communication		**Self-Awareness**		**Cognitive Ability**		**Physical Ability**		**Receptive Vocabulary**		**Expressive Vocabulary**	
Program Effects	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
G1 vs. G2: Difference in magnitude of the Dosage effect at Baseline	.001	.011	.004	.013	.012	.014	-.001	.017	-.020	.011	.006	.014
DinDinD 1st Post-test	-.026 *	.013	-.015	.014	-.019	.013	.001	.024	.001	.015	-.027	.016
DinDinD 2nd Post-test	-.006	.012	-.006	.016	.002	.016	.007	.030	.004	.016	.008	.019
DinDinD 3rd Post-test	-.029 m	.015	-.038 m	.021	-.053	.018	-.028	.024	-.001	.016	-.035	.021
Omnibus Wald F-tests:	F (3, 18) = 2.67 m		F (3, 18) = 2.92 m		F (3, 18) = 5.38†		F (3, 18) = 6.19†		F (3, 18) = .113		F (3, 18) = 4.95*	
N participants:	219		217		219		218		218		217	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units. *Program Daycare group (G1) is in reference, therefore negative values of DinDinD denote an advantage for the Program Daycare group*. Standard errors were estimated via the heterogeneity consistent 'robust' White estimator (White, 1980). Further, standard errors here are clustered by daycare to adjust for the correlation in the residuals over time. Model specification details are provided in Section 5.1.3.

m denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed, 20 degrees of freedom)

[a] Exposure here is defined as the average hours spent weekly in daycare during the 4 months preceding evaluation. [b] G1 here indicates membership in the Program Daycare group, G2 indicates membership in the Comparison Daycare group.

Summary

The pattern of results is consistent with the idea that the new preschool daycare program had a protective effect with respect to amount of daycare exposure, but *the effect was specific to the standardized Communication Scores.* For this scale, greater levels of exposure to the Comparison Daycare group setting were associated with developmental changes that were correspondingly less positive, and in fact slightly negative. In contrast, level of exposure to the new preschool daycare program is estimated to be a neutral factor with a slightly positive and non-significant effect. The results do not point to a dramatically positive program effect. However, they suggest that for the four month period preceding the first post-test, the tested preschool daycare program had a positive effect relative to existing programs delivered in comparison daycares.

Admittedly, the dosage analyses did not corroborate all of the effects that were observed with the simple group comparisons. While the DinDinD estimate for the Expressive Vocabulary subscale was sizeable (if non-significant), none of the other estimates associated with the first post-test period showed a hint of the expected effect. The failure to find an exact replication of the pattern that was observed with the main impact analyses should not be counted as evidence against the validity of the initial set of results as the tests were not sufficiently powerful for detecting small effects. Nevertheless, for the reasons noted above, such findings would have bolstered the case in favour of their validity. As it stands, such positive evidence was obtained only for the Communication scale.

Fine-grained Definitions of Program Integrity: Daycare Program Fidelity and Quality

The hypothesis behind the main impact analyses reported in this chapter's first section was that enrolment in a new preschool Program Daycare would lead to superior development on school-readiness outcomes. This formulation of the research hypothesis implies sufficient exposure for the program to have its effect. In this section, membership in one of the three treatment groups *was replaced* in the regression-model specification by indices related to the level and quality of the tested program. These indices are part of a model developed by Dane and Schneider (1998) for assessing a program or intervention's integrity (for details on these measures see Section 3.5.6). The fidelity indices are interpreted as a proportion of program elements that are correctly in place. The quality indices make finer distinctions. Whereas fidelity indices pertain to the presence/absence of program components, quality indices reflect how well core program components were implemented along a seven-point scale where 1 indicates inadequate level of care and 7 indicates an excellent level of care.

We anticipate that the Program Daycares and the Comparison Daycares will differ on the dimensions of both program fidelity and program quality. This prediction is tested in the first series of analyses. Further, we anticipate that the degree to which daycares vary on the measured aspects of program fidelity and quality will be predictive of developmental outcomes. This hypothesis is tested in a second series of analyses. If it is confirmed, support will have been obtained for the contention that the dimensions of program fidelity and quality targeted by the program have an impact on the development of children. As indicated in Section 5, these analyses may be construed as a more sensitive version of the DinD analyses by treatment group. The results of analyses on program fidelity and the analyses on program quality are presented in separate sections.

Specification Details & Presentation of the Results

Two series of analyses were conducted that were qualitatively different in type. The ***first series of analyses*** takes daycare site as the unit of analysis. Its purpose is to evaluate the extent to which the analyses by treatment group represents a comparison of daycares that truly differ with respect to the fidelity and quality dimensions considered here. In other words, it serves as a type of manipulation check for the program tested, a formal test of differentiation (see also the analyses of differentiation provided in the *Project Implementation Report*, Bérubé et al., 2014). The ***second series of analyses*** takes observations from individual children and may be considered a replication of the analyses by treatment group. Its distinctive feature is that, instead of treatment group membership, it is the estimates of the fidelity and quality of program delivery that are employed in the calculation of the DinD estimator. We expect this alternative way of conceptualizing the effect of the tested preschool daycare program to produce the same pattern of results as in the treatment group analyses. Finally, we predict that differences among the two daycare groups will be dramatically reduced once program fidelity and quality are controlled, validating the result.

Accordingly, a ***first series of analyses*** was conducted to evaluate a) the fidelity and the quality with which the tested program was implemented in program daycares; and b) the extent to which the program group daycares distinguish themselves from the comparison daycares in terms of fidelity and quality indices. As indicated in Section 3.5.6, we distinguish between two types of fidelity and four types of quality indices. The selected fidelity dimensions were Structural and Content Fidelity, while the selected quality dimensions were Structural Quality, Educative Quality, Educator Sensitivity, and Quality of Reading[70]. For fidelity, we were in a position to estimate the degree of change for the program group daycares over time. In all cases, we were able to estimate a difference between the program daycares and the comparison daycares.

Because issues of clustered sampling are not relevant for these analyses, the effects were tested using unadjusted tests of statistical significance. Both the parametric and non-parametric rank-based tests are reported as a check against the violation of distributional assumptions with this small sample. A distinguishing feature of the non-parametric tests is that they are sensitive to the relative rank of the quality scores but not to their absolute differences[71]. Correlated t-tests (parametric) and Wilcoxon rank-sum tests (non-parametric) are reported for tests for change over time. Welch's test (a t-test that does not assume homogeneity of variance or equal group sizes) and the Mann-Whitney U rank-order statistic (non-parametric) are reported for differences between independent groups[72]. Pearson r (parametric) and Kendall's tau (non-parametric) are reported as measures of association.

[70] Details as to the calculation of these indices can be found in Section 3.5.6 as well as in the companion document *Project Implementation Report* (Bérubé et al., 2014) submitted to HRSDC in May 2010.

[71] Large discrepancies in the effect size estimated using these two methods would indicate that the relative differences between scores is relevant to determining the size of the effect. Such a result is informative but does not necessarily indicate a problem with the result of the parametric test. Visual inspection indicated that the distribution of scores within each cell was quite reasonable given the small number of observations.

[72] Independent tests were conducted for each time period. We chose not to perform a more synthetic analyses (e.g., repeated measures ANOVA) due to the fact that the fidelity scores for Comparison daycares are invariant for the pre-test and the first post-test period due to the imputation strategy adopted here.

The small number of daycares (N = 15) means that the significance tests lack the power to reliably detect real effects[73]. To compensate for this lack of power, the nominal probability level for statistical significance was fixed at .10 rather than .05 for these analyses. Nevertheless, null effects should be interpreted with additional care in this situation. Finally, we emphasize that the results presented in this sub-section may differ from those presented elsewhere[74]. This source of this discrepancy arises from the fact that the present analyses are based on the sub-sample of communities included in the impact analyses.

In a ***second series of analyses***, we estimated the impact of the fidelity and quality indices on the developmental outcomes of children. For these analyses, the exact same specifications and analytical strategy were employed as in the analyses by treatment group. The only difference is that the fidelity and quality indices replaced the treatment group factor in the estimation of effects. The impact of each quality/fidelity index was estimated in an independent set of analyses, which means that we did not control for the common effects of these indices. We were primarily interested in determining whether one of the indices would distinguish itself in predicting developmental outcomes. For this section, the results of a secondary series of analyses are also reported where we entered DinD effects simultaneously for both the treatment group effect and the fidelity or quality indices. By controlling for the shared predictive variance of treatment group and quality/fidelity, we intended to test the degree to which the treatment group effects reported earlier would be reduced with the inclusion of the fidelity/quality indices. If the treatment group effect is the result of manipulating the dimensions tapped by fidelity and quality, then a substantial or total reduction in the effects should be observed.

Inter-correlations of the Fidelity and Quality Indices

Before presenting the results of the preliminary and secondary impact analyses, an examination of the observed correlations among these indices is informative (see Table 5.5). We observe positive correlations among all quality indices, which is consistent with the idea that they are all tapping elements of a common construct. Note that many of the correlations are small and none of the correlations approach unity ($r > .70$), which supports our initial strategy of considering these indices separately as predictors of development.

[73] In the *Project Implementation Report* (Bérubé et al., 2014), descriptive analyses were conducted at the level of individual classes rather than daycares. Another strategy was required for the quantitative analyses reported in this document because classes were nested within daycare site, which if ignored, could lead to over-estimated precision in effect estimates. The issue was resolved by aggregating class data to obtain a single score for each daycare site.

[74] *Project Implementation Report* (Bérubé et al., 2014). In the present report, we employed the 'flexible' means described in the aforementioned report, and aggregated this classroom data to the level of daycare site.

Table 5.5: Correlations among the Fidelity and Quality Indices

	1	2	3	4	5	6
Fidelity						
1- Structural	1	.64*	.39	.64*	.23	.66*
2- Content	.55*	1	.16	.59*	.36	.44
Quality						
3- Structural	.39	.16	1	.41	.51	.41
4- Educative	.64*	.59*	.42	1	.61*	.21
5- Educator Sensitivity	.26	.36	.51	.61*	1	.09
6- Reading	.66*	.44*	.41	.21	.09	1

Note: Daycares are the unit of analysis (N = 15). Fidelity estimates are based on data collected between second and third evaluation. Pearson r correlations are reported in the lower diagonal, while the non-parametric Kendall tau correlations are reported in the upper diagonal.

First Series of Analyses — I: Structural and Content Fidelity across Program and Comparison Daycares

The descriptive statistics and significance tests for comparisons of the fidelity indices across daycare groups are reported in Table 5.6. As noted earlier, both parametric and non-parametric tests of association are reported.

Examination of Table 5.6 reveals the lack of statistical power in analyses testing for change relative to the baseline period for the program daycares (see the bottom part of Table). We emphasize that the change in question captures the evolution from early to later implementation and not the true difference between pre- and post-evaluation. Quantitative estimates of these changes are intrinsically interesting therefore we report the results of these analyses despite their low power.

To begin, we consider the change in the program daycares over time. In this case, we observe that on average the program daycares increased their Content Fidelity by .05 and their Structural Fidelity by .15 relative to the baseline by the first post-test. These effects indicate gains of 5 and 15% respectively in terms of the total number of program elements that are in place in the program daycares. By the second post-test, the estimated gain in Content Fidelity dropped negligibly by .01 while the gains in Structural Fidelity dropped to .08. In sum, we can observe that program daycares show gains on the fidelity indices over time and that these gains dropped off following the initial post-test period.

Examining the contrasts comparing the two types of daycare, we note that the parametric tests indicate that the program daycares were substantively different from the comparison daycares just a few months after the start of the program. The difference was .12 for Content Fidelity and .39 for Structural Fidelity. This result indicates that, shortly after the program began to be implemented, the program daycares had significantly more of the core elements of the tested program in place than would be expected based on the estimates derived from the comparison daycares.

The difference between daycare groups was estimated again using the indexes computed from observations of the first post-test period for both groups. For both fidelity indices, the first post-test period showed the most separation between the groups. For this time interval, the difference between groups increased to .15 for Content Fidelity and .54 for Structural Fidelity. Both differences were statistically significant at the .10 level according to the parametric tests. The Structural Fidelity effect in particular was attested by the statistical significance of the non-parametric test. The effect then is most robust for the first post-test period.

For the second post-test period, the average difference between daycares groups remains unchanged at .15 for the Content Fidelity index. Similarly, the average difference between groups on the Structural Fidelity index drops only slightly to .53. However, relative to the first post-test period, the pattern of statistical significance changes. Specifically, the test based on Content Fidelity indicates that it is no longer statistically significant. Inspection of the standard errors suggests that this change is due to increased variability for the program group daycares on this measure. The difference between groups based on Structural Fidelity remains statistically significant for both the parametric and non-parametric test.

Table 5.6: Program Fidelity between Groups and Across Time for the Program Daycare Group

	Fidelity Dimensions			
	Content		***Structural***	
Between Daycare Groups	**Mean**	**SD**	**Mean**	**SD**
Baseline[c]				
Program Daycare (N = 4)	0.67	0.01	0.78	0.10
Comparison Daycare (N = 11)	0.55	0.14	0.39	0.15
Welch's[a]	F (1, 10) 7.94[*]		F (1, 8) = 31.63 [†]	
Mann_Whitney U[b]	Z = 1.31		Z = 2.76	
First Post-Test				
Program Daycare (N = 4)	0.72	0.10	0.93	0.05
Comparison Daycare (N = 11)	0.57	0.14	0.39	0.15
Welch's[a]	F (1, 7) = 5.30[m]		F (1, 12) = 102.10[†]	
Mann_Whitney U[b]	Z = -1.63		Z = -2.89[†]	
Second Post-Test				
Program Daycare (N = 4)	0.71	0.13	0.86	0.14
Comparison Daycare (N = 11)	0.56	0.12	0.33	0.14
Welch's[a]	F (1, 5) = 3.95		F (1, 5) = 40.24[†]	
Mann_Whitney U[b]	Z = -1.64		Z = -2.89[†]	
Third Post-Test	NA	NA	NA	NA

	Fidelity Dimensions			
	Content		*Structural*	
Between Daycare Groups	**Mean**	**SD**	**Mean**	**SD**
Within the Program Daycare group				
Program Group: Baseline vs. 1st Post-Test	paired t (3) = 0.93 Wilcoxon = 2.00, Z = .45		paired t (3) = 2.10 Wilcoxon = 9.00, Z = 1.46	
Program Group: Baseline vs. 2nd Post-Test	t (3) = 0.56 Wilcoxon = 7.00, Z = .73		t (3) = 1.18 Wilcoxon = 2.00, Z = .45	

Note: Data were available only for the first post-test period for the comparison daycares. Given the absence of intervention for these daycares, we assumed stability across time and imputed the pre-test values based on those of the post-test. For both daycare groups, fidelity data was only available up to the second post-test period.

[a] Welch's test is robust to the bias that can arise from a combination of unequal sample sizes and heterogeneous variance.

[b] Test compares the mean ranks of the scores for the two groups, making no assumptions about the distribution of these scores.

[c] For this set of comparisons only, the baseline period scores do not constitute a true pre-test. Elements of the tested program were already in place by this time. Though in place, there would have been little opportunity for an effect to be observed on the children at this point. The validity of the pre-test evaluation is therefore not in question.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed).

In sum, these preliminary analyses of the fidelity estimates indicate that between 70 and 80% of the core program elements were in place in program daycares at baseline and this proportion increased by 5 to 15% the first post-test with some decrements observed in the second post-test. The difference between program daycares and comparison daycares increased for both measures relative to the baseline period and remained relatively stable to the second evaluation. The conclusions that can be drawn from this observation are limited given that the number of daycares that received the tested program was insufficient to provide a fair test. Nevertheless, the findings indicate that the program daycares had significantly more of the core elements of the tested program in place than the comparison daycares.

First Series of Analyses — II: Quality Between Program and Comparison Daycares

While the fidelity indices present a very crude picture of daycare quality in terms of whether or not certain key elements of the tested program are in place, quality estimates allow for finer distinctions between daycare groups with respect to how the target program elements were implemented. Quality estimates reflect the average quality of the implementation of different core components of the program along a seven-point scale ranging from a child care program of inadequate quality (1) to a child care program of excellent quality (7). Fours dimensions are considered: a) Structural Quality, b) Educative Quality, c) Educator Sensitivity, and d) Quality of Reading[75]. One estimate for each dimension was available for the initial 12-month period post-intervention. Descriptive statistics and significance tests of the difference between daycare groups on these indices are reported in Table 5.7.

Inspection of the group averages indicates a Program Daycare group advantage for all quality variables. This apparent difference in average quality was attested by the parametric tests for the Structural and Educative Quality dimensions at the .10 level of statistical significance. The strongest and most statistically significant effect was associated with the Quality of Reading. The test was not sufficiently sensitive to detect the difference between groups on Educator

[75] A more detailed description of quality indices can be found in Section 3.5.6.

Sensitivity, which showed much more variability within groups than the other dimensions (see the standard deviation estimates). The results are consistent with the claim that on average, children enrolled in daycares offering the new preschool program received a higher quality program than those enrolled in comparison daycares, especially with respect to how reading activities are conducted.

Table 5.7: Comparing the Program Daycares and Comparison Daycares on Quality

	Quality Dimensions							
	Structural		**Educative**		**Educator Sensitivity**		**Reading**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Program Daycares (N = 4)	6.42	0.42	5.71	0.82	5.88	2.25	5.88	1.44
Comparison Daycares (N = 11)	5.73	0.81	4.38	1.42	4.55	2.45	1.27	0.90
Welch's[a]	F (1, 10) = 4.56 [m]		F (1, 9) = 5.03 [m]		F (1, 5) = .975		F (1, 3) = 35.90[†]	
Mann_Whitney U	Z = 1.46		Z = 1.60		Z = 1.04		Z = -3.35[†]	

[a] Welch's test is robust to the bias that can arise from a combination of unequal sample sizes and heterogeneous variance.
[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed)

Second Series of Analyses — I: Impact of Program Fidelity on Standardized ÉPE–AD Scales

As in the analyses by treatment group, we are interested in evaluating the impact of program fidelity on the developmental trajectories of our daycare groups. In technical terms, this involves estimating a DinD effect for the two facets of fidelity, which is to say Structural and Content Fidelity. We were also interested in verifying whether the observed DinD impacts of fidelity are redundant with those associated with treatment group. In technical terms, this involved re-estimating the treatment group impact in a model simultaneously with the effects associated with fidelity.

Estimates of the tested preschool daycare program effects based on the fidelity indices are reported separately for each outcome variable. The Communication scale analyses are based on participants who belonged to one of the daycare treatment groups. The sample used in the analyses of the remaining scales was slightly smaller as it excludes those children who did not complete the ÉPE–AD in French at least twice during the first year. For reasons of economy, we report only the results that were obtained using the full set of covariates (Model 5 as outlined in Section 5.1.3).

Communication Scale (Standardized): Fidelity indices were entered as predictors of Communication scores. Unlike previous analyses, the standardized Communication scores estimates are only presented once, for the full-sample analyses (see Table 5.8). The pattern of results is quite similar for both set of analyses, as we would expect given that the difference in samples is only 6 children.

The estimated effect of both types of fidelity is negative for the baseline period, but is statistically significant for Structural Fidelity only. The negative estimates of fidelity could be

explained by way of a selection bias whereby children with weaker communication skills are disproportionately enrolled in daycares characterized by high Structural Fidelity. Evidently, the covariates that were inserted into the analyses were not sufficient to correct for this selection bias in the case of Structural Fidelity. In any case, the direction of association reverses by the first post-test period where the DinD estimates indicate that gains in Communication score from the baseline period to the first post-test are positively associated with both fidelity indices[76]. As with the main impact analyses (reported in Section 5.1.2), the DinD estimates for subsequent time periods are not statistically significant.

The magnitude of the effect can be better appreciated if we solve the regression equation for a given difference in Structural Fidelity. This is easily achieved by multiplying the desired difference in Structural Fidelity with the DinD estimate reported in the table. When this calculation is performed for the effect of Structural Fidelity (first post-test), we observe a .068 standard deviation effect for a .10 difference in Structural Fidelity (i.e., .068 = .68 x .10). Similarly, a difference of .30 on the Structural Fidelity index corresponds to an estimated treatment impact of approximately .20 standard deviations, which is considered a small effect by conventional benchmarks. When the same calculation is performed using the Content Fidelity estimate, the magnitude of the effects is estimated to be much larger on average. A difference of only .20 on the Content Fidelity index is required for an equivalent estimated gain of .20 standard deviations on the Communication scale.

Table 5.8: DinD Estimate of the Fidelity Effect on Standardized Communication Scores (Time x Fidelity Indices) — French Version

	Fidelity Dimensions – Errors Clustered on Daycare			
	Structural Fidelity		**Content Fidelity**	
	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	-.482*	.184	-.367	.518
DinD 1st Post-test	.680*	.243	1.04 m	.554
DinD 2nd Post-test	.368	.224	-.575	.598
DinD 3rd Post-test	.175	.208	-.400	.753
Wald F	F (3, 12) = 2.28		F (3,12) = 3.57 m	
N	154		154	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

m denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed, 14 degrees of freedom).

Once it was established that the main impact results were replicated using fidelity estimates as a predictor, we considered whether controlling statistically for fidelity would eliminate the statistically significant effect that was observed for the first post-test period based on the comparison of daycare groups (i.e., $DinD_{1stPost\text{-}test}$ = -.314*, SE = .151). If the tested program has its effect via this route, then we would expect this to be the case. For these secondary analyses, we used a Global Index of Fidelity (Structural Fidelity and Content Fidelity are merged in one index). As expected, the DinD estimate of the Time by Treatment Group effect was substantially

[76] Here we are interpreting the Content Fidelity effect as statistically significant because it is marginal at .10 for the analyses with the complete sample and significant at .05 for the sub-sample analyses.

reduced and was not statistically significant when we controlled for the global fidelity estimate, [adjusted $DinD_{1stPost\text{-}test}$ = -.101, SE = .233].

Self-Awareness Scale (Standardized): In the same manner as in the preceding analyses, fidelity indices were entered as predictors of Self-Awareness scores. The observed pattern of results mirrors that obtained in the analyses of Communication scores. The only notable exception is the fact that the fidelity measures do not have a statistically significant effect at baseline for these analyses. The estimated fidelity impacts that resulted from these analyses are reported in Table 5.9.

Table 5.9: DinD Estimate of the Fidelity Effect on Standardized Self-Awareness Scores (Time x Fidelity indices) — (French Test-Takers Only)

	Fidelity Dimensions – Errors Clustered on Daycare			
	Structural Fidelity		**Content Fidelity**	
	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	-.410 [m]	.202	-.292	.541
DinD 1st Post-test	.674[†]	.197	1.017[*]	.463
DinD 2nd Post-test	.423 [m]	.200	-.605	.605
DinD 3rd Post-test	.040	.225	-.383	.879
Wald F	F (3, 12) = 5.91[†]		F (3, 12) = 5.73[*]	
N	147		147	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

[m] denotes significance at p < .10, [*] p < .05, [†] p < .01 (all tests two-tailed, 14 degrees of freedom).

Given that the significant DinD effects that were observed in the analyses by treatment group have been replicated here, a logical extension of these analyses is to verify whether the fidelity effect is redundant with the treatment group effect reported previously, [$DinD_{1stPost\text{-}test}$ = -.337[†], SE = .084]. A Global Index of Fidelity was used in these secondary analyses. Results indicated that controlling for fidelity causes a reduction in the magnitude of this effect which becomes non-significant, [adjusted $DinD_{1stPost\text{-}test}$ = .169, SE = .169]. This is the result that would be expected if treatment group membership exerts its effect by way of the fidelity of the program's implementation.

Cognitive Ability Scale (Standardized): Fidelity indices were then entered as predictors of Cognitive Ability scores. The estimated fidelity impacts that resulted from these analyses are reported in Table 5.10. The results partially replicate those found in the main impact analyses reported in Table 5.2.

Table 5.10: DinD Estimate of the Fidelity Effect on Standardized Cognition Scores (Time x Fidelity Indices) — (French Test-Takers Only)

	Fidelity Dimensions – Errors Clustered on Daycare			
	Structural Fidelity		**Content Fidelity**	
	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	-.429	.251	.108	.521
DinD 1st Post-test	.480*	.210	.064	.535
DinD 2nd Post-test	.264	.236	-.344	.797
DinD 3rd Post-test	.221	.230	-.447	.808
Wald F	F (3, 12) = 1.99		F (3, 12) = .184	
N	148		148	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed, 14 degrees of freedom).

For these analyses, the effect is specific to Structural Fidelity. Accordingly, we verified whether the effect observed in the main impact analyses for the first post-test period, [$DinD_{1stPost\text{-}test} = -.279^{\dagger}$, SE = .293], remained statistically significant once we controlled for Structural Fidelity. The adjusted DinD estimate shows signs of a suppression effect (the estimated magnitude of the effect is larger) albeit statistically non-significant, [adjusted $DinD_{1stPost\text{-}test} = -.496$, SE = .293]. A suppression effect occurs when one regressor (in this case Structural Fidelity) suppresses error variance in another to improve the overall fit of the regression model. This secondary issue is not interpreted further here beyond acknowledging that the test does not provide corroboration of our original hypothesis.

Physical Ability Scale (Standardized): Fidelity indices were also entered as predictors of the Physical Ability scores. The estimated fidelity impacts that resulted from these analyses are reported in Table 5.11.

As in preceding analyses, there is no impact on Physical Ability for the first post-test period. This absence of an effect is consistent with expectations and with the main finding of the treatment group analyses. Contrary to expectations, we observe an effect of Structural Fidelity for the third post-test. This effect might be dismissed as an aberration except that the DinD estimates of Structural Fidelity systematically increase with the passage of time. That said, the Structural Fidelity index does not represent any obvious correlate of the development of Physical Ability.

Table 5.11: DinD Estimate of the Fidelity Effect on Standardized Physical Ability Scores (Time x Fidelity Indices) – (French Test-Takers Only)

	Fidelity Dimensions – Errors Clustered on Daycare			
	Structural Fidelity		**Content Fidelity**	
	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	-.213	.335	.315	.692
DinD 1st Post-test	.074	.361	-.491	.623
DinD 2nd Post-test	.404	.325	-.374	.773
DinD 3rd Post-test	.640*	.245	.212	1.024
Wald F	F (3, 12) = 1.98		F (3, 12) = 2.85	
N	148		148	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

m denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed, 14 degrees of freedom).

Receptive and Expressive Vocabulary Scales (Standardized): Lastly, fidelity indices were entered as predictors of the development of the Receptive and Expressive Vocabulary scores. The estimated fidelity impacts that resulted from these analyses are reported respectively in Tables 5.12 and 5.13.

The results fail to indicate that fidelity was reliably associated with Receptive Vocabulary scores. The effect at baseline and the DinD estimates are not statistically significant. This result is consistent with the finding in the main impact analyses that the Program Daycare group and the Comparison Daycare group could not be distinguished based on this variable.

Table 5.12: DinD Estimate of the Fidelity Effect on Standardized Receptive Vocabulary Scores (Time x Fidelity Indices) – (French Test-Takers Only)

	Fidelity Dimensions – Errors Clustered on Daycare			
	Structural Fidelity		**Content Fidelity**	
	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	.180	.431	-1.02	.681
DinD 1st Post-test	-.384	.695	.315	.789
DinD 2nd Post-test	.349	.537	.659	.830
DinD 3rd Post-test	.061	.350	1.019	.668
Wald F	F (3, 12) = 2.88 m		F (3, 12) = .882	
N	148		148	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

m denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed).

In contrast, the analyses of Expressive Vocabulary scores indicate a strong and statistically significant advantage associated with increasing levels of fidelity for the first post-test DinD estimate. A .20 difference in fidelity corresponds roughly to a small (.20 standard deviation) and

moderate (.40 standard deviation) boost in the developmental gains of children on the Expressive Vocabulary measures. As with the main impact analyses reported in Section 6.1.2, initial gains are observed for the first post-test period only and are not found for the second or third post-test evaluation for either fidelity estimate.

Table 5.13: DinD Estimate of the Fidelity Effect on Standardized Expressive Vocabulary Scores (Time x Fidelity Indices) – (French Test-Takers Only)

	Fidelity Dimensions – Errors Clustered on Daycare			
	Structural Fidelity		**Content Fidelity**	
	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	-.697*	.247	-.454	.712
DinD 1st Post-test	1.17†	.299	2.07*	.767
DinD 2nd Post-test	.454	.333	-.941	.763
DinD 3rd Post-test	.124	.247	-.582	.905
Wald F	F (3, 12) = 9.06†		F (3, 12) = 9.71†	
N	147		147	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

m denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed).

Next, we verified whether statistically controlling for the observed fidelity effects substantially reduce or eliminate the DinD effect that was reported in the earlier analyses by treatment group, [$DinD_{1stPost\text{-}test}$ = -.543†, SE = .144]. The Global Index of Fidelity was used for these secondary analyses. Results showed that controlling for fidelity causes a reduction in the magnitude of this effect on Expressive Vocabulary measures, which becomes non-significant, [adjusted $DinD_{1stPost\text{-}test}$ = -.060, SE = .238].

Impact of Quality Indices on Standardized ÉPE–AD scales

In this section, we repeat the analyses reported in the preceding section this time estimating the impact of various Quality indices. Estimates derived from regression models of each outcome variable are reported in turn. A special status was accorded to Quality of Reading as a predictor; it is therefore reported last.

Communication Scale (Standardized): The estimated impact of the three quality indices for the full sample of participants who were enrolled in a francophone daycare at a given time period is reported in Table 5.14[77]. The magnitude of the impact is invariably largest for the first post-test period and it is statistically significant for the Educative Quality and Educator Sensitivity indices. This result is consistent with our expectation that these indices would be the most consistently related to the development of communication skills. In both cases, a two-unit difference in quality score is required to achieve a small effect (i.e., an effect of .20 standard deviations). Results from these secondary analyses suggest that controlling for global quality

[77] As in the analyses by fidelity, we do not report the analyses based on the sub-sample who completed the remaining ÉPE–AD scales in French because the two samples differ, in this case, by only a single case.

neutralizes the DinD effect for the first post-test period observed in the main impact analyses comparing daycare groups, [adjusted $DinD_{1stPost\text{-}test}$ = -.166, SE = .097].

Table 5.14: DinD Estimate of the Quality Effect on Standardized Communication Scores (Time x Quality Indices) – French Version

	Quality Dimensions – Errors Clustered on Daycare					
	Structural		**Educative**		**Educator Sensitivity**	
	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	-.032	.102	-.075	.052	-.026	.028
DinD 1st Post-test	.218 [m]	.115	.148[†]	.048	.098[†]	.024
DinD 2nd Post-test	-.025	.114	.014	.073	.004	.034
DinD 3rd Post-test	-.141 [m]	.056	.038	.081	.011	.035
Omnibus Wald F	F (3, 12) = 6.18[†]		F (3, 12) = 4.12[*]		F (3, 12) = 10.62[†]	
N	154		154		154	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units. Model estimated with full set of covariates, but excluding group membership, and the group by Time interaction.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed, 14 degrees of freedom).

Self-Awareness Scale (Standardized): Quality indices were entered as predictors of the development of Self-Awareness scores. The estimates that result from these analyses are reported in Table 5.15. Again, a statistically significant impact for all quality indices is observed for the first post-test period. Results from the secondary analyses suggest that controlling for global program quality produced an estimated treatment effect that is reduced in magnitude by 50%, but remains statistically significant at the .05 level, [adjusted $DinD_{1stPost\text{-}test}$ = -.271, SE = .085]. In other words, approximately half of the treatment effect is attributable to global program quality, but the effect is not completely reducible to this variable.

Table 5.15: DinD Estimate of the Quality Effect on Standardized Self-Awareness Scores (Time x Quality Indices) – French Test Takers Only

	Quality Dimensions – Errors Clustered on Daycare					
	Structural		**Educative**		**Educator Sensitivity**	
	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	.057	.168	-.082	.061	-.010	.032
DinD 1st Post-test	.192*	.072	.110†	.036	.061*	.028
DinD 2nd Post-test	-.013	.090	-.015	.061	.001	.027
DinD 3rd Post-test	-.150m	.076	.015	.088	.015	.033
Wald	F (3, 12) = 9.09†		F (3, 12) = 4.20*		F (3, 12) = 1.71	
N	147		147		147	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

m denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed).

Cognitive Ability Scale (Standardized): In the same manner as in the preceding analyses, Quality indices were entered as predictors of the development of Cognitive Ability scores. The estimates that result from these analyses are reported in Table 5.16. No statistically significant relationship was discerned. The effects observed with the treatment condition comparisons are therefore unattested by the quality measures. Re-estimation of the earlier treatment effect controlling for global quality was not attempted because quality was not a significant predictor of Cognitive Ability.

Table 5.16: DinD Estimate of the Quality Effect on Standardized Cognitive Ability Scores (Time x Quality indices) – French Test Takers Only

	Quality Dimensions – Errors Clustered on Daycare					
	Structural		**Educative**		**Educator Sensitivity**	
	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Average Effect	.044	.151	-.034	.056	.019	.021
DinD 1st Post-test	-.058	.080	.003	.042	-.037	.028
DinD 2nd Post-test	-.119	.107	.042	.076	-.023	.034
DinD 3rd Post-test	-.125	.092	-.017	.049	-.041	.031
Wald	F (3, 12) = .582		F (3, 12) = .913		F (3, 12) = .778	
N	148		148		148	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

m denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed).

Physical Ability Scale (Standardized): In the same manner as in the preceding analyses, quality indices were entered as predictors of the development of Physical Ability scores. The estimates that results from these analyses are reported in Table 5.17. As with most of the analyses reported previously for this outcome, no relationship was discerned.

Table 5.17: DinD Estimate of the Quality Effect on Standardized Physical Ability Scores (Time x Quality indices) – French Test Takers Only

	Quality Dimensions – Errors Clustered on Daycare					
	Structural		**Educative**		**Educator Sensitivity**	
	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	.160	.161	.031	.054	.002	.042
DinD 1st Post-test	-.132	.118	.018	.064	.014	.035
DinD 2nd Post-test	.056	.126	.026	.078	.003	.039
DinD 3rd Post-test	.000	.130	.063	.051	.010	.046
Wald F	F (3, 12) = 1.24		F (3,12) = .523		F (3,12) = .145	
N	148		148		148	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed).

Receptive and Expressive Vocabulary Scales (Standardized): Lastly, quality indices were entered as predictors of the development of the Receptive and Expressive Vocabulary scores. The estimated impacts for program quality that resulted from these analyses are reported respectively in Tables 5.18 and 5.19. The results indicate that both Structural and Educative Quality are related to Expressive Vocabulary scores. It is interesting to note that Educator Sensitivity emerged as a significant predictor of Communication scores, but is not of Expressive Vocabulary scores. Conversely, Structural Quality was not found to be predictive of Communication scores, but was found to be strongly related to Expressive Vocabulary scores in the present analyses. None of the quality indices were significantly related to Receptive Vocabulary scores.

Table 5.18: DinD Estimate of the Quality Effect on Standardized Receptive Vocabulary Scores (Time x Quality indices) – French Test Takers Only

	Quality Dimensions – Errors Clustered on Daycare					
	Structural		**Educative**		**Educator Sensitivity**	
	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	.016	.109	-.136	.087	-.056	.037
DinD 1st Post-test	.060	.099	.089	.099	.121 [m]	.065
DinD 2nd Post-test	.266 [m]	.142	.145	.109	.097 [m]	.046
DinD 3rd Post-test	.054	.080	.097	.087	.055	.032
Wald F	F (3, 12) = 1.43		F (3, 12) = .600		F (3, 12) = 1.52	
N	148		148		148	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed).

Table 5.19: DinD Estimate of the Quality Effect on Standardized Expressive Vocabulary Scores (Time x Quality indices) – French Test Takers Only

	Quality Dimensions – Errors Clustered on Daycare					
	Structural		**Educative**		**Educator Sensitivity**	
	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	-.173	.169	-.095	.055	-.006	.034
DinD 1st Post-test	.419†	.119	.203†	.052	.094	.058
DinD 2nd Post-test	.033	.140	.016	.083	.017	.037
DinD 3rd Post-test	-.033	.078	.057	.088	.040	.032
Wald F	F (3, 12) = 4.83*		F (3, 12) = 4.41*		F (3, 12) = .955	
N	147		147		147	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

m denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed).

A secondary set of analyses was conducted to verify whether controlling for the global program quality would successfully neutralize the DinD effect on Expressive Vocabulary observed in the main impact analyses by treatment group. Results revealed a weaker treatment effect that remained statistically significant at the .05 level, [adjusted $DinD_{1stPost\text{-}test}$ = -.437, SE = .177]. Nevertheless, this represents a 36% reduction in the magnitude of the original effect. As with the Self-Awareness scale, we conclude that quality partially accounts for a third of the observed treatment effect, but does not eliminate it completely.

Communication, Cognition, & Expressive Vocabulary (Standardized) with Quality of Reading: Quality of Reading was according a special place in the analyses because of its importance to the tested program both in theory and in practice (see Table 5.7 and the *Project Implementation Report*, Bérubé et al., 2014). We report in Table 5.20 estimates of its effect on the ÉPE–AD outcomes it was most likely to affect: the Communication, Cognitive Ability and Expressive Vocabulary scales.

Quality of Reading was not found to influence the Communication scores of children who attend daycare, but their Cognitive Ability and Expressive Vocabulary were significantly impacted. As in previous analyses, the effect is specific to the first post-test. The two affected outcomes are comprised of items that tap early literacy skills and, of course, Expressive Vocabulary.

Table 5.20: Communication, Cognitive Ability, & Expressive Vocabulary Scores as a Function of Quality of Reading (Time x Quality) – French Test Takers Only

	ÉPE–AD Outcomes					
	Communication		**Cognitive Ability**		**Expressive Vocabulary**	
	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Effect at Baseline	-.028	.020	-.034	.029	-.062	.029
DinD 1st Post-test	.041	.030	.057*	.026	.113†	.026
DinD 2nd Post-test	.006	.035	.015	.039	.011	.040
DinD 3rd Post-test	.000	.031	.023	.040	.005	.031
Wald F	F (3, 12) = .55		F (3, 12) = 4.21*		F (3, 12) = 7.29†	
N	146		146		146	

Note: The outcome has been standardized, which means that the effects can be interpreted in terms of standard deviation units.

m denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed).

As in the preceding analyses, we re-estimated differences between the daycare groups using quality as a covariate. The adjusted program effects for the first post-test fell short of statistical significance for both the Cognitive Ability scale [adjusted $DinD_{1stPost\text{-}test}$ = -.176, SE = .160] and the Expressive Vocabulary subscale [adjusted $DinD_{1stPost\text{-}test}$ = -.397, SE = .432]. The tested program's affect on these outcomes is therefore mediated by the quality of literacy activities it offers.

Summary

Several measures of daycare program fidelity and quality were taken and found to be weakly to moderately correlated with each other. A first series of analyses indicated the program daycares differentiated themselves from those of the comparison-group daycares on most measured dimensions. The superior quality of reading activities offered at program daycares was perhaps the most notable aspect to the results. Taken together, the results indicate that if the children attended these two types of daycare experience measurably different outcomes, then this finding may be attributable to the implementation of the tested program instead of some other factor. This supposition was confirmed in a second series of analyses where (a) the key results from the treatment group comparisons were replicated and (b) subsequently eliminated or substantially reduced by statistical control of program fidelity and program quality. Where such confirmation was obtained, we are confident that the observed results are due to the measured dimensions of the tested program and not some other factor. Interestingly, certain dimensions of quality and fidelity were found to be more predictive of certain child outcomes than others. This information may serve to inform changes to the program and areas to emphasize when considering where to focus investment of resources.

Fine-grained analyses: Linguistic Characteristics of the Sample

As with any intervention, it is interesting to consider whether the effectiveness of the treatment depends on some characteristics of the child. Particularly germane to the Readiness to Learn project is the assumption that increased exposure to a high quality daycare program

delivered in French will positively impact the developmental outcomes of children particularly in terms of linguistic skills. It is reasonable to hypothesize that the tested program will have its largest effect on children for whom the program has created the greatest change in exposure to French/quality of exposure to French when compared with the baseline period. We presume that children with a high degree of exposure to English at baseline fit this description. In short, we expect those children who were least exposed to French at baseline to respond most strongly to the tested program (other program evaluations have obtained similar results with language outcomes, Maltais, 2007).

The linguistic characteristics of the child can be captured in various aspects using indicators of the linguistic characteristics of families. This list includes the continuum of French spoken by the child with those in his or her immediate environment, the language of literacy activities, the linguistic composition of the household (i.e., Endogamous Francophone vs. Other), the continuum of French spoken by the mother to the child,[78] and the continuum of French spoken by the father to the child. We report the consistently high correlation among all the linguistic variables in Table 5.21.

The household composition (henceforth "Household Type") is a variable that makes a distinction between homogeneous households where both parents are francophone (i.e., Endogamous Francophone), and those families where parents are more linguistically diverse (i.e., other). Two "household"-type variables were considered in the analyses to follow. The first was "household" based on the FOLS (First Official Language Spoken), which is defined by the official languages known by the parents, their mother tongue, and the language they use in the home. The second was an alternative version of 'household' based exclusively on the language with which the parents reported speaking to the child (Child-Relative). The advantage of the second definition is that the focus is exclusively on conversations that involve the child. These variables were reduced to dichotomies (i.e., Endogamous Francophone versus Other) in order to simplify the analyses and accommodate the small cell sizes for some categories (e.g., Endogamous Anglophone families). For both binary variables, Endogamous Francophone households were coded as 1 and all other family types were coded as 0.

Even though the observed inter-correlations were strong, the measured linguistic variables can be ranked meaningfully along an important dimension: the child's active involvement. As a construct, exposure to French can be conceived as a continuum ranging from completely passive (i.e., a conversation within earshot that does not involve the child as a participant) to completely active (i.e., a conversation that involves a child as active participants, where French is the language that is both understood and produced). We anticipate that linguistic variables that most reflect the active participation of the child will interact most strongly with the program effect. This prediction was founded on the fact that active (versus passive) exposure to linguistic materials is recognized as being the most effective method of transferring such knowledge/skills (Whitehurst, et al., 1988; Hargrave & Sénéchal, 2000). Children who use French more consistently invite further exposure to the language from other people in what has been described as a chain reaction (Pearson, 2007). This conceptualization of "active" involvement suggests that

[78] An alternative definition would be "Language Continuum Spoken to the child by the Francophone Parent." The two definitions of this variable are correlated in excess of .90. In fact, the change in definition would affect the data of only seven families in the present sample. We opted for the simpler definition and the definition that is consistent with the literature indicating the characteristics of the mother are the strongest predictors of child outcomes.

the best candidate for moderating the program effect is the Language continuum spoken by the child.

The distinction between passive and active exposure is relevant to evaluating the impact of the tested program because the active participation of children is an important component of both the tested preschool daycare program and the family literacy workshops. Indeed, according to the *Project Implementation Report* (Bérubé et al., 2014), it was in the quality of reading activities in the classrooms (Quality of Reading) where the largest differences were observed between program daycares and those of the comparison group. The highest scores are obtained on this dimension when the children are engaged to participate actively in French during classroom activities. If active participation using French is a key element of the tested program, then children who were most deficient in this area prior to receiving the intervention should experience the largest effect. The strength of this relationship should be stronger for variables that best capture the "active" component of exposure and weakest for those linguistic variables that are based on "passive" criteria, such as the original household variable based on the FOLS.

Finally, we anticipate that language of literacy activities in the home will be particularly useful in distinguishing the Program Daycare group from the comparison groups. This expectation is justified by the fact that the family literacy component of the intervention directly targeted this dimension and literacy activities are an important contributor to school preparedness (Neuman & Celano, 2001)[79]. This measure is also the only one tied to specific concrete behaviours, which is a property that is known to elicit more accurate responses (for discussion, see Shrigley, 1990). Finally, other interventions have found it to be a moderator of the effectiveness of similar programs (Maltais, 2007).

[79] The analyses reported in Section 5.2 do not reveal a general program effect for the language of literacy activities in the home outcome measure. Nevertheless, it is possible that the instrument used was not sensitive enough to detect the effect. If this is the case, then indirect effects on the outcomes of children are still possible.

Table 5.21: Zero-order Correlations among the Linguistic Profile Indicators

Variable Name	*1*	*2*	*3*	*4*	*5*	*6*
1 - Language Continuum (child)	1	.63	.56	.71	.66	.69
2 - Language of Litt. Activities	.86	1	.49	.65	.59	.64
3 - Household Type (FOLS)	.63	.57	1	.72	.62	.59
4 - Household Type (Child-Relative)	.80	.74	.72	1	.82	.80
5 - Language Continuum (Mother to Child)	.78	.72	.63	.81	1	.64
6 - Language Continuum (Father to Child)	.83	.82	.59	.76	.61	1

Note: Based on baseline survey, N = 232 to 243. Pearson r correlations are reported in the lower diagonal. Kendall's tau rank-based correlations are reported in the upper-diagonal. All correlations are statistically significant at the .05 alpha level.

FOLS = First Official Language Spoken.

Conditional (On Language Exposure) Impact of Program on Standardized ÉPE–AD Scales

We anticipate that the impact of the program will be largest with children whose baseline behaviour and environment resulted in exposure to the French language was weakest. The strongest test of this idea is provided by an outcome variable that represents the widest range of French-language proficiency. For the Readiness to Learn project, the ideal candidate is the Communication scale because for this scale valid observations are available for the entire sample, including children with the weakest ability to communicate in French (i.e., all children were evaluated on this domain regardless of their linguistic background or level of proficiency in French). The other ÉPE–AD outcomes may not show the importance of baseline language exposure as clearly. This anticipated lack of power is a consequence of the exclusion of those children who were least proficient in French by reason of their having completed the ÉPE–AD in English. Such children are more likely to be represented in the lower-end of the French exposure indicators reported in Table 5.21.

We investigated this hypothesis by computing three-way interactions among Treatment Group, Time, and the linguistic characteristic being examined, which were then added to the model specifications[80]. The estimates yielded by these regression analyses were used as a starting point for generating conditional estimates of the program effect for "high French exposure" children and "low French exposure" children; they define an equation that can be solved for different levels of language exposure. The conditional estimates of the program effect reported in this report were computed in this way.

[80] To properly test the three-way interaction, we also needed to add so-called "lower order" interaction effects. Accordingly, all possible two-way interactions involving the linguistic characteristic being analyzed, Treatment Group, and Time were inserted into the regression model specification.

It is important to note that the moderating role of each linguistic characteristic was verified in separate regression models. In other words, the pattern of results is free to vary for each linguistic characteristic, but we must keep in mind that the variables are highly correlated. Consistent results across all indicators would suggest that each indicator is tapping a common construct (e.g., exposure to French) and that this common variance is driving the results. Conversely, results that diverge markedly depending on the language variable that is considered would suggest that the variables are tapping a unique aspect of the construct (e.g., "active" as opposed to "passive" language exposure). We have already mentioned that the obtained results might depend on how "active" the exposure to French is captured by a given variable.

In what follows, we report estimates of the program effect conditional on Exposure to French for each ÉPE–AD outcome. The DinD estimates of the program effect are interpreted in exactly the same way as they were in previous analyses. The only difference is that in this case the effects are conditional on a particular characteristic of the individual. Such conditional effects are easiest to interpret when the test of the three-way interaction among time, treatment group, and linguistic characteristic is statistically significant. Such a significant interaction can be interpreted as a test of whether or not the magnitude of the program effect varies significantly as a function of linguistic characteristic.

Two types of three-way interaction tests are reported. The first type tests whether the magnitude of the program effects (i.e., DinD estimates) vary with changing levels of language exposure with all post-test periods taken together. This general hypothesis is tested using the Wald-F test of the global interaction, which is reported in the second-to-last row of the tables. The second type of three-way interaction evaluates whether individual DinD estimates vary according to language exposure. The results of this second type of test, which amounts to a simple interaction test or DinDinD test (see the analyses by daycare dosage reported in Table 5.4 as an example), is reported in the tables by way of the symbol Δ when the test is statistically significant ($p < .10$).

Communication Scale (Standardized): The hypothesis that the magnitude of the program effect varies significantly as a function of the linguistic characteristics of the sample was verified formally by way of a 3-way interaction. The interaction was significant at $p = .10$ for 5 of the 6 linguistic characteristics considered here according to the Wald-F test. The conditional estimates of the program effect are reported in Table 5.22. Significant "simple" interaction tests are indicated by the symbol Δ in the table.

The results that were obtained with each moderator converge to form a relatively clear picture of the relationship between linguistic characteristics and the magnitude of the program effect. Overall, program effect estimates were much weaker for the children who are highly exposed to French than those with less exposure. This result is consistent with our expectations. None of the tests associated with high-exposure children was statistically significant. We now take a closer look at the estimated treatment effects for the low-exposure condition.

For the low-exposure children, we observe that the program effect based on the daycare group comparison is similar to that reported in Section 5.1.2. Across all moderators, the DinD estimate for the first post-test is statistically significant. A program effect emerges for the third post-test when the Language Continuum Spoken by the Child and the Language of Literacy Activities are considered. This result is interesting given that both behaviours were directly targeted by the tested program and imply the active participation of the child.

Perhaps the most critical question is that answered by the DinDinD tests ("simple" three-way interaction tests, which, when statistically significant, are denoted in Table 5.22 by way of the symbol $^{\Delta}$): does the magnitude of particular DinD estimates vary significantly as a function of the linguistic characteristics of the child? The exact pattern varied according to the analyses, but in general the significant program effects associated with the Informal Care group for the second and third post-tests were significantly larger than their non-significant counterparts in the "high French exposure" condition. A single effect involving the comparison of the Program Daycare group and the Comparison Daycare group interacted with linguistic profile: the third post-test DinD effect involving the language continuum spoken by the child. All the other effects reported in Table 5.22 can be interpreted as positive evidence only insofar as they are in the expected direction.

Unlike the results reported in Section 5.1.2, we observe program effects on the Communication scale relative to the Informal Care group in these analyses. The magnitude of the effects is larger and more consistent over time in these analyses than previously. Across all indicators, the program effects are significant for the first post-test and the third post-test. What is more, we observe significant program effects for the second post-test for the first time in these analyses when the following moderators were used: Language Continuum Spoken by the Child, Household Type (Child-Relative), and Language Continuum Spoken by the Mother to the child. Taken together, the results tend to confirm the expectation that the program makes the most difference with those children who were least exposed to French at baseline, and that this phenomenon is best captured by linguistic variables that tap "active" rather than "passive" exposure.

Table 5.22: Estimated Program Effects for the Standardized Communication Scores (French) Conditional on Linguistic Profile — Full Sample

	Moderators of the Program Effect											
	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother- child)***		***Language Continuum (father-child)***	
High French Exposure	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1 vs. G2												
Baseline	.131	.129	.150	.169	.162	.162	.087	.149	.159	.153	.249	.114
DinD 1st Post-test	-.292	.223	-.216	.190	-.313	.230	-.336	.215	-.299	.188	-.315	.190
DinD 2nd Post-test	-.189	.163	-.135	.177	-.190	.177	-.192	.149	-.143	.208	-.158	.117
DinD 3rd Post-test	$-.017^{\Delta}$	.182	.012	.222	-.090	.199	-.056	.176	-.081	.181	-.030	.157
G1 vs. G3												
Baseline	-.077	.073	-.031	.089	-.100	.099	-.124	.106	-.088	.090	-.019	.060
DinD 1st Post-test	-.131	.195	$-.101^{\Delta}$	.148	-.126	.210	$-.127^{\Delta}$	.173	$-.136^{\Delta}$	.164	-.209	.164
DinD 2nd Post-test	.058	.144	$.062^{\Delta}$	.154	.039	.163	$.106^{\Delta}$	.126	$.119^{\Delta}$	.197	-.039	.144
DinD 3rd Post-test	$.088^{\Delta}$	.150	$.103^{\Delta}$	.197	$.115^{\Delta}$	.165	$.123^{\Delta}$	.131	$.085^{\Delta}$	.174	$-.012^{\Delta}$	.167

	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother- child)***		***Language Continuum (father-child)***	
Low French Exposure	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1 vs. G2												
Baseline	.266*	.119	.343*	.113	.416*	.169	.420*	.152	.348**	.119	.247	.158
DinD 1st Post-test	-.331**	.100	-.398**	.115	-.316*	.097	-.290*	.109	-.319*	.118	-.356*	.141
DinD 2nd Post-test	-.218	.187	-.251	.185	-.153	.220	-.161	.228	-.214	.160	-.168	.245
DinD 3rd Post-test	-.319*$^{\Delta}$	.112	-.295*	.140	-.166	.190	-.204	.174	-.173	.172	-.178	.151
G1 vs. G3												
Baseline	-.338†	.075	-.285*	.113	-.072	.145	-.046	.108	-.224*	.096	-.299*	.108
DinD 1st Post-test	-.405†	.095	-.394**$^{\Delta}$	.127	-.388*	.101	-.440*$^{\Delta}$	.098	-.389*$^{\Delta}$	.111	-.324*	.130
DinD 2nd Post-test	-.325^{m}	.170	-.298$^{\Delta}$	.186	-.212	.181	-.455$^{m\Delta}$	.220	-.442*$^{\Delta}$	.125	-.226	.226
DinD 3rd Post-test	-.499†$^{\Delta}$	.110	-.472**$^{\Delta}$	.119	-.443*$^{\Delta}$	.069	-.625*$^{\Delta}$	.131	-.527*$^{\Delta}$	.113	-.463*$^{\Delta}$	.133
3-way Interaction (Wald F)	F (5, 15) = 1.79		F (5, 15) = 3.88*		F (5, 15) = 2.38^{m}		F (5, 15) = 5.25**		F (5, 15) = 3.46*		F (5, 15) = 3.57*	
N	230		230		230		230		230		223	

Note: Heterogeneity robust standard errors were clustered on daycare. The household variables are dichotomous, therefore the estimates represent the contrast of Endogamous Francophone versus Other family types. For the remaining four variables, the table reports regression-model estimates of the program effects at arbitrarily selected levels of French language use. The upper-bound estimate was calculated for the maximum scale score, while the lower-bound estimate was estimated for the value falling one standard deviation below the mean. The Program Daycare is in reference therefore negative effects denote an advantage for the Program Daycare group. Continuous variables were mean-centered prior to analyses. The baseline evaluation and Program Daycare group are the reference categories for all DinD and DinDinD effects.

mp < .10, *p < .05, **p < .01,† p < .001 ; Δ denotes effects that vary significantly as a function of varying degrees of exposure to French p < .10.

Self-Awareness Scale (Standardized): The analyses of the standardized Self-Awareness scores revealed statistically significant interactions for only one of the six linguistic characteristics according to the global Wald F test. The lone significant moderator of the program effect was the Language Continuum Spoken by the Father, three-way interaction test: [Wald F (5, 16) = 3.68*]. None of the DinDinD effect tests of the simple interactions was statistically significant, which means that none of the individual DinD program-effect estimates varied significantly when baseline levels of exposure to French are manipulated. The full set of analyses is therefore not reported since it would be redundant with the results presented in Section 5.1.2 (available upon request).

Cognitive Ability Scale (Standardized): The standardized Cognition scores were analyzed in the same manner as above. The analyses revealed statistically significant three-way interactions for five of the six moderators according to the global Wald-F tests. However, the pattern of these interactions is reversed when compared with that reported for the Communication scale (Table 5.21). In this case, the high French exposure children in the Program Daycare group showed superior gains relative to their counterparts in the comparison groups for the third post-test specifically (as evidenced by the result of the DinDinD tests, denoted by the symbol Δ). Regardless of whether children entered the program with a high- or low-level of exposure, the initial positive program effect associated with the first post-test was statistically significant[81]. All the conditional (on language moderator) program effect estimates are reported in Table 5.23.

The simple three-way interaction tests yield consistent findings across all language moderators for the third post-test effects involving the daycare groups (as denoted by the symbol Δ). Parallel effects were observed relative to the Informal Care group for the two household moderators only. This is the period when most of the children in the sample first enrolled in school (52%), which makes these effects particularly interesting. Taken together, the results indicate that, by the time most children entered school, those with the *most exposure* to French at baseline benefit most from the program specifically in terms of their cognitive development.

[81] Note that in Table 5.23 some of the conditional estimates indicate a significantly negative program effect for the low-exposure children (i.e., the DinD coefficients are positive and significant). It is specifically those associated with the Household variables. These estimates should not be interpreted too strongly given that they are not corroborated by the other (continuous) moderators.

Table 5.23: Estimated Program Effects for Standardized Cognition Scores (French) Conditional on Linguistic Profile

	Moderators of the Program Effect											
	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother-child)***		***Language Continuum (father-child)***	
High French Exposure	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1 vs. G2												
Baseline	.246	.147	.277	.153	.113	.166	.139	.131	.265	.163	.301	.153
DinD 1st Post-test	-.363**	.126	-.366**	.155	-.210	.146	-.353*	.138	-.282	.172	-.387**	.132
DinD 2nd Post-test	-.230	.177	-.200	.169	-.187	.147	-.230	.173	-.133	.213	-.191	.162
DinD 3rd Post-test	-.390*Δ	.145	-.413**Δ	.154	-.336*Δ	.158	-.380*Δ	.147	-.265Δ	.210	-.309 mΔ	.150
G1 vs. G3												
Baseline	.283	.156	.236	.095	.242	.159	.267	.142	.237m	.112	.243	.150
DinD 1st Post-test	-.342**	.118	-.296**	.112	-.301*	.123	-.425**Δ	.117	-.256m	.133	-.337**	.109
DinD 2nd Post-test	-.260	.147	-.180	.124	-.314*Δ	.133	-.327*Δ	.145	-.155	.172	-.275 m	.132
DinD 3rd Post-test	-.451†	.093	-.404†	.097	-.451**Δ	.129	-.527†Δ	.115	-.330m	.170	-.414†	.099

	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother-child)***		***Language Continuum (father-child)***	
Low French Exposure	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1 vs. G2												
Baseline	.204	.160	.189	.169	.332	.213	.319	.207	.217	.212	.140	.181
DinD 1st Post-test	-.376†	.077	-.331**	.085	-.417*	.133	-.253*	.108	-.413*	.112	-.302*	.116
DinD 2nd Post-test	.023	.177	-.039	.155	.078	.200	.106	.171	-.030	.169	-.010	.225
DinD 3rd Post-test	.205$^{\Delta}$	.141	.239$^{\Delta}$	.141	.303$^{*\Delta}$	.141	.307$^{m\Delta}$	.148	.201	.148	.141	.135
G1 vs. G3												
Baseline	-.262	.185	.147	.459	-.024	.265	-.077	.274	-.007	.243	.059	.246
DinD 1st Post-test	-.262**	.084	-.451**	.148	-.011	.174	.214$^{\Delta}$	.152	-.055	.113	-.118	.304
DinD 2nd Post-test	-.157	.167	-.635^{m}	.319	.135$^{\Delta}$	.164	.069$^{\Delta}$	.117	-.320^{m}	.155	-.203	.207
DinD 3rd Post-test	-.319	.318	-.520	.330	.064$^{\Delta}$	.149	.104$^{\Delta}$	.238	-.156	.280	-.633**	.168
3-way Interaction (Wald F)	F (5, 15) = 1.54		F (5, 15) = 3.46*		F (5, 15) = 4.21*		F (5, 15) = 7.22**		F (5, 15) = 2.87*		F (5, 15) = 6.49*	
N	211		211		211		211		211		210	

Note: Heterogeneity robust standard errors were clustered on daycare. The household variables are dichotomous, therefore the estimates represent the contrast of Endogamous Francophone versus Other family types. For the remaining two variables, the table reports regression-model implied estimates of the program effects at arbitrarily selected levels of French language use. The upper-bound estimate was calculated for the maximum scale score, while the lower-bound estimate was estimated for the value falling one standard deviation below the mean. ***The Program Daycare group is in reference therefore negative effects denote an advantage for the Program Daycare group.*** Continuous variables were mean-centered prior to analyses. The baseline evaluation and Program Daycare group are the reference categories for all DinD and DinDinD effects.

$^{m}p < .10$, $^{*}p < .05$, $^{**}p < .01$,† $p < .001$; $^{\Delta}$ denotes effects that vary significantly as a function of linguistic profile.

Physical Ability (Standardized): As in the analyses reported in Section 5.1.2, we did not anticipate that the tested program would have an impact on motor-skill development given that the main target of the program was the development of French-language skills and culture. It is in this sense surprising that statistically significant three-way interactions ($p < .10$) are observed for all moderator variables according to the Wald-F test. The interaction is driven by a program effect relative to the Comparison Daycare group for the second and especially the third post-tests (where the simple interactions are significant across all moderators, as denoted by the Δ symbol in Table 5.24).

This unexpected result was investigated further in a series of follow-up analyses. First, we split the scale up into items measuring gross and fine motor-skill development. These analyses revealed that the gross motor skill items were driving the effect. No element of the tested program maps onto this particular outcome. In subsequent analyses, the regression model was re-estimated separately for each community (unreported). These analyses indicated that Cornwall and Durham samples were driving the effect. Follow-up analyses were conducted to verify whether this combination of factors (high-exposure, located in Cornwall or Durham) is associated with a group of children that varies systematically across the two daycare treatment groups. Various potential confounds (for a list of factors affecting physical development, see Venetsanou & Kambas, 2010) were examined as potential explanations for this effect such as whether or not the child reported a health problem for the time in question, the number of chronic conditions, family revenue, number of siblings, number and frequency of sports activities. None of these factors accounted for the finding, which should be interpreted with caution given how unexpected it is. We prefer to avoid advancing speculative interpretations for the result before verifying whether it is replicated with the second cohort.

Expressive Vocabulary (Standardized): Analyses of the Expressive Vocabulary scores yielded statistically significant 3-way interactions according to the Wald-F test for two of the five moderator variables: Language of Literacy Activities and the Language Continuum Spoken by the Father to the Child. However, none of the simple interaction tests (i.e., the DinDinD tests denoted by the symbol Δ) for particular program-effect DinD estimates was statistically significant. In other words, we cannot claim that the distinction between conditional estimates for high- and low-exposure children in Table 5.25 is statistically reliable. The conditional estimates are reported nonetheless as a descriptive "analyses by sub-sample" due to the fact that this outcome is of special interest.

In comparisons involving the daycare groups, we observe that the program effect for the *first post-test* is larger for the high French exposure children than the low-exposure children, though not significantly so. The association with language exposure here is in the opposite direction of what was predicted. However, the effect was observed for both high and low French exposure conditions, which means that these early gains are invariant across levels of language exposure. Language exposure moderated the program effect in the expected way for the third post-test for the following language variables: the Language Continuum Spoken by the Child, the Language of Literacy Activities, and the Language Continuum Spoken by the Father to the Child.

Relative to the Informal Care group, an effect is observed for the first post-test in the low-exposure to French condition. The corresponding effect is not observed for the high-exposure

condition. Thus, there appears to be a conditional benefit here, but it is not present by the 3rd post-test when most of the children entered school.

Table 5.24: Estimated Program Effects for Standardized Physical Ability Scores (French) Conditional on Linguistic Profile

	Moderators of the Program Effect											
	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother-child)***		***Language Continuum (father-child)***	
High French Exposure	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1 vs. G2												
Baseline	.325*	.165	.360*	.176	.184	.140	.196	.135	.277*	.123	.267	.150
DinD 1st Post-test	.106Δ	.209	.016	.245	.128	.270	.070	.240	.054	.219	.074	.203
DinD 2nd Post-test	-.401*	.169	-.432*	.200	-.359mΔ	.189	-.341mΔ	.195	-.331 m	.177	-.295 m	.159
DinD 3rd Post-test	-.458**Δ	.150	-.518**Δ	.171	-.502**Δ	.187	-.499*Δ	.179	-.398*Δ	.148	-.487*Δ	.175
G1 vs. G3												
Baseline	.059	.155	-.022	.156	-.138	.093	-.100	.114	-.060	.134	-.118	.134
DinD 1st Post-test	.094	.228	.144	.237	.188	.223	.158	.226	.147	.205	.164Δ	.222
DinD 2nd Post-test	-.160	.192	-.119	.207	-.110	.169	-.043	.193	-.014	.182	-.007	.197
DinD 3rd Post-test	.006	.185	.089	.240	.088	.225	.099	.221	.144	.225	.082	.260

	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother-child)***		***Language Continuum (father-child)***	
Low French Exposure	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1 vs. G2												
Baseline	-.071	.174	-.063	.164	.085	.242	.080	.228	-.015	.250	.037	.181
DinD 1st Post-test	-.299$^{\Delta}$	.163	-.129	.166	-.265	.141	-.183	.173	-.220	.161	-.245	.204
DinD 2nd Post-test	-.136	.234	-.111	.187	-.152$^{\Delta}$	.240	-.178	.216	-.169	.206	-.224	.264
DinD 3rd Post-test	.119$^{\Delta}$	.253	.183$^{\Delta}$	.246	.214$^{\Delta}$	.246	.170	.226	.133	.270	.087	.174
G1 vs. G3												
Baseline	-.460*	.217	.151	.514	-.153	.323	-.178	.270	-.205	.198	-.010	.233
DinD 1st Post-test	-.109	.629	-.428	.555	-.033	.469	-.068	.481	-.122	.561	-.392$^{\Delta}$	.399
DinD 2nd Post-test	.162	.393	-.235	.377	.187	.345	.068	.367	-.055	.378	-.034	.167
DinD 3rd Post-test	-.047	.477	-.479	.419	.121	.386	-.094	.329	-.204	.455	-.202	.151
3-way Interaction (Wald F)	F (5, 15) = 2.76^{m}		F (5, 15) = 2.33^{m}		F (5, 15) = 4.88**		F (5, 15) = 7.92**		F (5, 15) = 3.74*		F (5, 15) = 6.27*	
N	211		211		211		211		211		205	

Note: Heterogeneity robust standard errors were clustered on daycare. The household variables are dichotomous, therefore the estimates represent the contrast of Endogamous Francophone versus Other family types. For the remaining two variables, the table reports regression-model estimates of the program effects at arbitrarily selected levels of French language use. The upper-bound estimate was calculated for the maximum scale score, while the lower-bound estimate was estimated for the value falling one standard deviation below the mean. The Program Daycare group is in reference therefore negative effects denote an advantage for the Program Daycare group. Continuous variables were mean-centered prior to analyses. The baseline evaluation and Program Daycare group are the reference categories for all DinD and DinDinD effects.

mp < .10, *p < .05, **p < .01,† p < .001; Δ denotes effects that vary significantly as a function of varying degrees of exposure to French p < .10.

Table 5.25: Estimated Program Effects for Standardized Expressive Vocabulary Scores (French) Conditional on Linguistic Profile

	Moderators of the Program Effect											
	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother-child)***		***Language Continuum (father-child)***	
High French Exposure	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1 vs. G2												
Baseline	.420*	.179	.500*	.224	.468*	.189	.358*	.218	.412*	.199	.423*	.196
DinD 1st Post-test	-.644**	.181	-.623**	.204	-.632**	.158	-.669**	.195	-.644**	.188	-.576**	.180
DinD 2nd Post-test	-.227	.169	-.277	.184	-.311*	.137	-.245	.175	-.271	.201	-.154	.180
DinD 3rd Post-test	-.051	.209	-.030	.226	-.172	.169	-.073	.215	-.126	.196	.014	.209
G1 vs. G3												
Baseline	.081	.110	.146	.157	.059	.095	.005	.133	.042	.132	.060	.115
DinD 1st Post-test	-.145	.178	-.143	.156	-.056	.161	-.116	.168	-.157	.193	-.160	.158
DinD 2nd Post-test	.067	.199	-.037	.215	.098	.153	.067	.179	.063	.229	.052	.218
DinD 3rd Post-test	.129	.133	.140	.158	.193	.097	.167	.112	.227	.128	.169	.145

	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother-child)***		***Language Continuum (father-child)***	
Low French Exposure	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1 vs. G2												
Baseline	.236	.133	.262	.166	.333	.182	.333	.182	.354	.164	.311	.163
DinD 1st Post-test	-.384*	.110	-.432*	.081	-.454*	.157	-.454*	.157	-.370*	.086	-.556*	.147
DinD 2nd Post-test	-.272	.292	-.211	.287	-.106	.268	-.106	.268	-.135	.263	-.336	.317
DinD 3rd Post-test	-.360*	.137	-.355*	.164	-.116	.217	-.116	.217	-.153	.166	-.335m	.179
G1 vs. G3												
Baseline	-.279	.378	-.320	.528	.188	.288	.188	.288	.161	.229	-.129	.378
DinD 1st Post-test	-.464*	.400	-.175	.338	-.443*	.160	-.443*	.160	-.317	.272	-.353m	.193
DinD 2nd Post-test	-.198	.405	.216	.344	-.176	.295	-.176	.295	-.204	.320	-.125	.438
DinD 3rd Post-test	-.159	.419	-.245	.330	-.184	.203	-.184	.203	-.337	.426	.141	.150
3-way Interaction (Wald F)	F (5, 15) = 1.32		F (5, 15) = 4.51**		F (5, 15) = 2.01		F (5, 15) = 1.28		F (5, 15) = 1.55		F (5, 15) = 3.68**	
N	211		211		211		211		211		211	

Note: The household variables are dichotomous, therefore the estimates represent the contrast of Endogamous Francophone versus Other family types. For the remaining two variables, the table reports regression-model estimates of the program effects at arbitrarily selected levels of French language use. The upper-bound estimate was calculated for the maximum scale score, while the lower-bound estimate was estimated for the value falling one standard deviation below the mean. ***The Program Daycare group is in reference therefore negative effects denote an advantage for the Program Daycare group.*** Continuous variables were mean-centered prior to analyses. The baseline evaluation and Program Daycare group are the reference categories for all DinD and DinDinD effects.

$^{m}p < .10$, $^{*}p < .05$, $^{**}p < .01$,† $p < .001$; None of the reported effects varied significantly as a function of linguistic profile.

Summary

The results reported in this section confirmed that "active exposure" linguistic variables were most useful in explaining heterogeneous treatment effects. The key moderating effects were almost invariably obtained with such variables (e.g., Language Continuum Spoken by the Child). The exact pattern of results and the linguistic characteristics emerged as relevant depended on the outcome and evaluation period considered.

As expected, children with the most exposure to French at the start of the project were the least influenced in the development of their language skills (i.e., the Communication and Expressive Vocabulary scales), while the gains associated with the program were positive and statistically significant on these outcomes when exposure to French was low (for some linguistic characteristics). The pattern was partially reversed in the case of the Cognitive Ability scale where it was the high-exposure children who showed the most dramatic positive program effects, which were significant both the first and the third post-test, just as many children began enrolment in junior kindergarten. The interpretation of this apparent dissociation is taken up in the chapter summary and in Chapter 7 (General Discussion).

5.2. PARENT-LEVEL IMPACTS

The program evaluated here was designed to operate at two levels: children and parents. Estimates of the direct impact of the program on the developmental outcomes of children were reported in the previous section. We now turn our attention to evaluating whether the family literacy workshops had a significant impact on the self-reported knowledge, beliefs, attitudes, and behaviours of the parents.

Analyses are presented in two sections. The first section reports results related to the literacy activity behaviours of the parents and the language in which there are done. The focus of the second section is on the analyses of additional scales designed to assess the self-reported Knowledge, Efficacy, Modeling, and Beliefs of the parents (for elaboration, see Chapter 3 and the more detailed descriptions provided below). As in the analyses of children outcomes, we evaluate the role of dosage (i.e., parental attendance) and quality (i.e., the workshop quality indices) in moderating the impact of the workshops.

5.2.1. Frequency and Language of Literacy Activities

An important component of the literacy activity workshops was their emphasis on the importance of the frequency literacy activities in the home as well as the importance of conducting these activities in French so as to support the linguistic development of francophone children. In an initial series of analyses, we report the estimated program impacts for the self-reported frequency and language of literacy activity outcomes. These analyses are based on all time periods inclusively, and only included the four communities that have the three experimental groups (Cornwall, Orléans, Durham and Edmundston). The reader is reminded that family literacy workshops were offered only in this first year of the tested program.

Before turning our attention to results, we first link workshop delivery with evaluation periods to assist in the interpretation of the results. The outcomes were gleaned from

eight surveys: the baseline survey and seven follow-up surveys (the first follow-up survey of October 2007 was only administered to parents registered in the project before September 1, 2007). Data from the first two surveys were combined into a single baseline, which served as the pre-test measure (equivalent in time period to the first evaluation of children). The remaining follow-up surveys, starting with the third, map onto the second through seventh child evaluations periods, which is to say the first to sixth post-tests. The terms baseline and post-test therefore refer to surveys and not to evaluations in the analyses reported in this subsection.

Four family literacy workshops in total were delivered between November 2007 and early December 2008, corresponding to what we consider the first post-test evaluation period (or here to the second follow-up survey done in February 2008). The remaining six family literacy workshops were delivered between the end of January 2008 to March 2008, which corresponds to the first and second post-test evaluation period (or here to the second and third follow-up surveys done in February and June 2008). The first post-workshop period corresponds to the third evaluation period. The relationship between the evaluations and the surveys is described in Section 3.5.1.

Model Specification Details

The regression model specification that was employed in estimating the effect of frequency is almost identical to that described in Section 6.1.2. The differences between the two specifications are now described.

The present analyses distinguish themselves in three ways. First, all evaluation periods are included in the analyses[82]. As in previous analyses, each evaluation period was represented by a dummy variable placing the baseline evaluation in reference. Program effects were evaluated for the first post-test (Winter 2008), up until the sixth post-test (Fall 2009). These post-test periods correspond to the third and seventh evaluations. These dummy variables were crossed with dummy variables representing treatment group membership to obtain the variables that were used to produce the DinD estimate. Second, dummy variables representing school enrolment were included in the baseline model (as in Section 7.1.2). Third, the standard errors for these analyses were clustered by child instead of daycare (see Section 4.3). This decision resulted in a substantial increase in the degrees of freedom for the analyses. These additional degrees of freedom allowed the overall interaction between time and treatment group to be tested with a high degree of power despite the loss of degrees of freedom that resulted from including all measurements in the same analyses.

Frequency of Literacy Activities

The results of the DinD panel analyses of the Frequency of Literacy Activities are reported in Table 5.26. First, note that the Program Daycare group is superior to the Comparison Daycare group at baseline. Note also that the magnitude of the effect is small (about .30), which translates in real terms to 1 item point (individual items were scored on a scale from 1 to 5). The practical implication, if any, associated with this difference is negligible. Findings underline that none of the DinD estimates for the Program Daycare group relative to the Comparison Daycare group.

[82] Standard Errors were not clustered at the level of daycare in these analyses, which dramatically increased the degrees of freedom for the analyses, effectively removing the obstacle that motivated the split between Year 1 and Year 2 in the analyses of children outcomes.

Specifically, the DinD estimates are very small for the most part, which indicates parallel change for the two groups on this outcome. In contrast, most of the DinD estimates are significant for the Program Daycare group versus the Informal Care group. The pattern of results suggests that families whose children are enrolled in a francophone daycare show more important gains relative to baseline than those whose children were assigned to the Informal Care group. We cannot attribute this effect specifically to the impact of the family literacy workshops or more generally to participation in the evaluated program, because the two daycare groups were found to be statistically equivalent.

While non-significant, interpretation of these findings would benefit from being nuanced. First, the internal consistency of the scale was relatively low (Cronbach alpha > .60), which indicates that measurement error could be attenuating the impact estimates. More fundamentally, questions for this scale were anchored to very specific frequencies (1-Never, 2 – Once a week or less, 3- A few times a week, 4- Once or twice a day, 5- Three or more times a day). The theoretical maximum score for the scale was 25, which means that to achieve even a score of 20, parents would have to report a 4 on all five items of the scale, for example. In short, the issue is whether there are enough hours in a week for parents to invest so much time in all of these activities. The upper end of the scale may represent a range of frequencies that is unrealistic for the behaviour being measured, creating an empirical ceiling effect. The observed treatment group means based on the unstandardized version of the scale ranged from 14.39 (Comparison Daycare group at Baseline) to 17.35 (Program Daycare group for the sixth post-test). If the scale had been more internally consistent or if the response options had allowed for a finer discrimination at the low- to mid-range of the frequency spectrum, then size of the impact estimates may have been larger.

Table 5.26: Difference in Difference (DinD) Group Effects for Reported Frequency of Literacy Activities (Standardized) in the Home

	Incremental Inclusion of Covariates — Errors Clustered by Child									
	(1)		**(2)**		**(3)**		**(4)**		**(5)**	
Program Effects	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Group Difference at Baseline (G1 vs. G2) [a]	-.380*	.151	-.376†	.140	-.409†	.141	-.379†	.141	-.293*	.135
DinD 1st Post-test Mid-Workshop	-.018	.155	-.026	.154	-.025	.155	-.024	.155	-.040	.157
DinD 2nd Post-test 1st Post-Workshop	-.056	.158	-.074	.156	-.084	.158	-.084	.158	-.098	.160
DinD 3rd Post-test 2nd Post-Workshop	-.209	.167	-.241	.161	-.251	.162	-.254	.161	-.270	.164
DinD 4th Post-test 3rd Post-Workshop	.026	.170	-.015	.154	-.010	.154	-.011	.155	-.027	.153
DinD 5th Post-test 4th Post-Workshop	-.052	.172	-.089	.161	-.089	.161	-.103	.160	-.099	.159
DinD 6th Post-test 5th Post-Workshop	-.012	.193	-.066	.182	-.063	.182	-.068	.180	-.081	.178
Group Difference at Baseline (G1 vs. G3) [a]	.009	.160	.033	.165	-.023	.165	.051	.171	.018	.161
DinD 1st Post-test Mid-Workshop	-.340 [m]	.177	-.408*	.179	-.411*	.180	-.418*	.180	-.384*	.182
DinD 2nd Post-test 1st Post-Workshop	-.244	.169	-.297 [m]	.171	-.302 [m]	.172	-.306 [m]	.172	-.300 [m]	.174
DinD 3rd Post-test 2nd Post-Workshop	-.503†	.175	-.544†	.178	-.551†	.180	-.568†	.181	-.561†	.184
DinD 4th Post-test 3rd Post-Workshop	-.247	.188	-.304	.195	-.323 [m]	.195	-.335 [m]	.194	-.295	.189
DinD 5th Post-test 4th Post-Workshop	-.400*	.202	-.439*	.205	-.463*	.204	-.480*	.204	-.414*	.200
DinD 6th Post-test 5th Post-Workshop	-.408 [m]	.208	-.467*	.213	-.495*	.210	-.480*	.210	-.395 [m]	.203
Omnibus Wald F-tests (Time x Group Interaction):	F (12, 231) = 1.09		F (12, 228) = 1.14		F (12, 227) = 1.20		F (12, 227) = 1.22		F (12, 220) = .996	
N participants:	243		240		239		239		232	

Note: The outcome has been standardized, which means that the DD effects can be interpreted in terms of standard deviation units. ***Program Daycare group (G1) is in reference, which means that negative values of DinD denote an advantage for the Program Daycare group***. Standard errors were estimated via the SPSS implementation of the heterogeneity consistent 'robust' White estimator (White, 1980). Further, standard errors were clustered by Child to adjust for correlation in the residuals over time. Model specification details are provided in Section 5.1.2. Note that whether or not the child is enrolled in school is entered as a covariate in the base model.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed).

[a] G1, G2, and G3 denote Program Daycare group, Comparison Daycare group, and Informal Care group respectively.

Language of Literacy Activities

The result of the impact analyses for the Language of Literacy Activities outcome is reported in Table 5.27. These analyses serve to verify whether the family literacy workshops achieved its goal of encouraging parents to provide an environment for their children that fosters their intellectual development within the French language. Here the language used in conducting literacy activities was employed as a proxy for this concept.

The DinD estimates reported in the table indicate that the rate of change for this variable was similar for the two daycare groups. In contrast, a significant difference was observed between the Program Daycare group and the Informal Care group for the first post-workshop period (second post-test). A corresponding effect was not observed for the mid-workshop period (first post-test). The specificity of this result is consistent with the fact that the subject at hand, language of literacy activities, was addressed in a workshop that was presented concurrently to the administration of the first post-test (end of January to early February 2008). It makes sense that the observed effect would not emerge until the following survey when all participants had received the training specific to this outcome. Of course, if this truly were an effect that is specific to the Family Literacy Workshop, we would expect the same effect to be observed when the two daycare groups are being compared. A less ambiguous result would have been a significant effect relative to both the Comparison Daycare group and the Informal Care group. In short, the analyses provided little definitive evidence that families whose children are enrolled in a Program Daycare distinguish themselves from families assigned to the other comparison groups.

Table 5.27: Difference in Difference (DinD) Group Effects for Reported Language of Literacy Activities in the Home (Standardized Scores)

	Incremental Inclusion of Covariates — Errors Clustered by Child									
	(1)		**(2)**		**(3)**		**(4)**		**(5)**	
Program Effects	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Group Difference at Baseline (G1 vs. G2) [a]	.064	.135	.007	.024	.008	.025	.001	.027	.003	.029
DinD 1st Post-test Mid-Workshop	-.029	.095	-.026	.094	-.025	.094	-.025	.094	-.030	.096
DinD 2nd Post-test 1st Post-Workshop	-.089	.097	-.095	.095	-.097	.095	-.097	.095	-.099	.097
DinD 3rd Post-test 2nd Post-Workshop	-.124	.109	-.128	.105	-.118	.105	-.118	.105	-.125	.107
DinD 4th Post-test 3rd Post-Workshop	.209	.135	-.005	.060	-.003	.061	-.005	.061	.002	.063
DinD 5th Post-test 4th Post-Workshop	.171	.138	.014	.055	.021	.054	.017	.055	.023	.056
DinD 6th Post-test 5th Post-Workshop	.077	.142	-.058	.078	-.048	.079	-.060	.079	-.058	.081
Group Difference at Baseline (G1 vs. G3) [a]	.264 [m]	.141	.021	.025	.022	.027	.009	.031	.016	.032
DinD 1st Post-test Mid-Workshop	-.114	.092	-.132	.088	-.129	.088	-.129	.087	-.148	.090
DinD 2nd Post-test 1st Post-Workshop	-.330†	.101	-.303†	.091	-.299†	.091	-.299†	.091	-.313†	.094
DinD 3rd Post-test 2nd Post-Workshop	-.146	.114	-.110	.101	-.106	.101	-.103	.101	-.099	.104
DinD 4th Post-test 3rd Post-Workshop	-.213	.164	-.084	.072	-.076	.072	-.069	.073	-.082	.075
DinD 5th Post-test 4th Post-Workshop	-.313 [m]	.169	-.127 [m]	.069	-.113	.068	-.106	.069	-.112	.072

	Incremental Inclusion of Covariates — Errors Clustered by Child									
	(1)		**(2)**		**(3)**		**(4)**		**(5)**	
Program Effects	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
DinD 6th Post-test 5th Post-Workshop	-.191	.175	.003	.102	.018	.103	.053	.103	.047	.108
Omnibus Wald F-tests (Time x Group Interaction):	F (12, 231) = 2.62†		F (12, 228) = 1.99*		F (12, 227) = 1.87*		F (12, 227) = 1.89*		F (12, 220) = 1.96*	
N participants:	243		240		239		239		232	

Note: The outcome has been standardized, which means that the DinD effects can be interpreted in terms of standard deviation units. ***Program Daycare group (G1) is in reference, which means that negative values of DinD denote an advantage for the Program Daycare group***. Standard errors were estimated via the SPSS implementation of the heterogeneity consistent 'robust' White estimator (White, 1980). Further, standard errors were clustered by Child to adjust for correlation in the residuals over time. Specifications: Details are provided in the body of the text.

[m] denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed).

[a] G1, G2, and G3 denote Program group, Comparison Daycare group, and Informal Care group respectively.

5.2.2. Pre-post effects: Self Reported Knowledge, Self-efficacy, Modeling Behaviours

Other purposes of the workshops were to a) inform the parents concerning issues related to child development, strategies to help better prepare the child for school, and the availability of francophone services in their community (Knowledge), b) increase their sense of Self-efficacy in their role as a parent as first educator of their child (Efficacy), and c) increase the frequency with which they modeled literacy activities for their child (Modeling). Analyses included participants from all six participating communities who were assigned to the Program Daycare group. The results of analyses comparing the participant (Participant) and non-participants (Non-participant) families (family literacy workshops) are reported in Table 5.28, while tests of change are reported in the text. The results for each indicator are discussed in detail in separate sections.

Specification Details

The effects reported in this section are unadjusted by the inclusion of an extensive list of covariates by reason of the prohibitively small number of participants (Green, 1991). There is nonetheless reason to believe that substantive differences exist between the Participant and Non-participant families. Both the parametric and non-parametric rank-based tests are reported as a check against the violation of distributional assumption violations. A distinguishing feature of the non-parametric tests is that they are sensitive to the relative rank of scores but not to their absolute differences[83]. Welch's F test (a t-test that does not assume homogeneity of variance or equal group sizes) and the Mann-Whitney U rank-order statistic (non-parametric) were conducted to evaluate differences between independent groups[84].

Table 5.28: Comparison of Parental Self-report Measures Pre- and Post-workshop and Participant vs. Non-participant

	Point-of-Reference of the Parent's Estimation				
	Pre- (Retrospective)		**Post-**		
Efficacy	Mean	SD	Mean	SD	N
Participant	18.75	3.34	21.20	2.34	89
Non-Participant	See post	See post	20.07	1.64	14
Welch's F (Participant vs. Non-participant)	F (1, 33) = 5.48* , Z = 1.18		F (1, 22) = 5.05*, Z = -2.05*		
Knowledge	Mean	SD	Mean	SD	N
Participant	17.5	3.26	20.81	2.20	90
Non-participant	See post	See post	17.73	3.31	15
Welch's F (Participant vs. Non-participant)	F (1, 18) = .06, Z = .32		F (1, 16) = 12.12† , Z = 3.70†		

[83] Large discrepancies in the effect size estimated using these two methods would indicate that the relative differences between scores is relevant to determining the size of the effect. Such a result is informative but does not necessarily indicate a problem with the result of the parametric test. Visual inspection indicated that the distribution of scores within each cell was quite reasonable given the small number of observations.

[84] Independent tests were conducted for each comparison.

	Point-of-Reference of the Parent's Estimation				
	Pre- (Retrospective)		Post-		
Modeling	Mean	SD	Mean	SD	N
Participant	18.73	3.28	18.88	3.03	88
Non-Participant	18.17	3.24	19.13	2.20	15
Welch's F (Participant vs. Non-Participant)	F (1, 14) = .32, Z = .49		F (1, 24) = .16, Z = -.14		

Note: Pre-test values are reported for the non-participants only when they were available (Prospective). Specification details are reported in the body of the text.

Z = Z-value of the Distribution-free Wilcoxon rank sum test.

^{m} denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed).

Self-reported Knowledge

The pattern of results obtained for the Knowledge scale is consistent with the conditions that were outlined for a "true" workshop effect in Section 4.2.4 of the report. The pre-test estimates provided by Participant families are equivalent to the post-test estimates of the Non-participant families, and there is consistent evidence of a positive program effect.

First, a repeated-measures t-test confirmed that those parents who participated in the workshop provided estimates that were significantly higher than the baseline estimates by 3.31 points on average, which is more than half the range of an item, [t (89) = 13.26, SE = .25, $p < .001$; Wilcoxon signed rank = 3,160, Z = 7.75, $p < .001$]. None of the participants reported a post-test estimate lower than the pre-test estimate and only 11 reported no change. Second, in Table 5.28, we observe an equivalency between the Participant and Non-participant families for the pre-test estimates and a statistically significant advantage for the participants when their post-test estimates are used as the basis for comparison. The post-test effect is in the order of 3 points, which corresponds well with the treatment effect estimate based on the change score. Taken together, these results are consistent with the idea that the two groups were similar to each other prior to the workshops and that those who participated in the workshop received a statistically significant boost. No positive evidence of bias in the estimates was observed.

Self-reported Self-efficacy

Self-efficacy scale results were only partially consistent with the conditions outlined previously as being most reflective of a real impact for the family literacy workshops. First, a repeated measures t-test indicated that the change relative to the baseline estimates was 2.45 points and statistically significant, [t (88) = 11.62, $p < .001$; Wilcoxon signed rank = 2,775, Z = 7.52, $p < .001$]. None of the participants reported a decrease in their sense of Self-efficacy and only 15 participants provided equivalent estimates for the baseline and post-test period.

Inspection of Table 5.28 indicates that the workshop participants reported lower Self-efficacy prior to the workshops than the non-participants. This difference was significant using the parametric test and non-significant by the non-parametric test. This initial difference may be the result of an underestimated pre-test state by participants (Taylor, Russ-Eft, & Taylor, 2009) or it may reflect an over-estimate on the part of non-participants (Moore & Tananis, 2009). The difference between groups could also be real and due to a self-selection process whereby parents who feel more efficacious in their role as parents are less likely to feel that the workshops are

worth their time. In any case, the results indicate that by the post-test period workshop participants reported higher Self-efficacy than the non-participants. This effect cannot be explained by the types of biases that are normally associated with an estimate made prior to/without intervention (Moore & Tananis, 2009) or with those made post-intervention (Taylor, et al., 2009). The size of the effect is in the order of a single point, a fraction of the maximum item score. Taken together, the results are consistent with the idea that the workshops had a positive impact on the participant's sense of Self-efficacy, but a) the size of the estimated effect is not as strong as with self-reported knowledge nor b) do the results concord as well with the ideal case where Participant and Non-participant families were statistically equivalent based on the pre-test estimates.

Self-Reported Modeling Behaviours

Analyses of the self-reported Modeling scale indicated that the participants and non-participants were similar both on the baseline estimates and the post-test estimates. The effect of the workshops on this outcome was negligible, even based on the highly sensitive repeated-measures analyses, [effect = .01, t (86) = .03, SE = .34]. A total of 37 participants reported positive change, 39 participants reported negative change, and 11 reported no change from baseline. No evidence of a program effect was detected for this measure, but this may be due to the anchor points that were selected for the response options. As with the Frequency of Literacy Activities scale, the modeling questions were anchored to very specific frequencies (1- Never, 2- Once a week or less, 3- A few times a week, 4- Once or twice a day, 5- Three or more times a day), which means that response-shift bias is not a problem (Howard, 1980). The scale includes 5 items, which means that the theoretical maximum for this scale is 25. As it stands, the reported means approach 20 (see Table 5.28) for both Participant and Non-participants on the pre-test, which is excellent. To achieve a score greater than 20 would require a frequency in excess of once or twice a day for all activities. It is unclear to what extent it is reasonable for parents to engage in each literacy-modeling activity three or more times a day. Had the response options provided to parents been better calibrated (i.e., particularly in mid-range), it is possible that some change would have been observed on this measure. Finally, the data for the prospective pre-test were collected after the workshop began for some participants, which may have contributed to the attenuation of the workshops effect[85].

Self-Reported Beliefs and Attitudes on Child Development

Four targeted questions regarding parental beliefs and attitudes were posed along with the other measures. The purpose of these questions was to evaluate the degree to which attending the workshops influenced their opinions (4-point scale: Totally Agree, Agree, Disagree, Totally Disagree). We observed that both pre-test (prospective, 2nd follow-up survey) and post-test (post-program survey) questions generated distributions that were highly skewed toward the responses that were considered "ideal" by the workshop program. In interpreting this result, it is important to note that some of the "pre-test" responses were collected after the workshops began (November 2007) and the content that was likely to influence these opinions was presented in the

[85] According to our records, at least 94.28% of parents completed the survey after the workshops began (November 2007). We lack this information for the remaining 6 parents. This problem has been corrected for the second cohort. An unbiased estimate of the program effect will be obtained from this second wave of data.

first few sessions. In other words, the pre-test comparison and the comparison with the non-participants may both be contaminated by exposure to the workshop.

Who is responsible for my child's education? Two questions addressed this issue. The first question was related to whether or not the respondent believed that the early-childhood educators or teachers were primarily responsible for the education of their child. Overall, 74.4% of the sample disagreed with this statement before the workshop (28.9% totally disagreed) and 72.2% after the workshop (34.4% totally disagreed). Of those who reported agreeing with the statement, a minority reported strong agreement (6.7% Pre, 4.4% Post). Given that this theme was addressed in the first workshops delivered in November 2007, it is unsurprising that the compromised 'prospective' pre-test indicated that parents were unlikely to endorse this belief before the workshop began (Mean = 2.03, SD = .87, out of an 4 point scale: totally disagree to totally agree) and that this belief did not change significantly following the workshop (Mean = 1.98, SD = .87), [correlated t (89) = .65, Wilcoxon = 329, Z = -.65]. Of the Participant respondents, 17 showed positive change while 21 showed negative change (52 participants reported no change in their opinion). The Participants were virtually identical to the Non-participants for the post-test comparison (Non-participant Mean = 1.93, SD = .96), [Welch's F (1, 18.07) = .028, Mann-Whitney U = 647.50, Z = -.266].

The second question addressed the issue of whether or not the parent worked together with the educators to help their child develop and be better prepared for school. Only 9% of Participants disagreed with this statement on the pre-test (6.7% totally disagreed) and only 10% disagreed after the workshops (3.3% totally disagreed). Among those who agreed, 45.6% reported total agreement before the workshop and 55.6% after. The apparent shift in responses is corroborated by a stronger trend toward a positive program impact when Participants are compared across time from pre-test (Mean = 3.30, SD = .82) to post-test (Mean = 3.43, SD = .77), t (88) = 1.29, [Wilcoxon = 499, Z = 1.30]; and when they are compared with Non-participants on the post-test, non-participants (Mean = 3.07, SD = 1.16), [Welch's F (1, 16.07) = 1.31; Mann-Whitney U = 582, Z = -.953]. In total, 25 participants show positive change, 15 showed negative change, and 49 showed no change.

Is my child too young to be read to? Another question asked parents whether they believed their child was too young to be read stories. For the pre-test, 92.2% of Participants were in total disagreement and 100% of Non-participants were in total disagreement. For the post-test, these numbers shifted only slightly to 95.6% for the Participants, while the Non-participants remained unchanged. No tests are reported for this measure due to lack of variance, but it is interesting to note that four participants reported that they 'totally agreed' with this statement prior to the workshop, but none did afterward.

Literacy materials are in French? Parents were asked whether books, toys, music and photos in the home reflected francophone culture. Among Participants, 26.4% disagreed prior to the workshop (3.4% totally disagreed) and 27.3% following the workshop (6.8% totally disagreed). Of those who agreed, 31% and 36.4% endorsed 'totally agree' for the pre-test and post-test period respectively. Given how little the frequencies changed following the workshops, it is perhaps not surprising that formal tests of change in opinion for the Participants indicated no significant change, Pre-test Mean = 3.01, SD = .84, Post-test Mean = 3.05, SD = .94; [t (85) = -.035; Wilcoxon = 573, Z = .40]. Similarly, Participants (n = 90) did not differ significantly from Non-participants (n = 15), [Welch's F (1, 16.07) = 1.31; Mann-Whitney U = 582, Z = .953].

5.2.3. Moderators of the Parent-level effects

In this section, we consider the impact of factors that could plausibly enhance the effect of the family literacy workshop. In the analyses to follow, we explore the role of workshop quality and fine-grained estimates of dosage in affecting the magnitude of the difference between the pre-test and post-test estimates provided by those parents who participated in the workshop three or more times (N = 90). These analyses are limited to the outcomes that were the object of the analyses reported in Section 5.2.2[86]. We report estimates of the strength of the association between change scores and estimates of workshop quality in Table 5.29[87].

Specification Details

The effects reported in this section are reported unadjusted by the inclusion of an extensive list of covariates by reason of the prohibitively small number of participants (Green, 1991). There is reason to believe that substantive differences exist between the participants and the non-participants. However, Welch's F-tests based on the covariates described in Section 5.4 failed to reveal any reliable differences between the groups.

Both the parametric and non-parametric rank-based tests of association are reported as a check against the violation of distributional assumption violations. A distinguishing feature of the non-parametric tests is that they are sensitive to the relative rank of scores but not to their absolute differences[88]. Zero-order Pearson r (parametric), partial Pearson r, and Kendall's tau (non-parametric) are reported as measures of association. In all cases, the partial Pearson r is the correlation between a change score and a quality measure after statistically adjusting for the pre-test score.

Because pre-test data were available for all workshop participants, a DinD estimator was used to estimate the effect of a fine-grained indicator of parental participation (i.e., proportion of workshops attended). In this case, both pre-test and post-test estimations are included in the model. Time was represented by way of a dummy variable and was crossed with the parental participation variable to obtain the indicator that was used to estimate the treatment effect. The sole covariates entered in analyses were dummy-variables representing community affiliation.

---

[86] These outcomes, based on the complete sample of the Readiness to Learn project, provided a sample (n = 90) that approached conventional benchmarks for obtaining stable estimates in a regression analyses (approximately N = (104 + k), where k is the number of predictors, Green, 1991). In contrast, the sample size of program participants for the literacy activity analyses was about 2/3 of the size by reason of the exclusion of Edmonton and St-Jean for these analyses.

[87] At least two methods of estimating the relationship between workshop quality and change scores (Maris, 1998). The first would involve simply correlating the change score (post-test minus pre-test) with the variable of interest. The second would enrich the analyses by statistically controlling for the pre-test value when analyzing the post-test scores. The two methods answer slightly different questions. The first tests whether on average the change scores are associated with the other indicators. The second tests whether the relationship would be significant if all participants were equivalent on the pre-test measure. Both methods yield biased estimates under certain conditions related to the presence of measurement error and the causal model relating the pre-test score to selection for treatment (Glymour, Weuve, Berkman, Kawachi, & Robins, 2005). Therefore, we report results based on both strategies. Further, given the small number of communities involved in the analyses (n = 4), we report non-parametric estimates of association to verify whether the distribution of quality scores may be playing a disproportionately important role in determining the value taken by the estimates. We consider that greater credibility should be given to the results when all methods of estimation lead to the same conclusion.

[88] Large discrepancies in the effect size estimated using these two methods would indicate that the relative differences between scores is relevant to determining the size of the effect. Such a result is informative but does not necessarily indicate a problem with the result of the parametric test. Visual inspection indicated that the distribution of scores within each cell was quite reasonable given the small number of observations.

Workshop Quality associated with Change Scores

In Table 5.29, quality effect estimates are highlighted in grey if they emerge consistently across tests as associates of participant change scores. The two that best predicted gains scores were: a) the degree to which the session length was judged to be appropriate and b) the style of the delivery of the family-literacy practitioner (Practitioner Style). The estimate of global quality, over all other variables except those based on parental reactions (positive, negative), confirmed this association. The Material Coverage indicator was associated with change scores when they were not adjusted for pre-test values. The associated relationship is ambiguous because it is confounded with pre-test scores. In other words, the effect of treatment in this case appears to depend on the estimated state of the participants prior to treatment. Another source of discrepancy is the difference between the parametric and non-parametric estimators. When the parametric and non-parametric tests (based on rank order) disagree markedly, it is a signal that the magnitude of the interval between scores on either the change score or the quality measure either exaggerated or masked the association between the variables. We conclude that the success of the family-literacy workshops varies depending on the aptitude of the person delivering the workshops and whether the material is delivered within the allotted time period.

Table 5.29: Change in Self-reported Knowledge, Self-efficacy, and Modeling as a Function of Workshop Quality

Change Score	Global Quality[a]	Material Coverage	Practitioner Style	Positive Reactions	Negative Reactions	Session Length
ΔKnowledge (n = 90)	.26* (.19*)	0.21* (.18*)	0.19 [m] (.16[m])	0.03 (.05)	0.06 (.20*)	0.22* (.15[m])
ΔKnowledge (partial) (n = 90)	.21*	.08	.22*	-.10	-.07	.16[m]
ΔSelf-efficacy (n = 89)	0.39[†] (.30[†])	0.24* (.20*)	0.32[**] (.26[**])	0.17 (.16*)	0.08 (.25[**])	0.34[†] (.22[**])
ΔSelf-efficacy (partial) (n = 89)	.23*	.04	.28**	.03	-.18*	.14[m]
ΔModeling (n = 87)	-0.05 (-.08)	0.32** (-.19)	-0.06 (-.14)	0.14 (-.08)	0.14 (.10)	-0.12 (-.03)
ΔModeling (partial) (n = 87)	-.18*	.19*	-.24*	-.07	.07	-.21*

Note: Findings that converge across analytical strategy are highlighted in grey. Change scores as a function of quality indices. Quality indices were derived from observations notes. Pearson r correlations are reported in each cell along with Kendall's tau, a non-parametric estimate of association reported in parentheses. The latter makes no assumptions about the shape of the distribution of scores. Rows featuring the designation '(partial)' report partial Pearson correlations controlling for pre-test scores. A non-parametric equivalent to the partial correlation is not available.

[a] Global Quality is the average overall quality excluding the indices based on positive and negative reactions.

[m] $p < .10$, *$p < .05$, **$p < .01$, [†] $p < .001$ (two-tailed, N = 88 to 90).

Parental Attendance in Association with Change Scores

In previous analyses, a distinction was made between parents who participated in two or fewer workshops (Participants) and those who attended three or more (Non-participants). The latter group was classified as having not "participated" in the family literacy workshops. These initial analyses were taken a step further by estimating the extent to which the remaining

variation in the rate of parental participation within those classified as Participants was related to self-reported change on the indicators considered here. It is important to note that, by excluding the Non-participants from this calculation, the range of participation values has been restricted. The analyses to follow therefore are likely to represent underestimations of the magnitude of the total dosage effect. We report the results as a conservative test of the null hypothesis.

We entered a variable representing the proportion of workshops that were attended by at least one parent of the child as a regressor in analyses based on the DinD estimator. We made use of the heterogeneity-consistent standard error estimator for these analyses, but did not employ an adjustment for clustering[89]. The results of these analyses suggest that a parent with average attendance was expected to show an average change of b = 3.08 (SE = .43) from pre-test to post-test. The average change amounts to a .96 standard deviation increase relative to baseline, which is a large effect.

The amount of gains in Knowledge from pre-test to post-test of course varied depending on the family. Some of this variability was explained by parental attendance of the workshops, DinD (89) = 3.40, SE = 1.73, for an estimated .11 standard deviation expected increase in change score for each additional workshop attended (i.e., a .10 increase in proportion of workshops attended). According to this model, families where at least one parent of the child attended all 10 workshops are expected to benefit from a .63 standard deviation change relative to the baseline over and above the gains made by a parent who attended only four workshops either in person or indirectly via the participation of the child's other parent. The result corroborates earlier analyses comparing the participant and non-participant parents[90]. Increasing exposure to the workshops is associated with self-reported gains in Knowledge.

Similarly, an average rate of parental participation was associated with a significant increase relative to the baseline on the Self-efficacy measure. This result confirms results from analyses of change reported earlier using the repeated measures t-test, b = 2.22, SE = .38. However, the associated DinD estimate was not statistically significant, DinD = 1.80, SE = 1.41, indicating that the rate of change for this outcome did not vary as a function of dosage. As in previous analyses, the effects associated with the Modeling scale did not approach statistical significance (unreported). In sum, it appears that the increase levels of exposure had a specific effect on self-reported gains in Knowledge.

5.3. SUMMARY

In this section, the program impacts for the first year of the Readiness to Learn project were presented for child- and parent-level outcomes. We summarize in turn the main findings from

[89] Fixed-effects of Community were entered to accommodate the positive correlation in the residuals introduced by the site of the workshops. This approach is more appropriate when the number of clusters is small as in this case (clusters < 10; Snijders & Bosker, 1999).

[90] It may be the case that parents who participated at higher rates tended to receive better quality workshops. The proposed mechanism behind this relationship is that parents should be more likely to participate if they perceived it to be a valuable use of their time. It is possible to compute and report estimates of association between parental attendance and workshop quality, but the results would be difficult to interpret. The nature of the quality indices are such that parents who attended more workshops may have contributed more information to the quality indices. As a result, an estimate of association computed between these two variables may reflect a methodological artefact.

each set of analyses, interpreting them summarily. A more detailed interpretation is provided in the Discussion (see Chapter 7).

Children Outcomes

For all ÉPE–AD scales except Receptive Vocabulary, the Program Daycare group began the project with a *disadvantage* relative to the Comparison Daycare group in the order of .10 to .30 standard deviations, but this disadvantage was only statistically significant for the Expressive Vocabulary measure[91]. The daycare intervention resulted in *a reversal of this disadvantage for the remainder of the first year*. DinD estimates of the program effect for Communication, Self-awareness, Cognitive Ability, and Expressive Vocabulary indicated that the Program Daycare group developed more rapidly relative to the Comparison Daycare group for all post-test periods. This advantage was statistically significant at four months post-intervention, but was not statistically detectable thereafter. In sum, despite starting in a disadvantaged position, the Program Daycare group showed either statistically better or statistically equivalent development when compared with the Comparison Daycare group over the first year of the project. Comparisons relative to the Informal Care group were equivocal.

The program effect observed four months post-intervention was buttressed by several converging lines of evidence. The results demonstrated discriminant validity in that the Program Group advantage was specific to targeted domains: Communication, Self-awareness, Cognitive Ability, and Expressive Vocabulary. No effect was observed where none was expected (i.e., Physical Ability). In a second series of analyses, dosage, program fidelity, and program quality were conceived as alternative operational definitions of the same concept: treatment intensity. Analyses of these indicators yielded a convergent set of findings[92]. All ways of conceptualizing treatment tend to converge on a coherent set of findings, supporting our contention that the observed treatment group effects are indeed valid.

To the extent that the Program Daycare group overcame an initial deficit to show equivalent development over the first year of the project, we can claim that the intervention had a positive impact on school readiness outcomes. This claim is based on the assumption that the initial disadvantage of the Program Daycare group is indicative of a pre-intervention developmental trajectory that was inferior to that of the Comparison Daycare group. We contend that had the Program Daycare group not received the intervention, they would have scored significantly below the Comparison Daycare group by the end of the first year on most of the observed measures. Nevertheless, it must be conceded that the results would have been stronger had the statistically significant positive program impacts been observed for all post-test periods. We attribute the null DinD estimates for the second and third post-tests to a lack of sensitivity in the measures used rather than to a real attenuation of the program effect (see below).

---

[91] The initial disadvantage suggests that the Program Daycare group was on course for a developmental trajectory that was inferior to that of the Comparison Daycare group prior to the intervention. If this assumption holds, then the observed treatment effects are underestimates of the true treatment effect. The observed results would then indicate that the developmental course of the Program Daycare group was "corrected" by the intervention and brought into line with that the Comparison Daycare group. According to this view, the statistically null DinD effects that were observed for Expressive Vocabulary for the second and third post-test constitute evidence of a positive treatment impact.

[92] These definitions of treatment validity were not explicitly available to the participants or to the experimenters and are therefore not as likely to be contaminated by various forms of bias. Experimenters were aware of the treatment status of participants and the expected effects of treatment status. The same cannot be said of the variables dosage and program quality/fidelity.

The fact that the observed gains due to the program are not statistically significant either 8 or 12 months post-intervention can be explained by: a) instrumentation bias caused by a lack of sensitivity in the measure as the children aged and b) a slight drop-off in the consistency with which the program was implemented after the initial four-month interval (see the *Project Implementation Report*, Bérubé et al., 2014) which means that programs provided by the two types of daycares grew more comparable over time. If the former cause is influencing the results, then we would expect program effects to emerge in analyses of the Year 2 data, from the sixth evaluation onward. From this point on, the ÉPE–AD was retooled so as to be more sensitive to individual differences in the mid to high range of ability (see Chapter 2, Section 2.5.2). Further, normalized measures of Receptive (ÉVIP–R) and Expressive Vocabulary (EOWPVT) were obtained during this period. These measures were obtained from other publishers and they discriminate among children of wide ranging linguistic ability. For these reasons, it is expected that any program effects that existed by the end of the first year will be more readily detectable in the follow-up analyses of Year 2 outcomes, especially in the last two evaluation periods. What is more, the new measures are expected to be more sensitive to the emergence of the delayed impacts of exposure to the program during the first year of the project. If the latter explanation is the cause for the observed pattern of results, then a replication of the intervention at other sites in the future might benefit from additional safeguards intended to ensure that standards of implementation are maintained at a consistently high level throughout the first year.

The Expressive Vocabulary subscale tended to show effects that were the largest. The items for this scale were gleaned from the Communication and Self-awareness scales of the ÉPE–AD. The effects observed with the latter two scales may therefore be driven in large part by items that tap expressive vocabulary. We note that the two daycare groups were found to be significantly different at the baseline period on Expressive Vocabulary[93]. As noted earlier, these baseline differences are consistent with the hypothesis of pre-existing differences in the developmental trajectories for the two groups. According to this view, the Program Daycare group was developing less quickly on this measure than the Comparison Daycare group prior to the intervention. This finding suggests that the assumption of parallel slopes that underlies the DinD estimator is unlikely to hold for these analyses. If we assume that the developmental trajectory of the Program Daycare group would have continued to be inferior to that of the Comparison Daycare group in the absence of intervention, then the DinD estimates represent conservative estimates or "underestimates" of the treatment effect. We consider that this assumption is reasonable and that the resulting DinD estimates are interpretable as a conservative test of the program impacts. In fact, the apparent correction of the Program Daycare group's developmental trajectory can be considered evidence of a positive program impact.

The sole ÉPE–AD outcome that defied our initial predictions at all evaluation periods was the Receptive Vocabulary subscale. We had expected to observe a positive program effect on this outcome due to the fact that this measure ostensibly taps the ability of children to understand French. The interpretation of this result must be tempered by the fact that this vocabulary measure was constructed from ÉPE–AD items and as a result its psychometric properties are not well documented. The correlation between it and a widely-used measure of the same construct is r = .43 (see Chapter 2, Section 2.5.2). Indeed, analyses conducted on the unstandardized version

[93] This initial difference might be explained by initial differences between the two groups in the level of French reported by the fathers in the study, which were not completely corrected by controlling for Household Type (exogamous, etc.).

of the scale indicated that this measure was insensitive to changes across time for the sample as a whole (first post-test DinD = 1.07, SE = 1.16). Thus, it is possible that the null effect reported here is the result of a lack of sensitivity rather than a real absence of an impact on Receptive Vocabulary[94].

The analyses by the child's linguistic profile confirmed that the program had the greatest impact on the ability to communicate in French when children are less exposed to the French language at the outset (i.e., Communication scale). The estimated program effects for the low-exposure condition were always consistent with a program effect with this outcome variable (i.e., in the right direction). More to the point, statistically significant program effects are observed for the low-exposure condition relative to both the Comparison Daycare group and the Informal Care group. The estimated conditional effects were also more consistent across time in these analyses as effects were observed for the first, second, and third post-test periods. The exact pattern observed depended on the language variable used as a moderator and the treatment group involved in the comparison. Nevertheless, it is clear that it is the low-exposure children who would benefit more in terms of the development of communication skills if the program were to be implemented on a wider scale.

The "new" program effects for the Communication scale in these analyses emerged most strongly for the third post-test, which took place at the beginning of the school year, immediately after summer vacation. It was for this evaluation that the language variables most consistently emerged as statistically significant moderators (e.g., for Communication scores). We would expect stronger program effects for this evaluation if the intervention was successful at increasing the quantity and quality of exposure to French during the summer months relative to the comparison condition. This would be the case, for example, if the program group parents changed their linguistic behaviours during this period (e.g., language spoken to the child; language of literacy materials) or if the Program Daycare group providers complied more consistently with the requirement that French should invariably be the language of interaction, even during the instability of the summer program. Both possibilities are in accord with the aims of the program and so the combined effect of both mechanisms may be at work.

It is worth noting that the positive program effects relative to the Comparison Daycare group on the Communication scale were corroborated by corresponding effects for the third post-test with the Expressive Vocabulary subscale (moderators: Language continuum Spoken by the child; Language of Literacy Activities), the other outcome thought to target communication specifically. The main difference between the two set of analyses is that with the Expressive Vocabulary subscale, the effects for the first post-test did not vary as a function of linguistic profile. In this case, both high- and low-exposure children experienced developmental gains for the first post-test. More generally, the moderating role of language exposure was weaker and invariably non-significant for this variable. These results are weak when compared with those obtained with the Communication scale, but we attribute this to a restricted range effect caused by the systematic missing data due to administration of the ÉPE–AD in English for some children.

[94] Examination of the distribution of this variable indicated a distinct negative skew at baseline (skewness = -1.19) that got progressively worse over the course of the study, third post-test skewness = -2.73. In contrast, the distribution of the Expressive Vocabulary subscale was quite normal even by the third post-test (skewness = -.66). We attribute this negative skew to the relative ease of the Receptive Vocabulary items and to the small number of items in this scale (n = 6). This property of the scale would explain its insensitivity in discriminating among participants.

As expected, the Language Continuum Spoken by the Child and the Language of Literacy Activities were the strongest moderators of the program effect relative to the Comparison Daycare group. The former variable represents a child's active use of the French language; the latter is an "active" variable as well and has the added advantage of being anchored to concrete behaviours, an attribute that is known to elicit more accurate responses. Findings suggest that the program provides a particular benefit for children with a deficit on these dimensions, which may be construed as an index for the child's bilingualism (i.e., the highest scores are obtained by children who speak only French, the lowest scores by those who only speak another language, usually English). By promoting the development of French-language skills, the program supports the development of additive bilingualism in the minority francophone population (see the discussion in Chapter 7). For its part, Language of Literacy Activities captures a dimension that was the direct target of the family literacy program; it is not surprising that it would be associated with observed treatment group differences.

The Language continuum spoken by the mother to the child was a particularly important moderator for the comparisons relative to the Informal Care group. When this dimension was considered, program effects on the Communication scale were observed for the first, second, and third post-tests. Even more encouragingly, the estimated program effects in these analyses got progressively larger over time, eventually culminating with a robust effect of half a standard deviation 12 months post-implementation. The language spoken by the mother may be an important moderator for comparisons with the Informal Care group because of the important contribution mothers make to the home environment.

The high French exposure children also benefited from the program, but in a different way. They began the program with a) more exposure to French, b) a corresponding better ability to communicate in French, and therefore c) the basic foundational skills necessary to learn new information in this language (Cummins, 1979). While all children showed initial (first post-test) gains on the cognitive domain of the ÉPE–AD, the high French exposure children experienced gains with the more advanced items in the scale, which were detected by the third post-test. With their French communication skills in place, these more advanced cognitive elements of the program (e.g., math and literacy skills) were in their zone of proximal development (Vygotsky, 1978). The proposed mechanism naturally accounts for the way in which the program effects have differentiated themselves as a function of baseline language exposure. In short, the results indicate that, by the time most of the sample (52%) enrolled junior kindergarten (i.e., third post-test), the high-exposure children showed a significant program effect relative to both comparison groups for the cognition domain of the ÉPE–AD.

Overall, the program improved the school readiness of all children though its effects were differentiated according to linguistic profile. The results support the contention that the program was effective at enhancing the school readiness of children who are less exposed to French by speeding up the development of their ability to communicate in French. For their part, children with a relatively high exposure to French at the start of the project appeared to derive cognitive benefits from the program primarily by the age of four. These children were manifestly better equipped to acquire such skills.

Parental Outcomes

Analyses of the parent-level outcomes provided little evidence that the family literacy workshops influence the self-reported Frequency and Language of Literacy activities of the participating families. There was some evidence of an effect for the second post-test period for Language of Literacy Activities specifically, but it was not maintained. We attribute the null effect for Frequency of Literacy Activities to a miss-calibration of the response choices provided to participants (e.g., the highest frequency available was three or more times a day, which might be considered unreasonable). However, statistically significant effects were obtained with other self-report measures. The most credible effects are associated with self-reported gains in Knowledge about child development, strategies to help him become school prepared and the availability of francophone resources in the community.

When participants and non-participants to family literacy workshops were compared on self-reported Knowledge, the resulting pattern did not show any positive signs of bias in the pre-test and post-test estimates. Moreover, we report effects whereby both workshop program quality (for certain aspects) and parental participation are associated with significantly more important gains in self-reported Knowledge. In sum, all analyses support the conclusion that the workshop component of the intervention had a real impact on this outcome. Statistically significant program impacts were also observed on self-reported Self-efficacy. This effect was attested by the program quality analyses, but not by the dosage analyses. No effect was detected based on the Modeling scale, which was intended to capture how often parents model literacy activities for their child (here we are dealing with frequency of modeling as opposed to the frequency the child is encouraged to/engages in the activity his or herself). As with the Frequency of Literacy Activities measure, we attribute this null effect to a scaling problem. An additional issue with this measure and the targeted parental opinion questions was the contamination of the pre-test measurement by exposure to the program. In sum, the reported impact effects are specific to self-reported Knowledge and Self-efficacy, while the null effects can be explained by methodological problems.

6. Impact Analyses: Second Year Follow-Up

This chapter reports analyses of developmental outcomes tracked during the second year of the study. Analyses of Year 1 outcomes concerned themselves with the immediate impact of the tested program, whereas analyses of Year 2 outcomes target instead the residual and delayed impact of exposure.

As in Chapter 6, we report the results of these analyses in two complimentary sections. The first concerns itself with analyses by treatment groups (Section 6.1.2). The second section concerns a series of secondary analyses by moderator variables (Section 6.2), including daycare dosage, program fidelity/quality, and linguistic profile. Six developmental outcomes were the object of each set of analyses: the four ÉPE–AD scales (standardized prior to analyses) and the two normalized vocabulary scales EOWPVT and ÉVIP–R.

6.1. VARIABLES RETAINED FOR ANALYSES

The variables retained for analyses of Year 2 outcomes are described in this section. We describe in separate sections the variables used as regressors and those used as outcomes.

6.1.1. Substantive Predictors and Covariates

The environment of children changed substantially during the second year necessitating the inclusion of an additional covariate in the regression-model specifications. Specifically, by the beginning of the second year of the project, some children entered school on a full-time basis, others on a part-time basis, while others were not enrolled in school at all. A categorical variable representing these varying degrees of school exposure was entered as a covariate in our regression analyses. This variable is coded as zero for the baseline period and as either one or zero starting with the fourth evaluation; the values attributed to each child are static for entire second year of the project, from the fourth to the seventh evaluation period.

For analyses of dosage effects, the Year 1 daycare exposure data was repurposed. We computed static variables on the basis of data representing daycare exposure during the first year of the study for the Program Daycare and Comparison Daycare groups. These variables represent the average hours per week spent in daycare for the period between the start of the intervention and the third evaluation. The data for daycare exposure during the summer months were not employed in this calculation by reason of the instability of a) summertime daycare arrangements and b) program delivery during this period. The Year 2 daycare exposure data were not included in analyses since the majority of children were enrolled in schools either full-time or part-time during the second year. As discussed in Chapter 4, to do otherwise would have been impractical.

Program fidelity and quality indices were identical to those used during the Year 1 analyses. As with the hours spent in daycare, the program fidelity/quality of the new preschool daycare program in the second year is no longer a meaningful source of information for most of the sample of children in Year 2. We considered it more meaningful to use the quality and fidelity ratings of the first year of the project as descriptors of individual daycares in the analyses. In other words, we were primarily interested in the effect of the programs children received in the first year of the project.

Fidelity and quality ratings were attributed to daycares for the Year 2 analyses in the following way. Fidelity values associated with each daycare for the fourth evaluation onward were based on observations collected during the interval between the second and third evaluations. Similarly, daycares were assigned the same quality ratings they were attributed in the analysis of Year 1 outcomes.

For the year two analysis, children were assigned to specific daycares in the following way. If children were not enrolled in school or were enrolled only part-time, the program quality/fidelity attributed to their current daycare at the time of evaluation was used. If enrolled in school full-time, the program quality/fidelity attributed to their last daycare before enrolling in school was used[95]. The same rules governing treatment group changes (e.g., delayed coding of changes that occur during the summer) were in effect for changes in daycare enrolment status.

6.1.2. Outcome Measures

In total, six outcomes were considered in the Year 2 impact analyses: Communication, Self-awareness, Cognitive Ability, Expressive Vocabulary (ÉPE–AD derived subscale), EOWPVT (a well-validated expressive-vocabulary scale) and the ÉVIP–R (a well-validated receptive vocabulary scale). The latter two measures were administered respectively during the sixth and seventh evaluations. As in analyses of Year 1 outcomes, the ÉPE–AD measures were standardized within time period prior to analyses (for additional details, see Section 6.1). Data for the Physical Ability scale and Receptive Vocabulary ÉPE–AD subscale were available for the fourth and fifth evaluations, but analyses based on these measures are not reported here — data from the fourth evaluation were analyzed and reported in Chapter 6. Moreover, results obtained from analyses of the Physical Ability and Receptive Vocabulary data from the fifth evaluation (unreported) yield comparable estimates to those based on the fourth evaluation.

The standardization of the outcome measures was also necessary for the Year 2 analyses because of the discrepancies in the scaling of variables noted in Section 5.1.1. Though not strictly necessary, we adopted the same transformation strategy with the Year 1 data. This decision was taken in order to maintain consistency in the presentation of the child outcome results. An incidental benefit to the standardization of our measures is that it allows us to meaningfully compare the relative size of program effects across outcome measures.

6.1.3. Results of Group Comparisons

The results of the Year 2 impact analyses are reported in this section. We first introduce the logic underlying the analyses conducted and the model-specification details associated with our regression models. The results of the treatment group analyses are then reported. For the sake of consistency, the results are presented using the same strategy as in Section 5.1.2. The sole additions to the list of outcomes are the EOWPVT and the ÉVIP–R vocabulary scales. We report analyses based on these outcomes separately from those based on the ÉPE–AD outcomes. As in all previous analyses, the results reported below are based on the French version of the ÉPE–AD. This general strategy of presentation is employed also in Section 6.2.

[95] Analyses based on daycare enrolment at the end of the first year of the study (third and fourth evaluation) yield very similar results. Thus, it is not possible to dismiss the reported effects as group composition changes.

Model Specification Details

The model-specification strategy employed in analyses of the Year 2 data is almost identical to that described in Section 5.1.2. We used a DinD estimator of the treatment effect combined with covariates whose inclusion was justified on the grounds that they would provide adjustments to compensate for variability over time in group composition. To the original list of covariates was added a set of dummy-variables representing school enrolment, which figured in the base model specification (Model 1). Otherwise, the analytical strategy remained the same as in Section 5.1.2. Of course, the evaluations considered in the analyses were different.

For the results reported below, the baseline evaluation period was used as a reference for evaluating program effects for the third post-test (Fall 2008) up until the sixth post-test (Fall 2009). These post-test periods correspond to the fourth and seventh evaluations respectively. We excluded the second and third evaluations from these analyses because their inclusion would make testing the treatment group by time interaction using the Wald-F test impossible by reason of too few degrees of freedom. This exclusion does not affect the validity of the other DinD estimates.

The main distinguishing feature of the Year 2 analyses is the use of an alternative regression model specification for analysis of the EOWPVT and the ÉVIP–R. Unlike the ÉPE–AD scales, only a single observation is available for each of the vocabulary scales. The EOWPVT was administered during the sixth evaluation and the ÉVIP–R during the seventh evaluation. Because no baseline observation is available for these measures, an ANCOVA estimator was employed to obtain the estimated treatment effect. For this purpose, the standardized Communication scores of the ÉPE–AD (Baseline evaluation) were used as a covariate in the base model specification[96]. This baseline measure is an adequate pre-test measure given that it correlates respectively at .75 and .79 (Pearson) with the ÉVIP–R and EOWPVT. The resulting estimates of the program effect are interpreted as the expected difference between groups on the vocabulary measures if the participants in each group began their participation in the project with equivalent Communication scores[97]. Because changes over time in group composition remain an issue in these analyses, we estimate treatment effects based on group membership at multiple time periods. For each time period, the children in the sample were grouped slightly differently based on their daycare arrangements. If these changes in group membership are unimportant to the results, then, regardless of how this issue is treated in the analysis, the obtained results should be essentially equivalent[98].

IMPACT RESULTS: Standardized ÉPE–AD Scales (in French)

The results of impact analyses for the ÉPE–AD scales are reported in this section: in Table 6.1 for the full-sample analyses of the Communication data and in Table 6.2 for the remaining

[96] We also considered using the ad hoc vocabulary scales that were created based on the ÉPE–AD for this purpose, but rejected the idea based on the fact that it would result in the exclusion of the 26 cases who did not meet the criteria for completing the ÉPE–AD in French a sufficient number of times during the first year of the project. As a result, using these variables as covariates would artificially restrict the range of French vocabulary scores, deflating the resulting estimates of association.

[97] In other words, baseline Communication scores are treated like any other static baseline characteristic that is included in the regression model. It was included in the base regression model because it serves to identify the treatment effect.

[98] In interpreting the results presented below, note that the largest changes in group composition (due to group switching) are observed for the fifth evaluation period (February 2009), see Table 3.2. The observed treatment effects however only emerge for the sixth and seventh evaluation periods.

ÉPE–AD scales. The results of the full-sample analyses are reported for each stage in the incremental inclusion of covariates (Models 1 through 5). For the remaining scales, we report only the results of the final model with the full set of covariates.

The results reveal a pattern that is consistent across all ÉPE–AD outcomes. Examination of both tables shows that effects for all variables emerge in one or both of the final two evaluations when the estimates are adjusted for covariates. These evaluations correspond with the switch from the original version of the ÉPE–AD to a revised version that excluded the easier items and included some additional items that were designed to be more difficult. In other words, statistically significant program effects begin to be detected when the measures used became more sensitive to individual differences among children who are moderate to strong performers. Stated in yet another way, we observe the expected treatment effects relative to the Comparison Daycare group when the measures used are more appropriately calibrated for the evaluated sample of children. The effects range in magnitude from approximately .20 standard deviations for the Communication scale to roughly .45 standard deviations for the Expressive Vocabulary score. The effects are therefore in the range of conventional benchmarks for small- and moderate-size effects. No program effect is observed relative to the Informal Care group reproducing findings reported in Section 5.1.2.

Table 6.1: Year 2 Difference in Difference (DinD) Program Effects for Standardized ÉPE–AD Communication Scores — French Version (Full Sample)

	Incremental Inclusion of Covariates — Errors Clustered by Daycare									
	(1)		**(2)**		**(3)**		**(4)**		**(5)**	
Program Effects	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Program Daycare vs. Comparison Daycare at Baseline (G1[a] vs. G2)	.205	.184	.181	.181	.203	.124	.244	.121	.204	.121
DinD 3rd Post-test	-.121	.247	-.084	.251	-.073	.204	-.069	.208	-.087	.202
DinD 4th Post-test	.091	.165	-.019	.135	-.038	.103	-.033	.110	-.039	.108
DinD 5th Post-test	-.013	.188	-.077	.146	-.121	.128	-.112	.129	-.129	.126
DinD 6th Post-test	-.205	.230	-.246	.163	-.286*	.134	-.266*	.133	-.282*	.132
Program Daycare vs. Informal Care at Baseline (G1 vs. G3)	-.065	.180	-.205	.168	-.153	.092	-.093	.085	-.151	.090
DinD 3rd Post-test	-.049	.361	-.013	.324	-.024	.249	-.026	.255	-.070	.259
DinD 4th Post-test	-.052	.380	.015	.280	-.007	.208	.019	.204	-.009	.221
DinD 5th Post-test	.059	.192	.193	.150	.173	.111	.178	.111	.138	.115
DinD 6th Post-test	-.140	.250	.013	.188	-.006	.151	.063	.123	.048	.125
Omnibus Wald F-tests:	F (8, 16) = .836		F (8, 16) = 1.32		F (8, 16) = 1.27		F (8, 16) = 1.24		F (8, 16) = 1.19	
N participants:	239		234		233		233		226	

Note: The outcome has been standardized, which means that the DD effects can be interpreted in terms of standard deviation units. ***Program Daycare group (G1) is in reference, which means that negative values of DD indicates a positive program effect***. Standard errors were estimated via the SPSS implementation of the heterogeneity consistent 'robust' White estimator (White, 1980). Further, standard errors were clustered by daycare to adjust for correlation in the residuals over time. The observed effect is only statistically significant once socio-demographic characteristics are inserted into the analyses. Model specification details are provided in Section 6.1.3.

[m] denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed).

[a] G1, G2, and G3 denote Program Daycare group, Comparison Daycare group, and Informal Care group respectively.

Table 6.2: Year 2 Difference in Difference (DinD) Program Effects for Standardized Subscale Scores of the ÉPE–AD (French Test-takers Only)

	ÉPE–AD scales - All Covariates Included — Errors Clustered by Daycare							
	Communication		**Self-awareness**		**Cognitive Ability**		**Expressive Vocabulary**	
Program Effects	DinD	Robust SE	DinD	Robust SE	DinD	Robust SE	DinD	Robust SE
Program Daycare vs. Comparison Daycare at Baseline (G1[a] vs. G2)	.209[m]	.113	.198	.168	.259	.141	.347	.175
DinD 3rd Post-test	-.064	.193	-.084	.175	-.042	.219	-.148	.159
DinD 4th Post-test	-.044	.105	-.111	.126	-.100	.191	-.209	.153
DinD 5th Post-test	-.118	.121	-.359[*]	.134	-.255[*]	.099	-.479[*]	.147
DinD 6th Post-test	-.248[*]	.118	-.294	.173	-.327[*]	.105	-.461[*]	.199
Program Daycare vs. Informal Care at Baseline (G1 vs. G3)	-.095	.083	-.082	.135	.185	.138	-.033	.135
DinD 3rd Post-test	.050	.253	.041	.118	-.174	.229	.074	.141
DinD 4th Post-test	.091	.201	.160[m]	.088	-.033	.170	.155	.201
DinD 5th Post-test	.106	.139	-.134	.135	-.250[m]	.121	-.102	.228
DinD 6th Post-test	.125	.102	-.196	.142	-.222	.138	-.169	.168
Omnibus Wald F-tests:	F (8, 16) = 1.06		F (8, 16) = 4.49[†]		F (8, 16) = 3.69[*]		F (8, 16) = 2.50[m]	
N participants:	209		208		209		208	

Note: The outcomes have been standardized, which means that the DinD effects can be interpreted in terms of standard deviation units. Analyses based on the sub-sample of participants who completed the ÉPE–AD in French at least twice during the first year of the study. Thus, the sample is skewed towards those children who are stronger in French. Expressive Vocabulary was the only scale to show a positive effect prior to the inclusion of covariates. Program Daycare group (G1) is the reference group, which means that ***negative values of DinD denote a positive program effect***. Model specification details are provided in Section 6.1.3.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed).

[a] G1, G2, and G3 denote Program Daycare group, Comparison Daycare group, and Informal Care group respectively.

EOWPVT and ÉVIP–R Vocabulary Scales (Unstandardized)

The estimated program effects for the EOWPVT and ÉVIP–R outcomes are reported in this section. The program effects were re-estimated multiple times on the basis of different treatment group assignments, reflecting group membership as it was for evaluations four through seven inclusive. These vocabulary measures were taken at a single point in time, but repeating the test using different versions of the treatment group variable based on past group membership allows an assessment of whether the observed results depend critically on group changing. If this is the case, then the validity of the results may be called into question. The results obtained via regression analyses of the EOWPVT scale with incremental inclusion of covariates are reported in Table 6.3. The results obtained via regression analyses of the ÉVIP–R scale with incremental inclusion of covariates are reported in Table 6.4.

The results of both series of analyses reveal a program effect relative to the Comparison Daycare group in the order of two to three words. The effect ceases to be statistically significant once the family composition variables are inserted as covariates (see the results for the fourth and fifth specifications). Despite this drop from statistical significance, the difference between groups is still estimated at approximately two words[99]. In other words, the inclusion of covariates resulted in a modest reduction in the size of the estimated program effect that was just sufficient to eliminate the effect.

The ÉVIP–R analyses distinguished themselves by showing signs of a program effect relative to the Informal Care group before all covariates where inserted into the analyses. In this case, little hint of a program effect remained by the time all covariates were inserted into the analyses (see the fifth specification). This observed dependence on the inclusion of covariates is consistent with the fact that the Informal Care group was revealed to be substantially different from the Program Daycare group on a number of characteristics in preliminary analyses in Section 5.4. Given this situation, it is unsurprising that the inclusion of these characteristics in the model should cause a substantial change in the estimated effect.

In sum, there is a hint in the data indicating a program effect relative to the Comparison Daycare group (EOWPVT, ÉVIP–R) and the Informal Care group (ÉVIP–R) when all covariates are included in the analyses. It remains to be seen whether these trends will manifest themselves as a significant effect when a more fine-grained approach is taken in analyses based on Year 1 exposure to treatment and program fidelity/quality indices. Realistically, the effect in question is in the order of only two to three words. Even if the effect was statistically significant, it would be difficult to argue that it represents a substantive impact for the intervention, even acknowledging that it represents of a real gain in a child's vocabulary of much more than three words.

[99] It is useful to remember that these two words represent an actual difference in total vocabulary that is much larger in terms of actual number of words known to the children. We do not have a 'conversion' rule that would allow ÉVIP–R scores to be expressed precisely as a function of total vocabulary size. Nevertheless, the two-word effect should not be mistaken as indicating a growth in real vocabulary of only two words.

Table 6.3: EOWPVT Scores (Taken at the Sixth Evaluation, June 2009) as a Function of Group Membership at the Four Evaluation periods of Year 2 (Unstandardized)

		Incremental Inclusion of Covariates — Robust Standard Errors Clustered by Daycare									
		(1)		**(2)**		**(3)**		**(4)**		**(5)**	
		Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1 [a] vs. G2	Groups at the 4th Evaluation	-3.95*	1.39	-3.42*	1.49	-3.04*	1.43	-1.85	1.60	-1.92	1.52
	Groups at the 5th Evaluation	-2.90m	1.42	-3.48*	1.67	-3.24m	1.63	-2.68	1.74	-2.84	1.74
	Groups at the 6th Evaluation	-3.21*	1.49	-3.48*	1.66	-3.49m	1.70	-2.69	1.85	-2.99	1.87
	Groups at the 7th Evaluation	-	-	-	-	-	-	-	-	-	-
G1 [a] vs. G3	Groups at the 4th Evaluation	-.821	1.10	-1.31	1.45	-.722	1.39	0.08	1.52	-0.36	1.44
	Groups at the 5th Evaluation	-2.28	2.15	-2.28	2.15	-1.88	2.12	-1.19	2.33	-1.87	2.29
	Groups at the 6th Evaluation	-3.06	2.20	-2.73	2.12	-2.55	2.20	-1.66	2.41	-2.42	2.35
	Groups at the 7th Evaluation	-	-	-	-	-	-	-	-	-	-
	N	236		231		230		229		223	

Note: EOWPVT scores denote the number of words correctly identified and were collected only once (the 6th Evaluation). Tests of the program effect were conducted using four alternative groupings: treatment group membership at the fourth, fifth, and sixth evaluations. The results with groupings from the seventh evaluation are not reported because this evaluation occurred after the administration of the EOWPVT. The purpose of testing these largely redundant tests was to evaluate the extent to which the migration of participants from one group to another may have influenced the results. ***Program Daycare group (G1) is in reference, which means that negative values of DD indicates a positive program effect***. Standard errors were estimated via the SPSS implementation of the heterogeneity consistent "robust" White estimator (White, 1980). Further, standard errors were clustered by daycare to adjust for the common effect of the daycare context. Model specification details are provided in Section 6.1.3.

[a] G1, G2, and G3 denote Program Daycare group, Comparison Daycare group, and Informal Care group respectively.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed, 20 degrees of freedom).

Table 6.4: ÉVIP–R Scores (Taken at the Seventh Evaluation, October 2009) as a Function of Group Membership at the Four Evaluation periods of Year 2 (Unstandardized)

		Incremental Inclusion of Covariates — Robust Standard Errors Clustered by Daycare									
		(1)		**(2)**		**(3)**		**(4)**		**(5)**	
		Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1 [a] vs. G2	Groups at the 4th Evaluation	-4.89*	1.74	-4.46†	1.43	-3.25*	1.41	-2.29	1.57	-2.45	1.51
	Groups at the 5th Evaluation	-2.25	1.78	-3.18*	1.46	-2.21	1.35	-1.93	1.55	-1.75	1.55
	Groups at the 6th Evaluation	-2.66m	1.51	-3.26*	1.38	-2.69m	1.40	-2.32	1.56	-2.37	1.40
	Groups at the 7th Evaluation	-2.66m	1.51	-3.34*	1.36	-2.75m	1.36	-2.21	1.71	-2.21	1.48
G1 [a] vs. G3	Groups at the 4th Evaluation	-1.32	1.73	-1.86	1.17	-.691	1.14	.306	1.26	.804	1.35
	Groups at the 5th Evaluation	-2.71	2.22	-3.31*	1.42	-2.12	1.27	-.58	1.28	-.215	1.53
	Groups at the 6th Evaluation	-3.35	2.04	-3.44*	1.25	-2.58*	1.23	-1.17	1.28	-1.01	1.34
	Groups at the 7th Evaluation	-3.35	2.04	-3.43*	1.23	-2.60*	1.20	-.912	1.35	-.725	1.37
	N	236		232		231		231		224	

Note: ÉVIP–R scores denote the number of words correctly identified and were collected only once (the 7th Evaluation). Tests were conducted using four alternative groupings: treatment group membership at evaluations four through seven inclusive. The purpose of testing these largely redundant tests was to evaluate the extent to which the migration of participants from one group to another may have influenced the results. ***Program Daycare group (G1) is in reference, which means that negative values of DD indicates a positive program effect***. Standard errors were estimated via the SPSS implementation of the heterogeneity consistent "robust" White estimator (White, 1980). Further, standard errors were clustered by daycare to adjust for the common effect of the daycare context. Model specification details are provided in Section 6.1.3.

[a] G1, G2, and G3 denote Program Daycare group, Comparison Daycare group, and Informal Care group respectively.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed, 20 degrees of freedom).

6.2. MODERATORS AND MEDIATORS OF THE INTERVENTION EFFECT

The purpose of the results presented in this section is to corroborate the findings presented earlier with a finer-grained set of analyses. We address here the issue of whether the amount of daycare exposure makes a difference over and above Year 1 treatment group membership and whether the intensity of exposure to the dimensions targeted by the new preschool daycare program, as indexed by the program quality and fidelity indices, are related to Year 2 outcomes. The results of these analyses are reported below.

6.2.1. Fine-grained Definitions of Dosage: Average Hours Spent in Daycare during Year 1

It is interesting at this point to ask whether the amount of Year 1 exposure affects Year 2 outcomes over and above what can be accounted for by treatment group membership. Another way to formulate this question is: Does the effect of dosage vary as a function of treatment group for the analysis of second-year outcomes? In what follows, we report the results of analyses based on the full set of covariates (Model 5) in separate sections for the ÉPE–AD outcomes and the two normalized vocabulary measures (i.e., EOWPVT, ÉVIP–R). The results of the Year 1 dosage analyses based on the full-set of covariates (Model 5) are reported in Table 6.5 for the ÉPE–AD standardized scales and in Table 6.6 for the normalised vocabulary measures.

Specification Details

As in analyses of dosage presented in Chapter 5, the effect of dosage here is estimated by way of a three-way interaction test: treatment group x dosage x time (DinDinD) indicators. Children with a minimum average of 10 hours per week of exposure were kept for analyses. Other children were excluded as outliers in the distribution. The reported analyses are therefore based on children with a minimum amount of daycare exposure only (for an example of a more complex approach for examining the effect of exposure on the *treated only* with a much larger sample size, see Behrman, Cheng, & Todd, 2004). In sum, the present analyses required a different approach from that reported for the Year 1 dosage data (see Section 5.1.3). The list of covariates was identical to those used for the analyses by treatment groups presented in the preceding section.

Year 1 Exposure Effects on Year 2 ÉPE–AD standardized outcomes

The Year 1 dosage effects after adjustment for the full set of covariates (Model 5) are presented in Table 6.5 for the standardized ÉPE–AD measures. Unlike in Section 5.1.2, the results for the Communication scale based on the full sample of participants are not reported, because they do not differ markedly from those reported here.

First, there was little evidence of treatment group differentiated DinDinD dosage effects (not reported), which is attested by the non-significance of the Wald-F values reported in Table 6.5. This null result may be attributable to the relatively small number of participants within each group ($n < 100$). The addition of the second cohort to the dataset may increase the stability of the associated estimates and, by extension, the probability of observing an impact.

Second, we observe here stronger treatment group effects than in analyses reported in Table 6.2. The former analyses allowed some group changing whereas the analyses reported in Table 6.5 fixed all participants in their Year 1 daycare[100]. Note that a significant DinD effect is observed for the Expressive Vocabulary measure (ÉPE–AD) for the fourth, fifth and sixth post-test, whereas in analyses that allowed group changing, the effect for the fourth post-test was non-significant. This discrepancy can be explained by the fact that the largest treatment group migration occurred prior to the fourth post-test period (see Section 2.4). We conclude that the group changing washed out this DinD effect in the preceding analyses. The results of these analyses confirm and re-enforce those reported in Table 6.2.

[100] Children were fixed in their Year 1 daycare so that the interaction between dosage (average hours spent in daycare) and the treatment group membership variable would be meaningful. It would not be desirable, for instance, to treat incorrectly hours spent in one of the program daycares as having been spent in a comparison daycare, all because of a group change that occurred in the second year of the project.

Table 6.5: Year 1 Treatment Group and Dosage as Predictor of Year 2 Outcomes (French Test-takers Only)

	EPE-AD scales – All covariates included ── Errors Clustered by Daycare							
	Communication		**Self-awareness**		**Cognitive Ability**		**Expressive Vocabulary**	
Year 1 Treatment Group Effects (G1 [a] vs. G2)	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
At Baseline	.197*	.086	.211m	.120	.281*	.107	.365*	.126
DinD At 3rd Post-Test	-.070	.164	-.085	.155	-.068	.177	-.195	.134
DinD At 4th Post-Test	-.144	.115	-.161	.123	-.060	.138	-.298*	.137
DinD At 5th Post-Test	-.209	.124	-.395*	.125	-.257*	.105	-.585*	.136
DinD At 6th Post-Test	-.272*	.129	-.302m	.152	-.348*	.092	-.514*	.183
2-way Interaction: Group x Time - DinDinD (Wald F)	F (4, 17) = 1.52		F (4, 17) = .622		F (4, 17) = 1.99		F (4, 17) = 5.01†	
N	148		148		148		148	
3-Interaction Test of Dosage Effect - DinDinD (Wald F)	F (4, 17) = 1.70		F (4, 17) = .622		F (4, 17) = 1.99		F (4, 17) = .339	

Note: ***Program group is in reference; therefore negative DinD values denote a positive program effect.*** Group differences at pre-test were observed, which are cancelled out by the DinD estimator. G1 refers to the Program Daycare group and G2 refers to the Comparison Daycare group. The confidence intervals of all G1 and G2 effects overlap, but G1 dosage is a significant predictor while G2 dosage is not. Children with fewer than 10 weekly hours of daycare exposure on average during the Year 1 were discarded from analyses (for an example dosage effects that are evaluated by considering only the treated, see Behrman, Cheng, & Todd, 2004). Specification: Details provided in the body of the text.

[a] G1 and G2 denote Program Daycare group and Comparison Daycare group respectively.

[m] denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed).

Year 1 Exposure Effects on Year 2 EOWPVT and ÉVIP–R vocabulary scales (Unstandardized)

The Year 1 treatment group and dosage effects on the EOWPVT and ÉVIP–R vocabulary scales are reported in Table 6.6. We report three tests in this table: a) treatment group effects (i.e., the difference between the program and comparison group), b) the effect of dose (an estimate that is specific to the G1 group, the Program Daycare group), and c) the Dose by Group interaction (DinD), which is the difference in the effect for dose for the Comparison Daycare group (i.e., G2) relative to the G1. If the first test is significant, it means that a statistically significant difference has been observed between the two groups. If the second test is significant, it means that a statistical significant effect of daycare exposure has been observed. Finally, if the latter test is significant, it means that estimate of the G1 dosage effect is statistically different than the G2 dosage effect. Alternatively, if the latter test is non-significant, it means that the effect of dose is statistically equivalent for the two groups. First we discuss the ÉVIP–R results and then those of the EOWPVT.

Note that the effects for Year 1 treatment group, Year 1 dose and the dose by treatment group interaction are not statistically significant for the ÉVIP–R for any specification. These results are consistent with those reported in Table 6.4. A trend for a small effect exists in Model 1 but disappears entirely by Model 5. In sum, the two set of analyses by treatment group are consistent: if there is treatment impact on this normalized vocabulary measure, it is subtle. Had a real pre-test been available, an estimate of the treatment effect based on the DinD estimator could have been used, which would have provided better control over the baseline characteristics of participants and a more powerful test of the program effect.

In a similar manner, the EOWPVT shows a null Year 1 treatment group effect. In this case, however, an effect of daycare dosage emerges but it is in the opposite of the expected direction. The estimate for the G1 dosage effect is a) significantly negative for Models 4 and 5 and b) is statistically equivalent to the dosage effect for G2. In sum, there is no evidence of a program impact here. Because there was no hint of this effect at baseline, its credibility depends entirely on whether the model is correctly specified (Maris, 1998; Glymour, et al., 2005). Inspection of the residuals did not reveal any obvious problems, but given the small sample size per group ($n < 100$), we choose not to interpret the direction of this effect too strongly. The insertion of data from the second cohort into the analyses should improve the stability of the estimates. Alternative model specifications that may shed more light on the issue are explored in the Section 7.2.3, where we report analyses by the linguistic profile of participants.

Table 6.6: Year 1 Daycare Exposure Effects as a Function of Daycare Type — ÉVIP–R and EOWPVT Scores (Unstandardized)

	Incremental Inclusion of Covariates — Errors Clustered by Daycare									
	(1)		**(2)**		**(3)**		**(4)**		**(5)**	
Program Effects	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
ÉVIP–R scores										
Year 1 Treatment Group	-2.39	1.741	-2.09	1.641	-1.33	1.96	-.691	2.22	-.851	2.11
Dosage [b] effect (G1)	.205	.161	.138	.176	-.071	.192	-.121	.208	-.205	.225
Dosage x Group (DinD) G1 [a] vs. G2	-.543[m]	.259	-.647	.293	-.398	.232	-.319	.244	-.223	.256
EOWPVT scores										
Year 1 Treatment Group	-2.46[m]	1.27	-2.16[m]	1.15	-1.77	1.35	-.828	1.44	-1.15	1.58
Dosage effect (G1)	-.004	.177	-.123	.136	-0.279	0.132	-.308[*c]	.119	-.344[*c]	.135
Dosage x Group (DinD) G1 [a] vs. G2	-.137	.206	-.131	.183	0.011	0.159	.055	.160	.090	.176
N	160		156		155		154		152	

Note: The Year1 dosage variables were centered about the grand mean prior to analyses. Standard errors were estimated via the SPSS implementation of the heterogeneity consistent "robust" White estimator (White, 1980). Further, standard errors were clustered by daycare to adjust for the common effect of the daycare context. Children with fewer than 10 weekly hours of daycare exposure on average during the Year 1 were discarded from analyses (for an example dosage effects that are evaluated by considering only the treated, see Behrman, Cheng, & Todd, 2004). Specification: Details are discussed in the body of the text.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed, 20 degrees of freedom).

[a] G1 and G2 denote Program Daycare group and Comparison Daycare group respectively.

[b] Dosage here is defined as the average hours spent weekly in daycare during the first year of the study for G1 and G2 daycares respectively.

[c] This effect lacks credibility given that a) no hint of it was detected in Model 1 and b) the number of observations per group is relatively small ($n < 100$).

6.2.2. Fine-grained Definitions of Program Integrity: Daycare Program Fidelity and Quality

Year 1 exposure to the tested daycare program was found to have a positive effect on development during the second year of the project when ÉPE–AD standardized outcomes were examined (by treatment group). In this section, we verify whether another way of defining intensity of treatment, the global fidelity and quality of program delivery, will be similarly associated with Year 2 outcomes[101]. We report the results of these analyses in separate sections for the ÉPE–AD standardized outcomes and the Vocabulary Scales.

Model Specification Details

Data were analyzed in the same way as in Section 5.2 of analyses of Year 1 outcomes. Estimates of the fidelity and quality of the new preschool daycare program were inserted in place of treatment group membership in the basic regression model specification for the DinD analyses (Model 5 for the ÉPE–AD outcomes; Models 1 through 5 for the normalized vocabulary scales). In secondary analyses, we verified whether the treatment group effects and the effects estimated on the basis of program fidelity/quality indices are redundant. In this case, the regression model included direct effects and DinD effects for treatment group membership controlling for the corresponding program fidelity or quality versions of these effects.

Global Program Fidelity/Quality effects on Year 2 Outcomes: ÉPE–AD scales (Standardized)

We report DinD estimates of the long-term effect of global program fidelity/quality on the Year 2 outcomes in Table 6.7. We do not report the results based on the full-sample analyses of Communication scores given that the results do not differ substantially from those reported below. As in analyses involving Year 1 daycare exposure effects (dosage), the results are based on the sub-sample of children enrolled in daycare only. We only report the results based on the global estimates of the tested daycare program fidelity and quality. This mode of presentation conveys all the relevant information succinctly (interested readers are directed to Chapter 6 where a more detailed approach to the analysis of these variables was taken).

The results reveal positive effects for Global Quality relative to baseline for the sixth and seventh evaluations (fifth and sixth post-tests) for the Communication and Cognitive Ability scales of the ÉPE–AD. For Global Fidelity, a corresponding effect is observed only for Cognitive Ability. In sum, for these two outcomes, the pattern of results that was observed in analyses by treatment groups is confirmed.

The Global Fidelity effect on Cognitive Ability indicates that a .20 increase on this variable is associated with a gain score advantage of .24 standard deviations, which is a small effect. The Global Quality effect indicates that a two-unit increase in the quality index is associated with a .26 standard deviation advantage in gain score, which is also a small effect. Though small, these effects were obviously statistically detectable. In sum, for these two outcomes we observe positive evidence of program impacts using three different operational definitions of treatment

[101] We leave out the results for specific dimensions of program quality (e.g., quality of reading) because doing so does not change the story of the Year 2 results in a substantive way.

intensity: a) treatment group membership, b) Year 1 dosage/exposure, and now c) global fidelity/quality of the new preschool daycare program.

Secondary analyses revealed that the treatment group DinD effect for the seventh evaluation that were observed in the treatment group analyses of Communication scores is dramatically reduced and is no longer statistically significant when Global Quality is statistically controlled, [adjusted DinD = -.070, SE = .127]. Similar reductions are observed in the adjusted DinD estimates for treatment group effect associated with Cognitive Ability, sixth evaluation [adjusted DinD = -.207, SE = .151]; seventh evaluation [adjusted DinD = -.287*, SE = .117]. The treatment group effects are almost completely eliminated when the effects associated with Global Fidelity are used as covariates, sixth evaluation [adjusted DinD = .045, SE = .277]; seventh evaluation [adjusted DinD = .047, SE = .211]. The results support the idea that the associated treatment group effects have their effect on the ÉPE–AD outcomes through program quality and/fidelity.

Table 6.7: Global Program Quality and Fidelity as Predictor of Year 2 Outcomes

	ÉPE–AD Scales — Errors Clustered by Daycare							
	Communication		**Self-awareness**		**Cognitive Ability**		**Expressive Vocabulary**	
	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
Global Fidelity								
At Baseline	-.607	.314	-.677	.429	-1.03†	.329	-1.00	.460
DinD At 3rd Post-Test	-.087	.526	-.060	.465	.276	.461	-.006	.405
DinD At 4th Post-Test	.106	.371	.370	.471	.576	.483	.466	.480
DinD At 5th Post-Test	.213	.353	.637	.441	.948†	.273	.977^{m}	.535
DinD At 6th Post-Test	.658	.459	.658	.459	1.18†	.255	1.15^{m}	.541
Omnibus Wald-F test	F (4, 11) = .772		F (4, 11) = 2.57^{m}		F (4, 11) = 5.32*		F (4, 11) = 3.64*	
N	148		148		148		148	
Global Quality								
At Baseline	-.023	.051	-.024	.068	-.012	.052	-.030	.063
DinD At 3rd Post-Test	-.040	.079	-.004	.071	-.101	.055	.059	.075
DinD At 4th Post-Test	-.019	.032	.065	.041	.037	.065	.096 ^{m}	.048
DinD At 5th Post-Test	.099*	.042	.049	.073	.132*	.049	.073	.087
DinD At 6th Post-Test	.143*	.051	.073	.072	.133*	.044	.118	.093
Omnibus Wald-F test	F (4, 11) = 8.23†		F (4, 11) = 1.32		F (4, 11) = 7.72†		F (4, 11) = 1.64	
N	148		148		148		148	

Note: The effect of new preschool daycare program quality and fidelity at baseline is an estimate of the pre-test association and does not constitute a program effect. This association is cancelled out by the DinD estimator. ***Positive DinD estimates denote positive effects for daycare program quality and/or fidelity.*** Daycare program quality and fidelity indices were derived from qualitative analyses. Fidelity scores range from 0 to 1 while quality scores range from 0 to 7. Interpretation of the coefficients must be adjusted according to these differences in scale. Results are based on daycare enrolment at the time of testing, but the results are comparable to analyses based on group membership at the end of the Year 1.

^{m} denotes significance at $p < .10$, * $p < .05$, † $p < .01$ (all tests two-tailed).

Global Program Fidelity/Quality effects on Year 2 Outcomes: EOWPVT and ÉVIP–R (Unstandardized)

We report estimates of the impact of global program fidelity/quality indices on EOWPVT and ÉVIP–R scores in Table 6.8. Results for the full set of analyses with incremental inclusion of covariates are reported (Models 1 through 5). In addition, as in preceding analyses with these two outcomes, we used an estimator for the treatment effect based on statistical control of the baseline Communication scores of children. The results reveal that, after adjusting statistically for the covariates in the model, a positive impact for Global Quality is observed for both the ÉVIP–R and the EOWPVT. In both cases, the effect is in the order of two words per unit change in program quality. In contrast, estimates of Global Fidelity of program were not found to be consistently related to measures of vocabulary.

As anticipated, analyses by quality proved to be more sensitive to the impact of program quality than the simpler analyses based on treatment group assignment. The result is important in that it lends credibility to our assumption that the treatment group effect that was observed relative to the Comparison Daycare group (but which did not persist with the inclusion of the full set of covariates) is a) real and therefore b) likely to emerge as statistically significant with an increased sample size. We do not report global program fidelity/quality adjusted estimates here because the treatment group effect was not statistically significant. It remains to be seen whether the insertion of data from the second cohort of children enrolled in the Readiness to Learn project into the analyses will confirm our expectation that the hint of an effect observed here is real.

Table 6.8: Year 2 Vocabulary Scores as a Function of Global Program Fidelity and Program Quality Indices — ÉVIP–R and EOWPVT Scores (Unstandardized)

	Incremental Inclusion of Covariates — Errors Clustered by Daycare									
	(1)		**(2)**		**(3)**		**(4)**		**(5)**	
Program Effects	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE	Effect	Robust SE
ÉVIP–R scores										
Global Fidelity	4.56	3.75	5.89	4.50	4.40	4.91	4.05	5.35	3.31	4.81
Global Quality	1.97[m]	1.04	2.20 [m]	1.10	2.44[†]	.697	2.13[*]	.789	1.79[*]	.724
EOWPVT scores										
Global Fidelity	6.12[*]	2.42	4.98	3.22	4.56	3.78	3.06	4.89	2.29	4.84
Global Quality	1.67[†]	.402	1.90[†]	.516	1.99[†]	.473	2.04[†]	.542	1.95[†]	.521
N	154		151		150		149		147	

Note: Global Fidelity and Global Quality indices were derived from qualitative analyses. Global program fidelity scores range from 0 to 1 while global program quality scores range from 0 to 7. Standard errors were estimated via the SPSS implementation of the heterogeneity consistent "robust" White estimator (White, 1980). Further, standard errors were clustered by daycare to adjust for the common effect of the daycare environment. Results are based on daycare enrolment at the time of testing, but are comparable to analyses based on group membership at the end of the Year 1. Specification: Details are discussed in Section 6.2.2.

[m] denotes significance at $p < .10$, [*] $p < .05$, [†] $p < .01$ (all tests two-tailed, 20 degrees of freedom).

6.2.3. Linguistic Characteristics of the Sample

A number of effects were reported in the main analyses of Year 2 outcomes that were marginally significant and/or whose magnitude was less important than anticipated (see the treatment group effects with the ÉVIP–R and EOWPVT). It makes sense at this stage to verify whether a targeted list of covariates might moderate the treatment group effects. As in analyses of Year 1 outcomes, we will consider here whether there is evidence that the baseline linguistic characteristics of the sample might be useful in identifying the optimal target population for the tested program.

The intention of the program was to support the development of francophone children in a minority context. The rationale of the intervention hinges on the premise that many children receive insufficient exposure to French language and culture to support their development as Francophones. As in Section 5.1.3, we anticipated that the children for whom this need is greatest would show more important positive impacts for the program. This possibility was explored in the analyses to follow.

The data were analyzed in the same manner as in Section 5.1.3, but in this case the results varied more widely here according to which moderator variable was considered. To compensate for this additional complexity, the presentation of the results was simplified. We report only estimates for the "low exposure to French" condition or the "high exposure to French" condition according to which condition showed an enhancement of the observed program effect. The conditional estimates that were eliminated intentionally from the tables are described summarily in the body of the text.

Communication Scale (Standardized): Unlike Section 5.1.3, we report only conditional estimates for the low exposure to French condition (see Table 6.9). Similar to the findings reported in Table 5.22, no significantly positive program effects were observed in the "high exposure" condition (unreported here). For some indicators, the results suggest that the group differences tend to be larger by the end of the second year for those whose exposure to French is low. However, the pattern of results is not consistent across linguistic profile indicators, which suggests that a) exposure to French as a unitary construct is a less relevant moderator of Year 2 development than Year 1 development and, conversely, b) it will prove useful to consider the idiosyncratic attributes of the indicators that show the effect in interpreting the results. For example, it was argued previously that the Language Continuum Spoken by the Child and the Language of Literacy Activities would be the most consistent moderators of the program effect. The characteristics of each comparison group may also be relevant to determining which moderator is most relevant.

Relative to the Comparison Daycare group, this moderating effect of linguistic profile was statistically significant for the fifth post-test when the Language Continuum Spoken by the Child, the Language of Literacy Activities, and Household Type (i.e., the FOLS defined purely on the basis of the language spoken to the child) were used. This effect was not statistically significant when it was estimated for the sample as a whole (see Tables 6.1 and 6.2). It would appear then that the Year 2 effect for the Communication scale emerges 4 months earlier when exposure to French is low (according to selected moderators). By the sixth post-test, the same effect is observed for the average of the sample as a whole (see Tables 6.1 and 6.2).

Relative to the Informal Care group, the magnitude of the program effect varied significantly as a function of the Language of Literacy Activities and the Language Continuum Spoken by the Mother to the Child for all three post-test periods reported in Table 6.9. The convergence of these two findings is sensible given that the mother is habitually the primary care giver for children in the Informal Care group. Evidence in favour of a program effect was strongest for the fourth and fifth post-test as the development of the Program Daycare group and the Informal Care group was estimated as statistically equivalent by the sixth post-test.

Table 6.9: Estimated Program Effects for Standardized Communication Scores (French) Conditional on Low Exposure to French — Full Sample

	Moderators of the Program Effect											
	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother-child)***		***Language Continuum (father-child)***	
	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1[a] vs. G2												
Baseline	.283*	.112	.330*	.100	.422*	.153	.438*	.153	.352*	.113	.137	.137
DinD 4th Post-test	-.216	.195	-.256	.199	-.204	.225	-.270	.187	-.172	.164	-.226	.141
DinD 5th Post-test	-.379$^{\Delta}$	.224	-.407$^{*\Delta}$	.195	-.200	.227	-.454$^{*\Delta}$	.198	-.294	.241	-.258	.155
DinD 6th Post-test	-.313^{m}	.177	-.510*	.174	-.209	.192	-.192	.213	-.264	.197	-.140	.169
G1 vs. G3												
Baseline	-.292*	.095	-.194^{m}	.111	.009	.103	.024	.116	-.192	.116	-.269	.106
DinD 4th Post-test	-.216	.283	-.698$^{*\Delta}$	.165	-.373$^{\Delta}$	.220	-.426$^{\Delta}$	.247	-.386$^{*\Delta}$	.172	-.307	.341
DinD 5th Post-test	.037	.189	-.286$^{\Delta}$	.171	.072	.195	-.403$^{*\Delta}$	.180	-.069$^{\Delta}$	.195	-.041	.304
DinD 6th Post-test	.018	.347	-.338$^{\Delta}$	.313	.076	.292	-.177	.319	-.158$^{\Delta}$	.354	-.002	.302
3-way Interaction (Wald F)	F (5, 15) = 3.63*		F (5, 15) = 6.45*		F (5, 15) = 6.12*		F (5, 15) = 18.96†		F (5, 15) = 3.73*		F (5, 15) = 4.73*	
N	221		221		221		221		221		215	

Note: The results for the third post-test are not reported but figured in the analyses. The household variables are dichotomous; therefore the estimates represent the contrast of Endogamous Francophone families versus Other family types. For the remaining two variables, the table reports regression model estimates of the program effects at arbitrarily selected levels of French language use. The upper-bound estimate was calculated for the maximum scale score, while the lower-bound estimate was estimated for the value falling one standard deviation below the mean. ***The program group is in reference therefore negative effects denote an advantage for the Program Daycare group.*** The baseline evaluation and Program Daycare group are the reference categories for all DinD and DinDinD effects.

[a] G1, G2, and G3 denote Program Daycare group, Comparison Daycare group, and Informal Care group respectively.

$^{m}p < .10$, $^{*}p < .05$, $^{**}p < .01$,† $p < .001$; $^{\Delta}$ denotes effects that vary significantly as a function of varying degrees of exposure to French $p < .10$.

Self-awareness Scale (Standardized): Analyses of the Self-awareness scale offered very little evidence that the magnitude of the program effects varied significantly and systematically as a function of linguistic profile and some of the results obtained in these analyses were contradictory. Rather that present a needlessly confusing picture, the results of these analyses are omitted from this report. In any case, this particular outcome was not expected to depend on linguistic profile in the same way as the measures of communication and vocabulary.

Cognitive Ability Scale (Standardized): In Year 1 analyses (Section 5.1.3), program effects in relation to Cognitive Ability were largest for children with a high degree of exposure to French. The Year 2 data presented a similar pattern. Estimates of the program effect for high-exposure children are reported in Table 6.10. Note that with one exception none of the estimates for the "low exposure" condition was statistically significant (unreported)[102].

Analyses presented here indicate that it was the high-exposure participants who were primarily responsible for the program effects reported in Table 6.2. For three of the six potential moderators, the magnitude of the program effect varied significantly as a function of family linguistic profile according to the global Wald test. Three of the six moderators showed statistical significance for the sixth post-test for the comparison of daycare groups (as denoted by the symbol Δ). Indeed, the Comparison Daycare group generates significant positive program effects for the fifth and sixth post-tests across all language moderators.

The findings reported in Table 6.10 also indicate that the program has a significant effect relative to the Informal Care group when it is conditioned on linguistic profile indicators. This latter effect was not observed with Language of Literacy Activities and it was specific to the fifth post-test. In other words, the more robust program effect is still obtained with the daycare comparisons, even in the present analyses.

[102] The exception was the DinD effects relative to the Informal Care group for the 5th and 6th post-test condition on the Language Continuum Spoken by the Father to the Child. The estimate effects were statistically significant and larger than for the high-exposure group. The source of this discrepant effect is unclear.

Table 6.10: Estimated Program Effects for Standardized Cognition Scores (French) Conditional on High Exposure to French

	Moderators of the Program Effect											
	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother-child)***		***Language Continuum (father-child)***	
	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1[a] vs. G2												
Baseline	.364*	.165	.390*	.153	.216	.155	.253^{m}	.124	.363*	.142	.392*	.150
DinD 4th Post-test	-.274	.259	-.304	.251	-.256	.212	-.246	.211	-.201	.243	-.234$^{\Delta}$	.236
DinD 5th Post-test	-.443**	.128	-.354*	.157	-.477**	.129	-.358**	.126	-.474**	.124	-.394**	.120
DinD 6th Post-test	-.513*	.200	-.500**	.142	-.584$^{**\Delta}$	.161	-.465**	.161	-.501$^{**\Delta}$	.144	-.501$^{**\Delta}$	.171
G1 vs. G3												
Baseline	.389^{m}	.207	.331*	.145	.330^{m}	.192	.352^{m}	.186	.336*	.146	.337	.206
DinD 4th Post-test	-.163	.256	-.037	.249	-.266$^{\Delta}$	.204	-.240$^{\Delta}$	.221	-.020	.217	-.104	.232
DinD 5th Post-test	-.441†	.100	-.173	.100	-.463†	.105	-.278**	.089	-.361†	.074	-.295*	.105
DinD 6th Post-test	-.345	.278	-.248	.226	-.524$^{*\Delta}$	.199	-.341	.208	-.249	.220	-.297	.244
3-way Interaction (Wald F)	F (5, 15) = 5.53**		F (5, 15) = .719		F (5, 15) = 2.50^{m}		F (5, 15) = 7.54†		F (5, 15) = 1.24		F (5, 15) = 1.98	
N	211		211		211		211		211		210	

Note: The household variables are dichotomous; therefore the estimates represent the contrast of Endogamous Francophone families versus Other family types. For the remaining two variables, the table reports regression model estimates of the program effects at arbitrarily selected levels of French language use. The upper-bound estimate was calculated for the maximum scale score, while the lower-bound estimate was estimated for the value falling one standard deviation below the mean. ***The program group is in reference therefore negative effects denote an advantage for the Program Daycare group.*** The baseline evaluation and Program Daycare group are the reference categories for all DinD and DinDinD effects.

[a] G1, G2, and G3 denote Program Daycare group, Comparison Daycare group, and Informal Care group respectively.

$^{m}p < .10$, $^{*}p < .05$, $^{**}p < .01$,† $p < .001$; $^{\Delta}$ denotes effects that vary significantly as a function of varying degrees of exposure to French $p < .10$

Expressive Vocabulary Scale (Standardized): Conditional estimates of the program effect are reported for the "low exposure to French" condition in Table 6.11. The pattern of results for the "high exposure to French" condition was very similar to that obtained when the average effect was estimated for this outcome based on the entire sample; it would therefore be redundant to report these estimates here (results available upon request). In these analyses, positive program impacts were observed for the fifth and sixth post-tests (see Table 6.2).

Two main findings presented in Table 6.11 set these results apart from those presented in Table 6.2. When the daycare groups are compared, the year-end positive program effect emerges 4 months earlier at the fourth post-test for the children with less exposure to French. Here the Linguistic Continuum Spoken by the Child and the Linguistic Continuum Spoken by the Father to the Child were statistically significant moderators of the effect (as denoted by the symbol Δ). The second interesting distinction is the fact that an effect relative to the Informal Care group emerges briefly for the fifth post-test period. As in the preceding analyses (i.e., Cognitive ability), the latter effect does not persist into the sixth post-test like that involving the two daycare groups.

Findings based on the FOLS ("Household" based primarily on the First Official Language Spoken) in Table 6.11 are clearly an outlier. We ascribe this peculiar result to the fact it is not as explicitly tied to the actual behaviour of people in a child's environment as the other indicators. Recall that all the other language variables used as moderators here are based on variable that are explicitly linked to interactions with the child.

Table 6.11: Estimated Program Effects for Standardized Expressive Vocabulary Scores (French) Conditional on Low Exposure to French

	Moderators of the Program Effect											
	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother-child)***		***Language Continuum (father-child)***	
	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1[a] vs. G2												
Baseline	.222	.148	.307*	.150	.295	.192	.417	.176	.347	.165	.302^{m}	.172
DinD 4th Post-test	-.502$^{**\Delta}$	.151	-.468*	.169	-.111	.128	-.366**	.122	-.219^{m}	.118	-.432$^{**\Delta}$	.134
DinD 5th Post-test	-.556†	.131	-.516*	.183	-.227$^{\Delta}$	.168	-.496**	.147	-.384*	.155	-.558**	.171
DinD 6th Post-test	-.396^{m}	.200	-.446*	.203	-.126$^{\Delta}$	.209	-.236	.233	-.233	.233	-.367^{m}	.204
G1 vs. G3												
Baseline	-.280	.380	-.202	.464	.129	.276	.245	.388	.172	.284	-.120	.444
DinD 4th Post-test	-.224	.384	-.515$^{\Delta}$	.340	.049	.184	-.233	.394	-.277	.370	-.184	.532
DinD 5th Post-test	-.441	.463	-.716^{m}	.346	-.516	.361	-.927$^{*\Delta}$	.501	-.830$^{*\Delta}$	.397	-.677	.628
DinD 6th Post-test	-.344	.586	-.530	.483	-.360	.488	-.792	.597	-.827	.495	-.580	.691
3-way Interaction (Wald F)	F (5, 15) = .615		F (5, 15) = 10.67†		F (5, 15) = 2.68*		F (5, 15) = 7.38†		F (5, 15) = 3.23*		F (5, 15) = 3.65*	
N	211		211		211		211		211		201	

Note: The household variables are dichotomous; therefore the estimates represent the contrast of Endogamous Francophone families versus Other family types. For the remaining two variables, the table reports regression model estimates of the program effects at arbitrarily selected levels of French language use. The upper-bound estimate was calculated for the maximum scale score, while the lower-bound estimate was estimated for the value falling one standard deviation below the mean. ***The Daycare Program group is in reference therefore negative effects denote an advantage for the Program Daycare group.*** The baseline evaluation and Program Daycare group are the reference categories for all DinD and DinDinD effects.

[a] G1, G2, and G3 denote Program Daycare group, Comparison Daycare group, and Informal Care group respectively.

$^{m}p < .10$, $^{*}p < .05$, $^{**}p < .01$,† $p < .001$; $^{\Delta}$ denotes effects that vary significantly as a function of varying degrees of exposure to French $p < .10$.

EOWPVT & ÉVIP–R: Analyses by linguistic profile were conducted as above for both the EOWPVT and the ÉVIP–R. Analyses were based on treatment group membership at the time of testing. The results for the ÉVIP–R are not reported here as there was very little evidence that linguistic profile was a moderator for this variable (other intervention studies have failed to observe program effects with this variable, see below; but for a counterexample, see Maltais, 2007). The exception was the variable Language of Literacy Activities. For children from families that scored low on this variable, a significant program effect is observed relative to the Informal Care group [Effect = -6.30*, SE = 2.28]. Otherwise, the estimated effects for this outcome were statistically indistinguishable from those reported in previous analyses where not effect was discerned (see Tables 7.3 and 7.4).

When compared with the ÉVIP–R, the EOWPVT proved more sensitive to the intervention when program effects were estimated for children with a low-exposure to French. This pattern of results is congruent with prior research findings that also found the EOWPVT to be the more sensitive measure (see Whitehurst, et al., 1988; Arnold, Loginan, Whitehurst, & Epstein, 1994; Hargrave & Sénéchal, 2000). In addition, the ÉPE–AD "Expressive Vocabulary" subscale suffered from the restricted range issue, whereas this was not a factor for the EOWPVT. Estimated effects based on the EOWPVT for the 'low exposure to French' children are reported in Table 6.12. As expected, the Linguistic Continuum Spoken by the Child at baseline is the most consistent moderator of the program effect, followed by the Language of Literacy Activities. Children who used French less often to express themselves during the baseline period appeared to benefit most from the intervention in terms of their expressive vocabulary development. This was true whether the Program Daycare or Informal Care groups were the basis for comparison. The effects are in the order of 4 to 6 words, which represents a relatively modest impact (roughly equivalent to those observed in a study evaluating the impact of the "Head Start" program on vocabulary, Wasik, Bond, & Hindman, 2006).

Table 6.12: Estimated EOWPVT Program Effects for "Low-exposure" Children

	Moderators of the Program Effect											
	Language Continuum (Child)		***Language of Literacy Activities***		***Household (FOLS)***		***Household (Child-Relative)***		***Language Continuum (mother-child)***		***Language Continuum (father-child)***	
	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
G1[a] **vs. G2**	-4.16$^{*\Delta}$	1.97	-3.43	2.22	-.948	2.52	-2.25	2.86	-1.81	2.66	-3.91	2.37
G1 vs. G3	-6.23$^{m\Delta}$	3.04	-9.24$^{\dagger\Delta}$	2.05	-1.81	4.07	-5.27	4.65	-5.27$^{\Delta}$	3.38	-7.47$^{m\Delta}$	4.26
Language x Group Interaction (Wald F)	F (2, 21) = 3.99*		F (2, 21) = 26.99†		F (2, 21) = 1.87		F (2, 21) = .82		F (2, 21) = 4.33*		F (2, 21) = 1.89	
N	222		222		222		222		222		215	

Note: The household variables are dichotomous; therefore the estimates represent the contrast of Endogamous Francophone families versus other family types. For the remaining two variables, the table reports regression model implied estimates of the program effects at arbitrarily selected levels of French language use. The lower-bound estimate is reported for the value falling one standard deviation below the sample mean. ***The Program Daycare group is in reference therefore negative effects denote an advantage for the Program Daycare group.*** The baseline evaluation and Program Daycare group are the reference categories for all DinD and DinDinD effects. Treatement groups were those at the time the data for the outcome variable were collected.

[a] G1, G2, and G3 denote Program Daycare group, Comparison Daycare group, and Informal Care group respectively.

$^{m}p < .10$, $^{*}p < .05$, $^{**}p < .01$,† $p < .001$; $^{\Delta}$ denotes effects that vary significantly as a function of varying degrees of exposure to French $p < .10$.

6.3. SUMMARY

Throughout this report, we have adopted a strategy based on the principle of converging operations whereby the concept "treatment intensity" or "treatment exposure" was defined in multiple ways so as to verify the extent to which these analyses produce a coherent set of results. In analyses of Year 1 outcomes, estimates of the program effect based on treatment group membership, dosage, and program quality led to the same conclusion: the tested program had a small but significant impact on the development of children enrolled in program daycares relative to those enrolled in comparison daycares that was detectable for the first post-test only, but which is consistent with a "catch-up" effect for all evaluations. The effect relative to the Informal Care group was not statistically significant in analyses by treatment group, a null result that could not be verified using the other two indices of treatment exposure which are specific to daycares. The level of agreement across estimates of treatment intensity varied depending on the outcome being considered though cross-validation of many effects was obtained. ***Analyses by linguistic profile of Year 1 outcomes*** indicated that by the time most children entered school (52% of the sample by the third post-test) the program had a positive impact. Those children who began the intervention with low exposure to French showed gains in their communication (Communication, ÉPE–AD) and expressive vocabulary (Expressive Vocabulary, ÉPE–AD) skills relative to their peers in the two comparison groups (Daycare, Informal Care). Conversely, those children who began the intervention with high exposure to French showed significant gains in their cognitive development by the time most children enter school for junior kindergarten (age four).

In analyses of the ***Year 2 outcomes***, the same strategy of converging operations was applied. Two sets of analyses by treatment group were conducted. The first of these included all three treatment groups and allowed participants to change groups (for the standardized ÉPE–AD scales, see Table 6.2) and the second of these was based on the daycare groups only — it was conducted within the context of analyses by dosage — and did not allow participants to change groups (for the standardized ÉPE–AD scales, see Table 6.5). The results were virtually identical: statistically significant program effects (based on the DinD estimator) for the fifth and sixth post-test. The effect on Expressive Vocabulary was much stronger and more persistent in the second analyses (fourth, fifth and sixth post-test). In sum, we observe in Year 2 that the Program Daycare group overcame initial disadvantage to ***demonstrate faster developmental gains relative to their non-program daycare peers***. Children in the Program Daycare group appeared to enter junior kindergarten "ready to learn" and made corresponding gains on the ÉPE–AD outcomes by the end of the school year. Analyses by treatment group failed to indicate a stable pattern of results in favour of a positive program impact relative to the Informal Care group.

The treatment group effects associated with the Comparison Daycare group were observed for all four Year 2 ÉPE–AD scales: Communication, Self-awareness, Cognitive Ability, and Expressive Vocabulary. We expected our analyses to gain in sensitivity when the SRDC-modified version of the ÉPE–AD began to be used. Indeed, the emergence of statistically significant program effects for most measures coincided with the switch to the modified ÉPE–AD beginning with the sixth evaluation. These effects were not attested by dosage effects differentiated by treatment group; in all cases, varying levels of daycare exposure had a similar effect regardless of treatment group.

Analyses by linguistic profile of ÉPE–AD Year 2 outcomes produced a pattern of results that was generally similar to that reported in Chapter 6. Evidence for positive program effects on communication and expressive vocabulary skills were strongest when baseline exposure to French was low. In contrast, evidence of a positive impact on cognitive abilities was strongest when baseline exposure to French was high. Unlike Year 1 analyses, the language-dependent effects relative to the Informal Care group did emerge, even if they were less consistent across analyses than those involving the daycare groups.

The treatment group effects for Communication and Cognition scores were attested by corresponding effects for Global Daycare quality. Global Fidelity also attested the cognition effect. For the version of the scales used at that period, the Communication and Cognition scales are heavily populated by items that tap "phonological awareness," which is a predictor of academic success. The result supports the contention that the levers manipulated by the program are having a positive impact on children participating in the program. In addition, Year 2 analyses featured two outcomes that were not based on the ÉPE–AD: the EOWPVT and the ÉVIP–R. The results featuring these outcomes were equivocal when treatment groups were compared, hinting toward a positive program impact.

However, a positive association for both vocabulary measures with Global Quality of the daycare program was detected which was robust to the inclusion of covariates, which again supports the idea that the levers manipulated by the program in an attempt to influence the development of children are valid. We take this issue up in Chapter 8 (Discussion & Conclusion). As expected, the program effects on these measures were more easily detected for children with lower levels of exposure to French (see analyses by linguistic profile). Children who started the project with relatively low levels of active exposure to the French language showed a program effect on the expressive vocabulary measure (EOWPVT), relative to both the Comparison Daycare and Informal Care groups. Similar to other daycare interventions (see Hargrave & Sénéchal, 2000, and several examples cited in that paper), the benefits in this study favoured expressive over receptive vocabulary.

7. Discussion and Conclusion

The goal of the present research was to evaluate the impact of a two-pronged intervention on young francophone minority children and their parents. The intention of the tested program was to promote the development of the linguistic abilities and school readiness of francophone children in a minority language context. Theorists have proposed that the driving force behind the acquisition of such competencies is the interaction of a child with his or her environment (Bronfenbrenner & Morris, 1998). Drawing upon Bronfenbrenner's (1979) ecological theory of child development, we conceived of a child's environment as a series of partially overlapping spheres of influence, which includes the home, the daycare (or school), and the community. As reviewed in the introduction, the majority language can influence speakers of the minority language within each of these spheres, including the two primary settings: the home and the daycare. Accordingly, a two-pronged approach was undertaken for maximum effect on child outcomes (Reese, et al., 2010): a) family literacy workshops to positively impact the environment in the home by way of the parents and b) a high-quality francophone daycare program to influence the children more directly.

The *combined effect* of these interventions on the development of children was evaluated by way of a quasi-experimental design comparing the families who received the intervention to two comparison groups: children in daycare (Daycare Comparison group) and children in an informal care setting (Informal Care group). The first comparison served to identify the effect of the intervention relative to the daycare services that are typically available to francophone children. The second comparison served to identify the effect of the tested program relative to informal care settings. A number of positive impacts for the intervention were observed in the analyses reported in Chapters 6 and 7, which are discussed in more detail in the sections to follow.

Each individual result on its own is informative, but might lack validity. Together, the findings converge so as to mutually reinforce each other, providing a high level of confidence in the findings that was made possible by our use of a mixed-methods approach to program evaluation. It entailed the use of a range of tools, both quantitative and qualitative in nature, from several information sources, all selected based on the research objectives. The wealth of the information thus collected facilitates the triangulation of research findings, a strategy that aids researchers in arriving at sound conclusions. Moreover, the complementarity of the data collected provides a more complete, more nuanced picture of the phenomenon under study. Among other sources, child assessments, parent surveys, observations in daycare classes and during family literacy workshops, as well as administrative data (e.g., the record of children's presences and absences from daycare), all served as fodder for regression analyses. The main findings of this report are presented summarily for child and parent outcomes respectively in Tables 7.1 and 1.2.

Table 7.1: Summary of the Main Findings of the Readiness to Learn project First Cohort Findings Report: Child Outcomes

	Year 1 Effects *(Baseline versus 1st, 2nd, 3rd post-test)*	**Year 2 Effects** *(Baseline versus 3rd, 4th, 5th, 6th post-test)*
Program integrity and child outcomes		
Fidelity/Quality/Dosage	***Program Fidelity and Quality (program differentiation):*** • Fidelity and quality were shown to **significantly** mediate early impacts of the program on G1 child outcomes relative to G2 children [a]. ***The importance of program fidelity and quality for impacts:*** • Fidelity and quality were shown to **significantly** mediate impact of the program on the G1 children relative to G2 children [a].	***The importance of program fidelity and quality for impacts:*** • Fidelity and quality were shown to **significantly** mediate the impact of the program on G1 child outcomes relative to G2 children [a].
Child Outcomes		
School-Readiness	***Program group children (G1) show:*** • As anticipated, **non significant** gains in physical ability relative to both the G2 and G3; ***As program group children (G1) entered JK (3rd post-test) [b, c]:*** • **Significant** gains in French communication skills for children with less active exposure to the language relative to G2 • As children enter JK, **significant** gains in cognitive skills (as per French test) for those children with more active exposure to the language relative to both G2 and G3 [a]	***As program group children (G1) enter SK (6th post-test) [b]:*** • Overall **significant** gains in communication, self-awareness, and cognitive ability (re-emerged with better-calibrated measures) relative to G2; • **Significant** gains in French communication skills for children with less active exposure to the language relative to G2, and G3; • **Significant** gains in cognitive skills (tested in French) for those children with more active exposure to the language relative to G3.
Vocabulary	***Program group children (G1) show:*** • **Significant** gains overall in expressive vocabulary in French relative to G2 ***As program group children (G1) entered JK (3rd post-test) [b, c]:*** • **Significant** gains in expressive vocabulary in French for children with less active exposure to the language relative to G2 (ÉPE–AD subscale)	***As program group children (G1) enter SK (6th post-test) [b]:*** • Overall **significant** gains in expressive vocabulary in French relative to G2 (ÉPE–AD) • **Significant** gains in French expressive vocabulary relative to both G2 and G3 (ÉPE–AD subscale; EOWPVT) for children with less active exposure to the language

Note: School readiness outcomes included: a) Communication, b) Self-awareness, c) Cognitive Ability, and d) Physical Ability. Vocabulary measures included: a) expressive and receptive vocabulary measures based on the ÉPE–AD (SRDC-generated) and b) normalized expressive (EOWPVT) and receptive (ÉVIP–R) vocabulary. Some effects were significant for some time periods and not others. **We highlight the significant results of interest, while non-significant results are implied.**

[a] G1 refers to the Program group; G2 refers to the Daycare Comparison group; G3 refers to the Informal Care group.

[b] Children were about 3 years old while in daycare, 4 years old when entering junior kindergarten (JK), and 5 years old when entering senior kindergarten (SK).

[c] Just over half the sample enrolled in JK.

Table 7.2: Summary of the Main Findings of the Readiness to Learn project First Cohort Findings Report: Parent Outcomes

	Year 1 Effects *(Baseline versus Post-test(s))*	**Year 2 Effects** *(Baseline versus Post-test(s))*
Program integrity and parent outcomes		
Fidelity/Quality/Dosage	• Positive effects for workshop participants are **significantly** related to elements of program implementation (attendance, quality)	
Parent Outcomes		
Literacy Activities	• **No significant effect** on frequency or language of use either for activities involving the child or for modeling of literacy behaviours: Attributed to measurement problems	• **No significant** differences on frequency or language of use between treatment groups; Measurement problems
Knowledge & Attitudes & Opinions	• For workshop participants, **significant** gains in sense of *Self-efficacy* in their role as parents and *knowledge* of child development and best practices relative to pre-test and a self-selected group of non-participant families. • **No significant effect** on opinions: attributed to measurement problems	• Not applicable

Note: We highlight the significant results of interest; most non-significant results are implied.

7.1. READINESS TO ATTEND SCHOOL IN FRENCH?

The main question this report attempted to answer is whether the intervention had a positive impact on the development of school readiness markers for participating children. The answer to this question appears to be yes when all sources of evidence are considered together. This is true whether the impact of the program is estimated for the entire sample or specifically for children with a given level of baseline exposure to the French language.

7.1.1. Average Program Effects

Overall, the comparisons indicated positive impact effects based on the school readiness measures for the Program Daycare group relative to the Comparison Daycare group for both Year 1 (Chapter 5) and Year 2 outcomes (Chapter 6). The same pattern of results was not, however, evidenced relative to the Informal Care group. The results were robust to the inclusion of covariates and in many instances were confirmed using one or more "alternative" definitions of "treatment intensity," such as program quality/fidelity and dosage. The family literacy workshops were found to have a positive influence on some parental outcomes, which suggests that a part of the program's impact on child development may have been achieved indirectly by way of the parents. This would be consistent with the philosophy used to develop the program and with previous research indicating that the combined effect of interventions directed at children and parents exceeds that of programs that focus on one aspect only (Brooks-Gunn, Berlin, & Fuligni, 2000; Reese, et al., 2010). We now turn to a more detailed account of the results.

Program Daycare versus Comparison Daycare

The focus of the study was on the ability of francophone children to meet their future educational challenges in French. Indeed, the school readiness of program participants (in French) was positively influenced by the new preschool program, as indicated by statistically significant effects for three of the four dimensions of the ÉPE–AD. The observed program effects were specific in that no effect was observed where none was predicted (i.e., with the Physical Ability dimension of the ÉPE–AD), and general in that the targeted outcomes variables were affected positively. The findings converge onto a consistent pattern of results for school-readiness outcomes, all of which depend heavily on mastery of the French language. The list of positively affected outcomes includes: Communication, Self-awareness, and Cognitive Ability. A program effect was also observed on the Expressive Vocabulary subscale composed of items from the ÉPE–AD. The effect sizes associated with these effects are in the order of .30 to .50 standard deviations, which falls within the range of effect sizes observed in comparable studies conducted in the United States, where effects were reported in the range .26 to .79 standard deviations (for a review, see Howes, et al., 2008).

In the present study, the Program Daycare group began the project with a *non-significant* disadvantage relative to the Comparison Daycare group on three of the four domains of the ÉPE–AD. The Program Daycare group overcame this initial disadvantage to show an initial developmental advantage, where superior developmental gains relative to the Daycare Comparison group were observed 4 months post-implementation on three of the four domains of

the ÉPE–AD and the Expressive vocabulary subscale (first post-test). These initial gains re-emerged, as expected, 16 and 20 months post-implementation (fifth and sixth post-tests).

As discussed in Section 6.3, we anticipated the re-emergence of the program effect based on the use for these evaluations of a version of the ÉPE–AD that was better able to discriminate among children of varying skills and abilities in Communication, Cognition and Self-awareness (i.e., the SRDC-modified ÉPE–AD). School enrolment (junior kindergarten, age four) was statistically equivalent across experimental groups at the start of the second year. Nevertheless, it was entered as a covariate to control for the impact of this variable that might result from the significantly inferior rates of school enrolment among children in the informal care group that were observed later in the year due to group changing. In short, the resurgence of the treatment effect in the second year cannot be attributed to an association between school enrolment and treatment group. It is, however, possible and even likely that better prepared children were better able to take advantage of the school environment. To the extent that the program contributed to increased school preparedness, this would translate into a positive program effect in the second year relative to the comparison groups, all other things being equal. Indeed, this is what was observed (see below).

The results support the conclusion that the children who received the intervention are more school ready than their peers in the Comparison Daycare group by the time they are five years old (senior kindergarten). Children enrolled in the Program Daycare group were found to make significantly greater gains on school readiness outcomes than they would have in the absence of intervention. Such gains are important because better school readiness is associated with smoother transition into the classroom (Cunha, Heckman, Lochner, & Masterov, 2006), where they will be more likely to fully take advantage of the social and academic environment it provides (Janus & Offord, 2000). Of course, we refer here specifically to school-readiness measures that were administered in French.

In future reports, we will investigate whether these gains in school readiness carry over to more immediate precursors of academic achievement, such as phonological awareness and alphabetic knowledge (Lonigan, Shanahan, Westberg, & The National Early Literacy Panel, 2008; Sénéchal, 2005). In the modified version of the ÉPE–AD (sixth and seventh evaluations), both concepts are represented in the Cognitive Ability scale, and two of the four items of Communication scale tap phonological awareness. Both scales showed positive impacts for the tested program. On this basis, we anticipate that positive program effects will be observed on validated instruments, designed specifically to measure these constructs. The results of such analyses will figure in future reports.

French vocabulary measures were taken to meet the French language objective of this research. The goal was to determine whether the tested program positively influenced the development of French vocabulary in a minority context. For the Expressive Vocabulary subscale that was constructed based on ÉPE–AD items, positive program impacts were observed for the fifth, sixth and seventh evaluations. The analyses were conducted first by acknowledging group changes in the second year of the project and second by fixing group membership as it was by the end of the first year. Normally, the two analyses yielded equivalent results, suggesting that group changes in the second year of the project did not have a meaningful influence on the results. Exceptionally, we observed an effect for the fifth evaluation that emerged only in analyses that ignored second-year group changes. In this particular case, the migration of

children away from the Program Daycare group immediately prior to the fifth evaluation obscured the tested program's true impact. All told, estimates associated with each administration of the ÉPE–AD in the second year of the program were consistent with a positive impact on the development of expressive vocabulary. No effect was observed on the ÉPE–AD-based Receptive Vocabulary subscale, but we attribute this to the poor psychometric properties of this measure.

Analyses based of the normalized measures of receptive and expressive vocabulary, the ÉVIP–R (seventh evaluation) and the EOWPVT (sixth evaluation) respectively, suggested a modest program effect of roughly two to three words that did not persist when all covariates were included in the model. The null result was obtained regardless of how treatment groups were defined (e.g., group membership at the end of the first year versus at the time of testing). The absence of an effect on these two vocabulary measures contrasts with the positive program effects observed with the ÉPE–AD Expressive Vocabulary subscales. We attribute this incongruence to a methodological difference in the two analyses. The normalized vocabulary scales were analyzed using an estimator based on statistical control; the ÉPE–AD scales were analyzed using the DinD estimator. The DinD estimator would have provided better control of baseline characteristics had it been an option with the normalized vocabulary scales (i.e., selection bias, Cronbach & Furby, 1970; Winship & Morgan, 1999). More encouraging results were obtained for children who began the project with less exposure to the French language (see below).

Program Daycare versus Informal Care

Despite the fact that the Informal Care group scored significantly higher on measures of French language and culture (Chapter 4), the Program Daycare group maintained a statistically equivalent developmental trajectory on the ÉPE–AD outcomes for the first two years of the project (before and after adjustment for linguistic characteristics). At no point, however, did the analyses comparing the two groups indicate a statistically significant program impact. This was true despite the fact that community effects were controlled experimentally and a variety of other family and child characteristics were controlled statistically. There is no evidence to suggest that the children who received the intervention distinguish themselves from the children in the Informal Care group in the main analyses of developmental outcomes. The expected positive program effects emerged instead in the analyses by linguistic profile.

7.1.2. Program Effects by Linguistic Profile

The literature on bilingualism makes a clear distinction between cases where the acquisition of a second language benefits the general development of a child and cases where development appears to suffer (e.g., Landry, Allard, & Deveau, 2009). These disparate types of bilingualism are called respectively "additive" and "subtractive." The critical distinction between these two types of bilingualism appears to be whether the *mother tongue* is developed sufficiently to support the acquisition of a second language without incurring delays in age-appropriate development of cognitive abilities or competency in the mother tongue (UNESCO, 2010). The issue is especially important when the mother tongue is also the language of instruction.

This minimal condition does not appear to be met for many francophone children living in a minority context, a population which has sometimes been known to lag behind their peers on

outcomes such as reading achievement (Bussière, et al., 2001; Chartier, et al., 2008; Conseil canadien sur l'apprentissage, 2008; Education Quality and Accountability Office, 2007). The causes behind this achievement gap are reasonably well understood. The literature on bilingualism explains it by appealing to a relatively straightforward mechanism: exposure to the language. For bilingualism to be additive, a minimal threshold of language exposure or use needs to be exceeded in the mother tongue (for a review, see Pearson, 2007). For various complex reasons (e.g., motivational, greater exposure to the language of the majority across multiple settings; Landry, et al., 2009), this minimal threshold is higher when the mother tongue is a minority language (Pearson, Fernandez, Lewedag, & Oller, 1997; Vihman, Lum, Tierry, Nakai, & Keren-Portnoy, 2006). Consequently, francophone children in a minority setting who are raised as bilinguals require special support if they are to improve their changes of achieving additive bilingualism.

It is in this spirit that we investigated the possibility that children who were exposed to languages other than French (usually English) would benefit most from the piloted preschool program. We refer here explicitly to the joint contributions of both the daycare and family literacy components in exerting an influence over the linguistic environment of children. We considered that such children are at risk of developing bilingualism of the subtractive type given the fact they live in a minority language context, anticipating the largest effects for the program with this population specifically because it was designed to address their needs. This prediction was confirmed is a series of analyses where program effects were estimated for children with varying levels of baseline exposure to the French language.

Program Daycare versus Comparison Daycare

By the end of the first year (i.e., third post-test), most of the children in the project (52%) were enrolled in school on either a part-time or full-time basis (age four, junior kindergarten). It is useful then to consider whether the children in the Program Daycare group were more "school ready" at this point than they would have been in the absence of treatment. The main analyses of program effects did not provide overwhelmingly positive evidence that this was the case. However, further analyses revealed that these mixed results were partly due to variable responses to treatment. Children who received the intervention do indeed appear to be more "school ready," but the dimension impacted by the program depends on the characteristics of the child.

We report evidence that children who were *less exposed* to French at the start of the project benefited most in terms of the acquisition of basic linguistic skills as measured by the Expressive Vocabulary and Communication scales. This is true when most of the sample entered school in junior kindergarten (age four) and the advantage was observed again when virtually the entire sample entered senior kindergarten (age five) by the end of the second year of the project (according to the ÉPE–AD and the EOWPVT). The *Readiness to Learn in Minority Francophone Communities: Project Implementation Report* (Bérubé et al., 2014) and the results reported in Chapter 6 of the present report clearly show that quality *of reading* activities is an important feature in distinguishing tested program from what is normally available in French-language daycare program. The observed effects of the program could be explained by this feature of the tested program. The seeming specificity of the effect to expressive language is consistent with other evaluations of preschool interventions (i.e., Hargrave & Sénéchal, 2000) that emphasize the importance of interactive language activities (e.g., dialogic reading), while the

apparent specificity of the program effect to language skills is consistent with studies of Francophones that have considered this variable as a moderator (i.e., Maltais, 2007).

In contrast, children who entered the study with a *high level of exposure* to French, and a correspondingly higher facility with the language, were in a position to develop the skills tapped by the Cognitive Ability scale. The data presented here suggest that basic language skills facilitate the achievement of secondary gains in cognitive skills (e.g., literacy). This explanation finds support with developmental theories of skill acquisition (Bloom, 1976; Vygotsky, 1978) and with the importance of language mastery for academic success (Cummins, 1979; Doherty, 1997; Hindman, Skibbe, Miller, & Zimmerman, 2010). It is also consistent with the findings of Maltais (2007) who reported that the effect of a full-day pre-kindergarten program on language development depended on the linguistic profile of the child, but that the broader cognitive benefits (i.e., gains in reading ability) were observed even for children whose exposure to French was high, and for such children the effects were stronger.

In sum, the tested program's high level of quality resulted in more effective transmission of skills and knowledge. It is simply the case that children derived benefits according to their readiness to learn particular types of skills in a French environment. A generally similar pattern of results was obtained in both the Year 1 and Year 2 analyses, which tends to indicate that the early effects persisted even when they escaped detection in certain statistical analyses. Taken together, these results have implications for how the program is implemented in other contexts, because the expected benefits of the program seem to vary both qualitatively and quantitatively according to the language exposure of the clientele.

Program Daycare versus Informal Care

At the outset, we had hypothesized that the program would result in positive program effects relative to both comparison groups. This expectation was based on the fact that the intervention's two components (daycare program, family literacy workshops) were designed to effect change respectively in the environment of children at daycare and at home. The hypothesized effects were not observed in the main analyses relative to the Informal Care group. The absence of an effect in these analyses is perhaps attributable to the fact the Informal Care group benefited from relatively high levels of exposure to French during the baseline period (see Table 4.4). Indeed, the expected effects were revealed when a more subtle analysis was conducted that allowed program effects to vary as a function of exposure to French.

By the end of the first year, "at-risk" children with lower exposure to French when they began receiving the intervention had developed more quickly on the Communication dimension than children in informal care with a similar level of exposure. Virtually the same pattern of results was obtained regardless of the variable that was used as an indicator of the child's exposure to French. However, the Language Spoken by the Mother to the Child yielded the most consistent "conditional" effects across time. The baseline language exposure of children in an informal daycare setting (e.g., at home) is a necessarily more important determinant of development than for children who spend a substantial amount of time in daycare. This accounts for the stronger role of this characteristic in moderating the program effects for comparisons relative to this group. In the second year of the project, the program group showed superior expressive vocabulary according to the EOWPVT. Again, the seeming specificity of the effect to

expressive language is consistent with other evaluations of preschool interventions (Hargrave & Sénéchal, 2000).

In sum, there is evidence of positive program impacts on the development of francophone children whether regular daycare or informal care children are taken as the basis for comparison. For children at the greatest risk of subtractive bilingualism, the program appears to have its most marked effect on language and communication skills. The children who were least at risk benefited from the better cognitive outcomes that come from exposure to a higher quality daycare program.

7.2. PARENT-LEVEL OUTCOMES

The amount of one-on-one attention that a daycare educator can devote to a given child is limited (Wasik, 2008), which is why some have argued that parents are an important resource in any childcare intervention (Reese, et al., 2010). The parent workshops were designed to enlist this resource in support of the development of French language and culture, and we considered a number of outcomes in gauging its success.

Robust effects were observed for self-reported Knowledge and Self-efficacy. The strength of these effects varied as a function of workshop quality measures (i.e., Practitioner's Style of Delivery and Session Length) and parental participation (for a similar finding, see Jordan, et al., 2000). The converging set of results for the workshop program that were generated in separate analyses by participation (Participant versus Non-participant) and by workshop program quality lends greater credibility to the program effect, firmly establishing the link between the parental workshops and change in the attitudes of those parents who received the intervention. Establishing this link is the first step in demonstrating that the workshops affected the development of children indirectly via the parents and the environment they created in the home.

Now that the workshop-parent link has been established, the research question could be extended to consider whether parental participation or parental change-score are significantly related to the developmental outcomes of children. Demonstrating such a link would complete the causal chain ending in positive effects on the outcomes of children. For now, we conclude that the link from workshop participation to parental attitudes has been established. We presume the link exists between parental attitudes and child development outcomes based on previous evaluations of the effectiveness of interventions, where the combination of both a child program and a literacy workshop component (Brooks-Gunn, et al., 2000; Reese, et al., 2010) yielded stronger effects than either component in isolation. A lack of statistical power precluded a formal test of the plausibility of the causal structure just described for the Readiness to Learn project sample. In subsequent reports, a formal mediation test will be conducted (e.g., Baron & Kenny, 1986) for the purpose of decomposing observed program effects according to their source, in this case the daycare programs and literacy workshops. Such analyses will be possible when data from the second cohort of children are inserted into the analyses[103].

Most of the remaining outcomes (i.e., Frequency of Literacy Activities, Modeling Behaviours, beliefs about child development) were distorted by experimental and measurement problems, which explains why no program effects were observed with these measures. The

[103] Presently, the sample of workshop participants from the four communities used in the impact analyses is less than 50.

implication is that a real change in literacy behaviours may have occurred with the associated benefits on child outcomes (Bus, van IJzendoorn, & Pelligrini, 1995; Scarborough & Dobrich, 1994). The equivocal results do not pronounce themselves one way or the other. A contributing factor to the null result may have been the fact that the topics presented during the workshops were too general. Literacy workshops generate consistent effects when they target the development of specific, concrete skills by way of a program that devotes more time to specific topics than the one offered by the Readiness to Learn project (Sénéchal, 2006; for a list of examples, see Reese, et al., 2010).

A positive effect relative to the Informal Care group was observed for the variable Language of Literacy Activities for the third post-test only. This effect was interesting in that it was not observed for the mid-workshop surveys, but only on the following survey. The result is consistent with the fact that the issue of language in literacy activities was covered towards the end of the series of workshops (i.e., workshop 6). The result would be easier to interpret as a program effect had a corresponding advantage for the Program Daycare group been observed relative to the Comparison Daycare group.

7.3. DIFFERENTIATED DOSAGE EFFECTS

The literature on daycares has begun addressing the issue of whether the effect of daycare exposure (e.g., hours per week spent in daycare) varies depending on the quality of the daycare program (Vandell, 2004). Indeed, there is some evidence that the high-quality daycares attenuate the negative behavioural effects associated with long hours in daycare (National Institute of Child health and Human Development Early Child Care Research Network, 2003). It is possible that quality has a similar moderating effect on the negative associations that are sometimes observed between dosage and cognition or language outcomes (Vandell, 2004). All stakeholders (i.e., parents, children, educators, evaluators) were aware of the program and its intended effects and who was being treated. This awareness could have contributed to the observed results in various ways. In contrast, stakeholders were all more or less blind to the variable “dosage” and its expected effects. Any effects that are observed with the dosage variable are therefore free of the more obvious forms of expectancy bias.

To the extent that the tested program resulted in a better quality daycare program, it is reasonable to ask whether the effect of daycare exposure varies as a function of whether children were enrolled in a daycare assigned to the program or comparison group. Indeed, analyses presented in Section 5.1.2 revealed that the difference between the two daycare groups on the Communication domain for the first post-test increased in direct proportion to hours spent in daycare. For this outcome at least, we observe dosage effects that are differentiated by treatment group, which lends additional credibility to the reported program impacts for this time period. In all other analyses (i.e., other outcomes in Year 1 and Year 2 analyses of all outcomes), the quantity of daycare exposure was not predictive of school-readiness outcomes *over and above what could already be explained by treatment group membership*.

7.4. DAYCARE FIDELITY AND QUALITY

All daycares vary with respect to the quality of the program they provide. The tested program was designed to address the specific needs of francophone families and children within the context of a high-quality program. We predicted that the daycares assigned to the new program would offer a higher quality program than those assigned to the comparison group. As we argued in Chapter 4, this division of daycares into two groups is a simplification of a continuous dimension that can be measured directly: program quality.

The better quality of instruction, the more likely it is that children exposed to the program will assimilate the targeted concepts (Howes, et al., 2008), in this case French language and culture. As a simplification, treatment group membership does not fully capture the variability in quality that exists across daycares, even though it explains much of it (see Table 5.7). For example, some comparison daycares provided a relatively high quality of service, while some program daycares implemented the intended program less effectively. Descriptive analyses indicated that both program and comparison daycares provided a quality of service that was in the upper range (*Project Implementation Report*, Bérubé et al., 2014). It could be argued on that basis that any change in quality caused by the intervention is superfluous. Such an argument would run counter to research indicating that the effect of quality on child development outcomes is non-linear; it is in the mid- to upper- range of quality that differences in daycare program quality have the strongest impact (Burchinal, Vandergrift, Pianta, & Mashburn, 2009).

Descriptive and inferential statistics confirmed that the program group daycares distinguished themselves positively on the dimensions of program fidelity and quality. Both Structural and Content Fidelity were higher in the Program Daycare group, but the most marked difference was observed for Structural Fidelity in this low-powered analysis. Among the quality indices considered here (Structural, Educative, Educator Sensitivity, and Reading), the largest advantages for the program over the comparison group daycares were observed for Structural Quality and Reading Quality. In sum, all quality and fidelity indices support the validity of our assumption that treatment group membership (i.e., program daycare versus comparison daycare) is an adequate stand-in for daycare program quality.

Despite this consistency, we nevertheless performed follow-up analyses which employed the fidelity and quality measures directly as predictors of the developmental outcomes of children. The purpose of these analyses was to contribute to the body of evidence in support of the construct validity of the study. Indeed, the results of these analyses tended to confirm the results of the analyses by treatment groups in Year 1 (Chapter 5). The most consistent effects were observed with *structural fidelity*, which seemingly contradicts the results of a study reporting that structural fidelity contributes little additional variance when a wide assortment of covariates are included in the model (Howes, et al., 2008). The same list of covariates was unavailable for the analyses presented here, which makes a definitive comparison impossible. We suggest that in the absence of adequate controls, the Structural Fidelity indices are capturing both structural fidelity and other concepts such as the level of “buy-in” within a given daycare. This explanation would account for the unexpected effect associated with Physical Ability for the third post-test, which was the only impact effect involving this dimension. Note also that Quality of Reading was strongly correlated with both Structural and Content Fidelity (see Table 5.5). In sum, these indices are not pure measures of the targeted constructs, but nonetheless support the study’s validity.

Content Fidelity was associated with the developmental outcomes of children as well. Content fidelity was associated specifically with Communication, Self-awareness, and Expressive Vocabulary. No association was observed with Cognitive Ability, which is consistent with the fact that the program content was geared specifically towards French-language outcomes rather than cognitive development per se.

More generally, *program quality effects* were observed for the Communication, Self-awareness, and Expressive Vocabulary scales. These associations with expressive-language skills are consistent with previous work with Anglophone children (Mashburn, et al., 2008). Interestingly, Structural Quality and Educative Quality were significantly related to Expressive Vocabulary, while Educative Quality and Educator Sensitivity were related to Communication. This mild dissociation supports the idea that educator sensitivity is mainly associated with the social aspects of communication (Peisner-Feinberg, et al., 2001), while various structural elements in the daycare are concerned with more concrete goals like building vocabulary. Educator sensitivity is important given the educator–child relationship is the primary conduit for delivering program resources to the children (Howes, et al., 2008). For instance, the negative effects associated with larger class sizes may be partially explained by the concomitant decreases in educator sensitivity that are observed in this context (Phillipsen, Burchinal, Howes, & Cryer, 1997). The impact of the tested program on this dimension suggests that developmental outcomes can be improved by targeting this dimension of classroom quality.

In the analysis of Year 2 outcomes, Global Fidelity and Quality of the daycare program (averages based on some of the more fine-grained indices discussed above) were predictive of development on the Cognitive Ability scale by midway through the second year. Given that this scale contains items that tap phonological awareness and alphabetic knowledge, we may conclude that program fidelity and program quality are predictive of important precursors to literacy acquisition and subsequent academic success (Belsky, et al., 2007). We attribute the association that was observed between Global Quality and Communication to a similar cause: the two phonological awareness items in this scale. Note that the two remaining items are shared with the Expressive Vocabulary subscale, which is unrelated to Global Quality. Finally, we observed an association between Global Quality and the normalized vocabulary measures of receptive and expressive vocabulary: ÉVIP–R and EOWPVT (for a similar finding, see Mashburn, et al., 2008). A similar association was not observed for Global Fidelity, suggesting that to be successful the program requires more than crossing off a list of basic requirements from a list. Standards of quality for the implementation of these requirements are also important. Taken together, the results support our contention that the tested program manipulates the correct levers for influencing developmental outcomes.

7.5. LIMITATIONS AND FUTURE WORK

Bronfenbrenner's (1979) ecological theory provides a rich framework for exploring various contingencies implicated in the success of a daycare program like the one evaluated here. In testing whether the magnitude of the program impacts varied according to language exposure, we have only begun to scratch the surface. Community-level measures of ethnolinguistic vitality may also contribute to moderating the program effects. The combination of low exposure to French and low vitality of the francophone community in a given location might conspire together to further enhance the benefits of the program. A fair examination of such complex

effects would require a larger sample than that available for the present report. Such questions could be examined more reasonably when data from the second cohort become available. Additional participants could also enable us to tease apart the independent contributions of the daycare and family literacy workshop components of the program using other types of data analysis strategies (e.g., mediation tests; path analyses).

A second limitation of the current study concerns the way school-readiness was measured. By their very nature, school-readiness tools such as the Early Development Instrument (EDI) and ÉPE–AD are not well calibrated for measuring program impacts. They are best suited for screening a population of children when the purpose is to identify those who are risk of developmental delay and who are in need of a "boost" to achieve a set of functional skills appropriate to their age range. All that is required to meet this objective is a general measure of development, which plays to the strengths of the EDI and the ÉPE–AD. However, this strength becomes a weakness when the purpose changes: evaluating the impact of a program. In this case, the research questions center on effects that can be more subtle (i.e., potentially small program effects rather than large delays or achievement gaps) and the specific skills impacted by the tested program are of interest. In short, a more precise picture of development is required, one that informs us not only about general development, but specifically on the skills (measured precisely) that best predict later outcomes (e.g., word acquisition and phonological knowledge inform us on later reading skills which in turn are a good predictor of school success; Hirsh-Pasek, et al., 2005).

A third limitation of the study concerns itself with external validity. The question of whether the reported program effects will be replicated in all other populations of francophone preschool children is limited by the very nature of the intervention. The tested program was implemented in francophone daycare centers, largely located in urban communities, and characterized by intermediate-to-high levels of ethnolinguistic vitality (except for Durham). Moreover, families participating in the project were Canadian, for the most part, with few representatives of newly immigrated Francophones (see Table 4.5 in Chapter 4). Future work is needed to determine the impacts of such a program for a rural francophone minority population, those living in highly minority settings, as well as the population of new francophone minority immigrants who in many cases claim multiple first languages.

Lastly, while extensive analyses were conducted to verify the robustness of results to various model specifications, the small sample size ($n < 100$ per treatment group) and small number of daycares ($n < 30$) mitigate our confidence in the findings. This is particularly true given that in the Readiness to Learn project, as with all quasi-experimental research, the possibility of an uncontrolled confound driving the results cannot be definitively ruled out. The overall recommendation of the present report is that re-analysis of the data including participants from both the first and second cohorts are necessary to confirm the stability of the estimates.

7.6. CONCLUSION

The impact analysis reported in this document revealed *positive program effects on both child and parent outcomes*. With respect to the child outcomes, the positive effects manifested themselves as superior developmental gains relative to the Comparison Daycare and Informal Care groups both at the first post-test and by midway through the second year of the project for

some children. The nature of these gains depended on the child's exposure to French at the start of the project. Children with the least exposure benefited most in terms of language development (e.g., expressive vocabulary), while those with the most exposure to French saw gains in their cognitive development. The component of the program directed at parents was successful in influencing some attitudinal dimensions. Parents reported gains in Knowledge and in Self-efficacy after participating in the workshops, which means that part of the observed child impacts may be attributable to the workshop component of the intervention.

We may conclude then that the tested program had a modest impact on the development of school readiness in minority-language francophone children. In forthcoming reports, the analyses will benefit from greater stability and statistical power by virtue of the addition of a second cohort of children to the sample. The focus of these forthcoming reports will be on whether apparent gains on dimensions of school-readiness and French language mastery will translate into benefits in the medium to long term. The first step in this line of investigation is to determine whether the program has had a positive effect on more immediate precursors to academic success, such as phonological awareness, alphabetical knowledge and vocabulary (Lonigan, 2008). Such is to be the object of the Readiness to Learn project in its second phase.

References

Abadie, A. (2005). Semiparametric Difference-in-Differences estimators. *Review of Economic Studies, 72*(1), 1–19.

Adams, G. R., & Ryan, B. A. (2000). *A longitudinal analysis of family relationships and children's school achievement in one- and two-parent families*. Ottawa, ON: Applied Research Branch, Strategic Policy, Human Resources Development Canada.

Arnold, D. H., Lonigan, C. J., Whitehurst, G. J., & Epstein, J. N. (1994). Accelerating language development through picture-book reading: Replication and extension to a videotape training format. *Journal of Educational Psychology*, *86*(2), 235–243.

Arnold, D. H., Zeljo, A., Doctoroff, G. L., & Ortiz, C. (2008). Parent involvement in preschool: Predictors and the relation of involvement to preliteracy development. *School Psychology Review*, *37*(1), 74–90.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Baylor College of Medicine. (2001). Guide for using the classroom observation tool.

Becker, G. S. (1993). Nobel lecture: The economic way of looking at behaviour. *The Journal of Political Economy*, *101*(3), 385–409.

Behrman, J. R., Cheng, Y., & Todd, P. E. (2004). Evaluating preschool programs when length of exposure to the program varies: A nonparametric approach. *The Review of Economics and Statistics*, *86*(1), 108–132.

Belsky, J., Vandell, D. L., Burchinal, M., Clarke-Stewart, K. A., McCartney, K., Owen, M. T., & The NICD Early Child Care Research Network. (2007). Are there long-term effects of early child care? *Child Development*, *78*(2), 681–701.

Benoît, S. (n.d.). *Questionnaire utilisé par CIRCEM dans l'étude de l'impact de l'alphabétisation familiale sur les familles vivant en milieu minoritaire*. Vanier, ON: Coalition francophone pour l'alphabétisation et la formation de base en Ontario.

Berger, L. M., Paxson, C., & Waldfogel, J. (2005). *Income and child development*. Center for Research on Child Wellbeing, Princeton, NJ: Princeton University.

Bertrand, J. (2007). Programmes d'enseignement efficaces au préscolaire. Commentaires sur Kagan et Kauerz et sur Schweinhart. In R. E. Tremblay, R. G. Barr, & R. de V. Peters (Eds.), *Encyclopédie sur le développement des jeunes enfants* [online] (pp. 1-7). Montréal, QC: Centre d'excellence pour le développement des jeunes enfants. Available online at: http://www.child-encyclopedia.com/pages/PDF/BertrandFRxp.pdf.

Bertrand, M., Duflos, E., & Mulliainathan, S. (2004). How much should we trust Difference-in-Differences estimates? *The Quarterly Journal of Economics*, *119*(1), 249–275.

Bérubé, A., Legault, L., Janisse, L., Carson, R., Saucier, M.-A., & Lefebvre, L. (2014). *Readiness to Learn in Minority Francophone Communities: Project Implementation Report*. Ottawa, ON: Social Research and Demonstration Corporation.

Bloom, B. S. (1976). *Human characteristics and school learning*. New York, NY: McGraw-Hill.

Bornstein, L., & Bornstein, M. H. (Eds.). (2007). Pratiques parentales et développement social de l'enfant. In R. E. Tremblay, R de V. Peters, M. Boivin, & R. G. Barr (Eds.). *Encyclopédie sur le développement des jeunes enfants* [online] (pp. 1-5). Montréal, QC : Centre d'excellence pour le développement des jeunes enfants. Available online at: http://www.enfant-encyclopedie.com/Pages/PDF/BornsteinFRxp.pdf.

Breton, R. (1964). Institutional completeness of ethnic communities and the personal relations of immigrants. *The American Journal of Sociology*, *70*(2), 193–205.

Brigham, J. C. (1991). Attitudes and persuasion. *Social psychology* (2nd ed., pp. 132–224). New York, NY: HarperCollins.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.

Bronfenbrenner, U., & Morris, P. A. (1998). The ecology of developmental processes. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development* (5th ed., Vol. 1, pp. 993–1028). New York, NY: Wiley.

Brooks-Gunn, J., Berlin, L. J., & Fuligni, A. S. (2000). Early childhood intervention programs: What about the family? In J. P. Shonkoff & S. J. Meisels (Eds.), *Handbook of early childhood intervention* (2nd ed., pp. 549-588). New York, NY: Cambridge University Press.

Brooks-Gunn, J., & Duncan, G. J. (1997). The effects of poverty on children. *The Future of Children*, *7*(2), 55–71.

Burchinal, M. R., Roberts, J. E., Riggins, R., Jr., Zeisel, S. A., Neebe, E., & Bryant, D. (2000). Relating quality of center-based child care to early cognitive and language development longitudinally. *Child Development*, *71*(2), 339–357.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A., (2009). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*. *25*, 166-176.

Bus, A. G., van IJzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research*, *65*(1), 1–21.

Bussière, P., Cartwright, F., Crocker, R., Ma, X., Oderkirk, J., & Zhang, X. (2001). *À la hauteur: la performance des jeunes du Canada en lecture, en mathématiques et en sciences*. Ottawa, ON: Human Resources Development Canada, Statistics Canada, Council of Ministers of Education (Canada).

Canadian Council on Learning. (2008). *Literacy among Canadian students in minority language contexts*. Unpublished report.

Centre for Family Literacy. (2002). *Statements of good practice for family literacy programs in Alberta*. Edmonton, AB: Author.

Charlebois, P., Brendgen, M., Vitaro, F., Normandeau, S., & Bourdreau, J.-F. (2004). Examining dosage effects on prevention outcomes: Results from a multi-modal longitudinal preventive intervention for young disruptive boy. *Journal of School Psychology, 42*(3), 201–220.

Chartier, M. J., Dumaine, J., Daudet-Mitchell, L., Gosselin, J., & Vielfaure, K. (2008). Vivre en français à la petite enfance et apprendre à l'école française en milieu minoritaire, y a-t-il un lien? *Enfants en santé Manitoba, Division scolaire franco-manitobaine*. Document submitted for publication.

Cleveland, G., Corter, C., Pelletier, J., Colley, S., Bertrand, J., Ontario Institute for Studies in Education/University of Toronto, & Jamieson, J. (2006). *A review of the state of the field of early childhood learning and development in child care, kindergarten and family support programs*. Toronto, ON: Atkinson Centre for Society and Child Development.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioural sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Commission nationale des parents francophones. (2005*). Le meilleur est...avenir! La petite enfance dans les communautés francophones en milieu minoritaire au Canada : une analyse transformative de la situation*. Ottawa, ON. Consulted February 2, 2009. Available online at: http://cnpf.ca/documents/SCAN_FRANCAIS_2005.pdf.

Conduct Problems Prevention Research Group. (1999). Initial impact of the Fast Track prevention trial for conduct problems: I. The high-risk sample. *Journal of Consulting and Clinical Psychology*, *67*(5), 631–647.

Connor, S., & Brink, S. (1999). *Understanding the Early Years: Community impacts on child development*. Ottawa, ON: Applied Research Branch, Strategic Policy, Human Resources Development Canada. Available online at: http://www.hrsdc.gc.ca/eng/cs/sp/sdc/pkrf/publications/nlscy/uey/1999-000092/w-99-6e.pdf.

Connors, L. J., & Epstein, J. L. (1995). Parent and school partnerships. In M. H. Bornstein (Ed.), *Handbook of parenting. Volume 4: Applied and practical parenting* (pp. 437–458). Mahwah, NJ: Lawrence Erlbaum Associates.

Conseil canadien sur l'apprentissage. (2007). *Rapport sur l'apprentissage au Canada : rapport sur l'état de l'apprentissage chez les jeunes enfants*. Ottawa, ON.

Corbeil, J.-P., & Blaser, C. (2007). *Le portrait linguistique en évolution : Recensement de 2006*. Ottawa, ON: Statistics Canada. Available online at: http://www12.statcan.ca/census-recensement/2006/as-sa/97-555/pdf/97-555-XIE2006001-fra.pdf.

Corbeil, J.-P., Grenier, C., & Lafrenière, S. (2007). *Les minorités prennent la parole : résultats de l'Enquête sur la vitalité des minorités de langue officielle 2006*. Consulted

December 11, 2007. Available online at: http://www.statcan.gc.ca/pub/91-548-x/91-548-x2007001-fra.pdf.

Corter, C., & Pelletier, J. (2005). Parents and community involvement in schools: Policy panacea or pandemic? In N. Bascia, A. Cumming, A. Datnow, K. Leithwood, & D. Livingstone (Eds.), *International Handbook of Educational Policy* (pp. 295–327). Great Britain: Springer.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change" – or should we? *Psychological Bulletin*, *74*(1), 68–80.

Cummins, J. (1979). *Cognitive academic language proficiency, linguistic interdependence, the optimum age question and some other matters*. Working Papers in Bilingualism, No. 19, 12–129.

Cunha, F., Heckman, J. J., Lochner, L., & Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (1st ed., Vol. 1, pp. 307–451). Amsterdam: North-Holland: Elsevier.

D'Amours, Y. (2010). *La scolarité des francophones et des anglophones, à travers les groupes d'âge, au Québec et en Ontario: données sociodémographiques en bref*. Québec, QC : Available online at: http://www.stat.gouv.qc.ca/publications/conditions/pdf2010/sociodemobref_fev10.pdf.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*(1), 23–45.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, *19*(2), 294–304.

Dearing, E., Berry, D., & Zaslow, M. (2006). Poverty during early childhood. In K. McCartney & D. Phillips (Eds.), *The Blackwell handbook of early childhood development* (pp. 399–423). Malden, MA: Blackwell Publishing.

Desrosiers, H., & Ducharme, A. (2006). *Commencer l'école du bon pied : facteurs associés à l'acquisition du vocabulaire à la fin de la maternelle*. Montreal, QC: Institut de la statistique Québec. Available online at: http://www.jesuisjeserai.stat.gouv.qc.ca/pdf/publications/feuillet/fascicule_ecole_bon_pied.pdf.

Dionne, A.-M. (2007). *Incidence des pratiques de littératie familiale sur les compétences en lecture de l'enfant dans les familles immigrantes*. Presented at the Actualité de la Recherche en Éducation et en Formation, Strasbourg, France.

Doherty, G. (1997). *Zero to six – The basis for school readiness*. Hull, QC: Applied Research Branch of Strategic Policy, Human Resources Development Canada. Available online at: http://www.hrsdc.gc.ca/eng/cs/sp/sdc/pkrf/publications/1997-002557/SP-241-02-01E.pdf.

Donald, S. G., & Lang, K. (2007). Inference with Difference-in-Differences and other panel data. *The Review of Economics and Statistics*, *89*(2), 221–233.

Dunn, L. M., Thériault-Whalen, C. M., & Dunn, L. M. (1993). *Échelle de Vocabulaire en Images Peabody, adaptation française du Peabody Picture Vocabulary Test-Revised: Manuel pour les Formes A et B* (p. 147): Pearson Canada Assessment Inc.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3–4), 327–350.

Education Quality and Accountability Office. (2009). *Ontario student achievement: EQAO's provincial report on the results of the 2008-2009 assessments of reading, writing and mathematics, primary division (grades 1-3) and junior division (grades 4-6), and the grade 9 assessment of mathematics. English-language students*. Toronto, ON. Disponible en ligne au http://www.eqao.com/pdf_e/09/Cprr_PJ9e_0909_web.pdf.

Engle, P. L., Black, M. M., Behrman, J. R., Cabral de Mello, M., Gertler, P., Kapiriri, L., ... International Child Development Steering Group. (2007). Child development in developing countries 3: Strategies to avoid the loss of developmental potential in more than 200 million children in the developing world. *The Lancet*, *369*, 229-242. Consulted March 23, 2010. Available online at: http://www.who.int/child_adolescent_health/documents/pdfs/lancet_child_dev_series_paper3.pdf.

Erikson, E. H. (1994). *Indentity and the life cycle*. New York, NY: W. W. Norton & Company.

Fédération canadienne pour l'alphabétisation en français. (2007). *Apprendre ça commence à la maison : guide de pratiques exemplaires en alphabétisation familiale en contexte francophone minoritaire*. Ottawa, ON: Author.

Forgues, É., & Landry, R. (2006). *Définitions de la francophonie en situation minoritaire : analyse de différentes définitions statistiques et leurs conséquences*. Ottawa, ON: Institut Canadien de recherche sur les minorités linguistiques.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, *51*(1), 115–134.

Gardner, M. (1979). *Expressive One-Word Picture Vocabulary Test: Manual and forms*. Novato, CA: Academic Therapy Publications.

Gauthier St-Onge, D. (s.d.). Questionnaire utilisés dans l'Évaluation du programme alpha District 3, Edmundston, NB.

Gilbert, A. (2003). *Une vision nationale : la petite enfance : porte d'entrée à l'école de langue française : Rapport final de la recherche*. Ottawa, ON: Centre for Interdisciplinary Research on Citizenship and Minorities, University of Ottawa in partnership with the Canadian Teachers' Federation.

Glymour, M. M., Weuve, J., Berkman, L. F., Kawachi, I., & Robins, J. M. (2005). When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *American Journal of Epidemiology*, *162*(3), 267–278.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioural Research*, *26*(3), 499–510.

Guimond, L. (2003). *Revue des travaux sur la vitalité des communautés francophones du Canada*. Ottawa, ON: University of Ottawa.

Hargrave, A.C., & Sénéchal, M. (2000). A book reading intervention with preschool children who have limited vocabularies: The benefits of regular reading and dialogic reading. *Early Childhood Research Quarterly*, *15*(1), 75–90.

Harms, T., Clifford, R. M., & Cryier, D. (1998). *Early Childhood Environment Rating Scale: Revised Edition (ECERS–R)*. New York, NY: Teachers College Press.

Haveman, R., & Wolfe, B. (1995). The determinants of children's attainments: A review of methods and findings. *Journal of Economic Literature*, *33*(4), 1829–1878.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behaviour Research Methods*, *39*(4), 709–722.

Hertzman, C., & Kohen, D. (2003). Neighbourhoods matter for child development. *Transition*, *33*(3), 3-5.

Hindman, A.H., Skibbe, L.E., Miller, A., & Zimmerman, M. (2010). Ecological contexts and early learning: Contributions of child, family, and classroom factors during Head Start, to literacy and mathematics growth through first grade. *Early Childhood Research Quarterly*, *25*(2), 235–250.

Hirsh-Pasek, K., Kochanoff, A., Newcombe, N. S., & de Villiers, J. (2005). Using scientific knowledge to inform preschool assessment: Making the case for "empirical validity." *Social Policy Report*, *19*(1), 1–20.

Horizon Research. (2000). Inside the classroom: Observation and analytic protocol. Consulted January 2008. Retrieved from: http://www.horizon-research.com/instruments/clas/cop.pdf.

Howard, G. S. (1980). Response-shift bias: A problem in evaluating interventions with pre/post self-reports. *Evaluation Review*, *4*(1), 93–106.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, *23*(1), 27–50.

Hox, J. (2002). *Multilevel analysis: Techniques and applications* (1st ed.). New York, NY: Routledge.

Human Resources and Skills Development Canada. (2006). *Projet pilote de garde d'enfants : principes directeurs du projet pilote, 2006*. Available online at: http://www.hrsdc.gc.ca/fra/sm/comm/ds/garde_enfants/2006/principes_directeurs.shtml.

Human Resources and Social Development Canada (HRSDC), & Statistics Canada. (2005). *Building on our competencies: Canadian results of the International Adult Literacy and Skills Survey*. Ottawa, ON: Statistics Canada.

Institut de la statistique Québec. (2003). *En 2002…j'avais 5 ans! Questionnaire informatisé rempli par l'intervieweure (QIRI) – E6*. Québec, QC.

Izzo, C. V., Weissberg, R. P., Kasprow, W. J., & Fendrich, M. (1999). A longitudinal assessment of teacher perceptions of parent involvement in children's education and school performance. *American Journal of Community Psychology, 27*(6), 817–839.

Janus, M., & Offord, D. (2000). Readiness to learn at school. *ISUMA: Canadian Journal of Policy Research, 1*(2), 71–75.

Japel, C, Tremblay, R. E., & Côté, S. (2005). La qualité, ça compte ! Résultats de l'Étude longitudinale du développement des enfants du Québec concernant la qualité des services de garde. *Choix IRPP, 11*(4), 1–46.

Jordan, G. E., Snow, C. E., & Porche, M. V. (2000). Project EASE: The effect of a family literacy project on kindergarten students' early literacy skills. *Reading Research Quarterly, 35*(4), 524–546.

Kézdi, G. (2004). Robust standard error estimation in fixed-effects panel models. *Hungarian Statistical Review, Special number 9*, 95–116.

Klebanov, P. K., Brooks-Gunn, J., & Duncan, G. J. (1994). Does neighbourhood and family poverty affect mother's parenting, mental health, and social support? *Journal of Marriage and Family, 56*(2), 441–455.

Lafrance, F. (1993). Les conditions sociolinguistiques de l'enseignement du français minoritaire et leurs conséquences sur la pédagogie du français langue maternelle en Ontario. *Revue du Nouvel-Ontario, 15*, 11–34.

Lamb, T., & Tschillard, R. (2003). *An underutilization design in applied research: The retrospective pretest*. Paper presented at the Annual meeting of the American Sociological Association. Available online at: http://www.allacademic.com/meta/p_mla_apa_research_citation/1/0/8/1/9/p108194_index.html.

Lamb, T. A., & Tschillard, R. (2005). Evaluating learning in professional development workshops: Using the retrospective pretest. *Journal of Research in Professional Learning*. Retrieved from: http://www.nsdc.org/library/publications/research/lamb.pdf.

Landry, R. (1994). Diagnostic sur la vitalité de la communauté acadienne du Nouveau-Brunswick. *Égalité, 36*, 11–39.

Landry, R. (2010). *Petite enfance et autonomie culturelle : Là ou le nombre le justifie… V.* Research report for the Commission nationale des parents francophones. Canadian Institute for Research on Linguistic Minorities. Moncton, NB.

Landry, R., & Allard, R. (1997). L'exogamie et le maintien de deux langues et de deux cultures : le rôle de la francité familioscolaire. *Revue des sciences de l'éducation, 23*(3), 561–592.

Landry, R., & Allard, R. (2000). Langue de scolarisation et développement bilingue : Le cas des acadiens et francophones de la Nouvelle-Écosse, Canada. *DiversCité Langues, V.* Retrieved from: http://www.teluq.uquebec.ca/diverscite/entree.htm.

Landry, R., Allard, R., & Deveau, K. (2007a). A macroscopic intergroup approach to the study of ethnolinguistic development. *International Journal of the Sociology of Language, 185*, 225–253.

Landry, R., Allard, R., & Deveau, K. (2007b). Bilingual schooling of the Canadian Francophone minority: a cultural autonomy model. *International Journal of the Sociology of Language*, *185*, 133–162.

Landry, R., Allard, R., & Deveau, K. (2009). Self-determination and bilingualism. *Theory and Research in Education*, *7*(2), 203–213.

Legault, L., Mák, G., Verstraete, M., & Bérubé, A. (2014). *Readiness to Learn in Minority Francophone Communities: Reference Report*. Ottawa, ON: Social Research and Demonstration Corporation.

Lemelin, J.-P., & Boivin, M. (2007). *Mieux réussir dès la première année : l'importance de la préparation à l'école*. Quebec, QC: Institut de la statistique Québec. Available online at: http://www.jesuisjeserai.stat.gouv.qc.ca/pdf/publications/feuillet/Fasc2Vol4.pdf.

LeTouzé, S. (2006). *Pour mon enfant d'abord : étude de l'impact de l'alphabétisation familiale sur les familles vivant en milieu minoritaire. Étape 2 : Cohortes 2 et 3 : 2005-2006*. Ottawa, ON: Centre for Interdisciplinary Research on Citizenship and Minorities.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: John Wiley & Sons.

Lonigan, C. J. (2008). Effective strategies for promoting young children's literacy skills. Paper presented at the Effective Early Learning Programs: Research, Policy and Practice, Banff, AB.

Lonigan, C. J., Shanahan, T., Westberg, L., & The National Early Literacy Panel. (2008). Identification of children's skills and abilities linked to later outcomes in reading, writing, and spelling. In *Developing early literacy: Report of the National Early Literacy Panel*. (pp. 55–101). Jessup, MD: National Institute for Literacy. Available online at: http://www.nifl.gov/publications/pdf/NELPReport09.pdf.

Maltais, C. (2007). Relation entre la littératie familiale en milieu francophone et l'incidence d'un programme de maternelle quatre ans à temps plein sur le développement du langage et de la lecture des enfants. *Éducation francophone en milieu minoritaire*, *33*(2), 6–18.

Martel, A. (2001). *Droits, écoles et communautés en milieu minoritaire : 1986 - 2002. Analyse pour un aménagement du français par l'éducation*. Ottawa, ON : Office of the Commissioner of Official Languages.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., ... Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749.

Masny, D. (2006). Le développement de l'écrit en milieu de langue minoritaire : l'apport de la communication orale et des habiletés métalinguistiques. *Éducation et francophonie*, *34*(2), 126–149.

Meisels, S. J., & Atkins-Burnett, S. (2006). Evaluating early childhood assessments: A differential analysis. In K. McCartney & D. Phillips (Eds.), *The Blackwell handbook of early childhood development* (pp. 533-549). Malden, MA: Blackwell Publishing.

Merriam, S. B. (1988). *Case study research in education: A qualitative approach*. San Francisco, CA: Jossey-Bass.

Monheit, A. C., & Steinburg Schone, B. (2004). How has small group market reform affected employee health insurance coverage? *Journal of Public Economics, 88*(1–2), 237–254.

Moore, T. (2005). *Towards improved support for families of young children: The role of communities*. Paper presented at the 9th Australian Institute of Family Studies Conference. Available online at: http://aifs.gov.au/institute/afrc9/moore1.pdf.

Moore, D., & Tananis, C. A. (2009). Measuring change in a short-term educational program using a retrospective pretest design. *American Journal of Evaluation*, *30*(2), 189–202.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics, 72*(2), 334–338.

National Institute of Child Health and Human Development Early Child Care Research Network, & Duncan, G. J. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development*, *74*(5), 1454–1475.

National Institute of Child Health and Human Development Early Child Care Research Network. (2003). Does amount of time spent in child care predict socioemotional adjustment during the transition to kindergarten. *Child Development*, *74*(4), 976–1005.

Neuman, S.D., & Celano, D. (2001). Access to print in low-income and middle-income communities: An ecological study of four neighborhoods. *Reading Research Quarterly, 36*(1), 8-26.

Office de la qualité et de la responsabilité en éducation. (2009). *Rendement des élèves de l'Ontario : rapport provincial de l'OQRE sur les résultats des tests en lecture, écriture et mathématiques, cycle primaire (de la 1ère à la 3e année) et cycle moyen (de la 4e à la 6e année) et du test de mathématiques, 9e année, 2008–2009. Élèves de langue française*. Toronto, ON.

Pearson, B. (2007). Social factors in childhood bilingualism in the United States. *Applied Psycholinguistics*, *28*(3), 399–410.

Pearson, B. Z., Fernandez, S., Lewedag, V., & Oller, D. K. (1997). Input factors in lexical learning of bilingual infants (ages 10 to 30 months). *Applied Psycholinguistics*, *18*(1), 41–58.

Peisner-Feinberg, E. S., Burchinal, M.R., Clifford, R.M, Culkin, M.L., Howes, C., Kagan, S.L., Yazejian, N. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development, 72*(5), 1534–1553.

Pelletier, J., & Corter, C. (2005). Design, implementation, and outcomes of a school readiness program for diverse families. *The School Community Journal*, *17*(3), 89–116.

Phillips, L. M., Hayden, R., & Norris, S. P. (2006). *Family literacy matters: A longitudinal parent-child interaction study*. Calgary, AB: Temeron Books.

Phillipsen, L. C., Burchinal, M. R., Howes, C., & Cryer, D. (1997). The prediction of process quality from structural features of child care. *Early Childhood Research Quarterly*, *12*(3), 281–303.

Reese, E., Sparks, A., & Leyva, D. (2010). A review of parent interventions for preschool children's language and emergent literacy. *Journal of Early Childhood Literacy*, *10*(1), 97–117.

Rockwell, S. K., & Kohn, H. (1989). Post-the-pre evaluation. *Journal of Extension*, *27*(2). Available online at: http://www.joe.org/joe/1989summer/a5.php.

Rogers, K. N., Fernandez, M., Thurber, L., & Smitley, A. (2004). Exploring differential attrition rates among system of care evaluation participants. *Journal of Community Psychology*, *32*(2), 167–176.

Sanders, M. R., & Morawska, A. (2006). Peut-on améliorer les résultats des enfants en modifiant les connaissances des parents, leurs attentes dysfonctionnelles et la régulation de leurs émotions? In R. E. Tremblay, R. de V. Peters, M. Boivin et R. G. Barr (Eds.). *Encyclopédie sur le développement des jeunes enfants* [online] (pp. 1-14). Montréal, QC : Centre d'excellence pour le développement des jeunes enfants. Available online at: http://www.enfant-encyclopedie.com/Pages/PDF/Sanders-MorawskaFRxp.pdf.

Scarborough, H. S., & Dobrich, W. (1994). On the efficacy of reading to preschoolers. *Developmental Review*, *14*(3), 245–302.

Sénéchal, M. (2005). Alphabétisation, langue et développement affectif. In R. E. Tremblay, R. G. Barr, R. de V. Peters (Eds.). *Encyclopédie sur le développement des jeunes enfants* [online] (pp. 1-6). Montréal, QC : Centre d'excellence pour le développement des jeunes enfants. Available online at: http://www.enfant-encyclopedie.com/Pages/PDF/SenechalFRxp.pdf.

Sénéchal. M. (2006). Testing the home literacy model: Parent involvement in kindergarten is differentially related to grade 4 reading comprehension, fluency, spelling, and reading for pleasure. *Scientific Studies of Reading*, *10* (1), 59–87.

Sénéchal, M., & LeFevre, J.-A. (2002). Parental involvement in the development of children's reading skills: A five-year longitudinal study. *Child Development*, *73*(2), 445–460.

Shrigley, R. L. (1990). Attitude and behavior are correlates. *Journal of Research in Science Teaching*, *27*(2), 97–113.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Statistics Canada. (2004). *Revue trimestrielle de l'éducation*. *9*(4). Available online at: http://www.statcan.gc.ca/pub/81-003-x/81-003-x2002004-fra.pdf.

Statistics Canada. (2006). *Enquête longitudinale nationale sur les enfants et les jeunes : Matériel d'enquête pour la collecte des données 2004-2005. Cycle 6. Livre 1 : Parent, enfant et jeune*.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.

Tamis-LeMonda, C. S., & Rodriguez, E. T. (2008). Rôle des parents pour favoriser l'apprentissage et l'acquisition du langage chez les jeunes enfants. In R. E. Tremblay, R. G. Barr, R. de V. Peters, M. Boivin (Eds.), *Encyclopédie sur le développement des jeunes enfants* [online] (pp.1-11). Montreal: Centre d'excellence pour le développement des jeunes enfants. Available online at: http://www.ccl-cca.ca/pdfs/ECLKC/encyclopedia/TamisLemondaRodriguezFRxpCSAJELangage.pdf.

Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Gilding the outcome by tarnishing the past: Inflationary biases in retrospective pretests. *American Journal of Evaluation, 30*(1), 31–43.

Thomas, E. M. (2006). *La disposition à apprendre à l'école pour les jeunes de cinq ans au Canada*. Consulted February 17, 2008. Available online at: http://www.statcan.gc.ca/pub/89-599-m/89-599-m2006004-fra.pdf.

Tremblay, R. E., Nagin, D. S., Séguin, P. D., Boivin, M., Pérusse, D., & Japel, C. (2004). Physical aggression during early childhood: Trajectories and predictors. *Pediatrics, 114*(1), e43–e50.

UNESCO. (2010). *Enhancing learning of children from diverse language backgrounds: Mother tongue-based bilingual or multilingual education in early childhood and early primary school years*. Available online at: http://www.ecdip.org/docs/pdf/UNESCO%20Mother-tongue%20based%20EY%202010.pdf.

Vandell, D. L. (2004). Early child care: The known and the unknown. *Merrill Palmer Quarterly Journal of Developmental Psychology, 50*(3), 387–414.

Venetsanou, F., & Kambas, A. (2010). Environmental factors affecting preschoolers' motor development. *Early Childhood Education Journal, 37*, 319–327.

Vihman, M. M., Lum, J., Tierry, G., Nakai, S., & Keren-Portnoy, T. (2006). The onset of word form recognition in one language and in two. In P. McCardle & E. Hoff (Eds.), *Childhood bilingualism: Research on infancy through school age* (pp. 30–44). Clevedon, UK: Multilingual Matters.

Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wasik, B. (2008). When fewer is more: Small groups in early childhood classrooms. *Early Childhood Education Journal, 35*(6), 515–521.

Wasik, B.A., Bond, M.A., & Hindman, A. (2006). The effect of a language and literacy intervention on Head Start children and teachers. *Journal of Educational Psychology, 98*(1), 63-74.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817–838.

Whitehurst, G. J., Falco, F. L., Lonigan, C., Fischel, J. E., DeBaryshe, B. D., Valdez-Menchaca M. C., & Caulfield, M. (1988). Accelerating language development through picture book reading. *Developmental Psychology, 24*(4), 552–558.

Williams, R. L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics, 56*(2), 645–646.

Willms, J. D. (2007). *Child Care Pilot Project: Final report on direct assessment measures and parent and community questionnaires*. Fredericton, NB: KSI Research International Inc.

Willms, J. D. (2009). *Évaluation Petite Enfance : Appréciation Directe, version allongée*. Fredericton, NB : KSI Research International Inc.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25*, 659–707.

Wise, J. C., Sevcik, R. A., Morris, R. D., Lovett, M. W., & Wolf, M. (2007). The relationship among receptive and expressive vocabulary, listening comprehension, pre-reading skills, word identification skills, and reading comprehension by children with reading disabilities. *Journal of Speech, Language, and Hearing Research, 50*, 1093–1109.

Appendix A: Evaluation Tools and Timetable

The timetable for the evaluation of children is presented in the table below. For each data collection wave, we report the construct measured (*in italics*) and the direct-assessment tool that was used to capture its development. We refer to the EYE-DA and the PPVT–R by their French appellations: the ÉPE–AD and ÉVIP–R respectively.

Table A1: Timetable for Child Evaluations

October 2007 (Baseline)	February 2008 (1st Post-Test)	June 2008 (2nd Post-Test)	October 2008 (3rd Post-Test)	February 2009 (4th Post-Test)	June 2009 (5th Post-Test)	October 2009 (6th Post-Test)
School Readiness						
ÉPE–AD[104]: Self-awareness (A)						
ÉPE–AD: Cognitive Ability (B)						
ÉPE–AD: Language and Communication (C)						
ÉPE–AD: Physical Ability and Motor Skills (D)						
Expressive Vocabulary						
ÉPE–AD subscale						
					EOWPVT–R	
Receptive Vocabulary						
ÉPE–AD subscale						
						ÉVIP–R

[104] Évaluation de la petite enfance — Appréciation directe (Willms, 2007).

Appendix B: Procedure for Administering the ÉPE–AD (Pre-intervention Measure)

The evaluators who administered the *ÉPE–AD* (the EYE–DA in English) to children were recruited starting in late summer 2007. SRDC provided evaluators with theoretical and practical training lasting approximately six hours in August, September and, for the community of Orléans, in October. In addition to presenting the test administration protocol, training provided an introduction to the Readiness to Learn project and procedures relating to confidentiality. They signed a contract whereby they agreed to adhere to the administration and confidentiality protocol. The complete steps of the protocol for test administration are as follows:

1. The evaluators call parents to schedule appointments for assessments at home or to notify them when the assessment will be conducted in daycares. The purpose of these calls is also to confirm the child's age in months and the answers to the questions on languages spoken with parents and friends for domain E (*Awareness and involvement in francophone culture).*
2. The evaluators and community coordinator get in touch with the participating daycares to define the schedule and arrange a place in the classroom that is favourable for a good assessment.
3. The evaluator applies the "medical" method, that is, she waits to check that the child is the right one before completing the identifying information on the paper questionnaire.
4. The evaluator addresses the child in his or her mother tongue first, then applies the protocol for determining the test language.
5. The evaluator follows the tool developer's scoring instructions, that is, she rounds the score to the lower whole number for the purpose of conducting a prudent assessment.
6. The evaluator encourages the child, but does not give any hints as to how to answer, unless the protocol indicates to do so.
7. If the child gets tired during the test, the evaluator stops and can start again later at the start of the domain where she left off.
8. At the end of the assessment, the evaluator gives the child a sticker to thank him or her for participating.
9. If the child really does not want to participate, the evaluator must try to assess the child at least one more time (another day).

Point 4 is definitely key to the test, as adherence or failure to adhere to this rule may substantially influence the results. The administration protocol for determining the test language suggested by Mr. Willms (presented on December 8, 2006, and revised in July 2007) is outlined in Table B.1.

Table B1: Decisional Tree for Determining the Language of Administration

Steps	***Protocol***
1. Administration of domain E	Administer the six questions for domain E directly to the child, and to the parent for questions E4 to E6, at the start of the test.
2. Decisional tree for determining the test language (domain E)	If the score is greater than 6 on domain E, administer the rest of the ÉPE–AD in French. If the score is less than or equal to 6 on domain E, assess domain C in French and English.
3. Decisional tree for determining the test language (domain C)	If the score in French for this domain is greater than 14, the rest of the assessment is conducted in French. If the score in French for this domain is less than or equal to 14 and the score in English is less than or equal to 14, the rest of the test is conducted in French also. If the score in domain C is less than or equal to 14 in French, but the score is greater than 14 in English, the rest of the assessment is conducted in English.

In addition, there are two starting points for the test depending on the child's age. If the child is less than four years old, the evaluator starts with the first item for the domain. Otherwise she starts further on in the test with the option of going back to the starting point if the child is struggling. This decision rule was not mentioned in the *Reference report* (Legault et al., 2014) because at the time of the baseline measurement all children were less than four years of age.

Appendix C:
Comparing Different Versions of the ÉPE–AD

This section concerns the content of three version of the ÉPE–AD (the EYE–DA in English) employed during the first two years of implementation of the Readiness to Learn project. New versions of the test were created in response to two issues noted in the fall of 2008 and following decisions taken in the wake of a meeting with the Consultative Committee of the HRSDC on December 17th, 2008. The first issue concerned the potential for a ceiling effect based on projected scores for the 5th evaluation (February 2009) derived from the performance of children on the 3rd and 4th evaluations. The test designer was hired by the HRSDC to create new items for the test that were more difficult, thereby allowing the developmental trajectory of children to be followed on Domains A, B, and C. The second issue concerned the need for a measure of language ability that makes finer discriminations among children. Though the ÉPE–AD measures school readiness well, it was not sensitive enough to capture the developmental dimension of children and only scratches the surface of their ability to communicate.

For the purpose of the discussion to follow, we give the name 'initial ÉPE–AD' to the version of the test administered in February of 2009, the name 'extended ÉPE–AD' to the version of the test administered in the winter of 2009 (containing the more difficult questions developed by Willms, and excluding the easiest questions), and finally the name 'modified ÉPE–AD' to the version of the test that was reworked by SRDC and used for the evaluations from June to October 2009.

Domains measured by the initial ÉPE–AD

The ÉPE–AD conceived by Doug Willms comprises four domains plus a fifth conceived specifically for the Readiness to Learn project:

- Domain A = Self-awareness;
- Domain B = Cognitive Ability;
- Domain C = Language and Communication;
- Domain D = Physical and Motor Skills; and
- Domain E = Awareness and Engagement in Francophone Culture.

Domains measured by the modified ÉPE–AD

The modified ÉPE–AD comprises seven dimensions, namely:

- Expressive Vocabulary;
- Self-awareness;
- Phonological Awareness;
- Numeracy;
- Memory for Personal Information;
- Alphabetic Knowledge; and
- Oral Reading.

These scales allow the trajectories of children to be pursued for the domains A, B, and C. The modified ÉPE–AD contains all items from Domain A (minus question A18). As for Domain B, all items were kept, either in the Phonological Awareness subscale, the Numeracy subscale, or the Alphabetic Knowledge subscale. Finally, Domain C is partially preserved within the Expressive Vocabulary subscale, with which it correlates at .87, confirming that it will be possible to continue tracking the trajectory of this domain.

The tables below serve to compare the items used in the three versions of the ÉPE–AD. The modified version of the ÉPE–AD was conceived in French only; the items listed in the following tables are taken verbatim from that evaluation. Note that items which were not retained for the modified ÉPE–AD are not presented in the table. The list of excluded items includes those that were eliminated by Willms for the extended ÉPE–AD.

Table C1: Expressive Vocabulary Subscale

Items	Initial ÉPE–AD	Extended ÉPE–AD	Modified ÉPE–AD
Expressive Vocabulary			
Comment s'appelle cet objet? ***un seau/une chaudière, des boutons, des chandelles, une horloge/un réveille-matin***	**A11**	**x**	**V1**
Pointez chacune des images, une à la fois. **Peux-tu me nommer quatre couleurs? Peux-tu me nommer quatre fruits? Peux-tu me nommer quatre animaux? Peux-tu me nommer quatre vêtements?**	**A12**	**x**	**V2**
Comment se nomme cette partie du corps? ***le menton, le coude, le poignet, l'épaule***	**A14**	**x**	**V3**
Qu'est-ce que cette personne fait comme métier? ***le policier, l'enseignant(e), le fermier/jardinier, le médecin***	**A15**	**x**	**V4**
Qu'est-ce qu'on utilise pour écrire? pour s'asseoir? pour se promener? pour balayer?	**C8**	**x**	**V5**
Nomme le plus d'animaux possible.	**C9 (nomme huit animaux)**	**x**	**V6**
Quel temps fait-il dans chacune des images? ***Il vente, il pleut, il neige, il fait soleil***		**A21**	**V7**
Comment s'appelle ceci? ***thermomètre, tasse à mesurer, règle, balance***		**A22**	**V8**

Note: The Expressive Vocabulary scale (six items) is correlated .87 with the items from Domain C of the extended version (C7 à C14). We kept only items asking children to name an object, so as to be as consistent as possible with the methodology of conventional vocabulary tests such as the EOWPVT.

Table C2: Self-awareness Subscale

Items	Initial ÉPE–AD	Extended ÉPE–AD	Modified ÉPE–AD
Self-awareness			
Que devrais-tu faire quand tu : **es fatigué, as faim, as froid, as soif?**	**A7**	**x**	**CS1**
La souris est petite; l'éléphant est ___ ***(gros ou grand).*** Cet oiseau est dehors : cet oiseau est _____ ***(en dedans ou à l'intérieur).*** Ce pot à biscuits est plein; ce pot à biscuits est ___ ***(vide).*** Cette échelle est courte; cette échelle est ___ ***(longue).***	**A8**	**x**	**CS2**
Placer des objets pour démontrer sa compréhension de : ***premier, dernier, devant, derrière.***	**A10**	**x**	**CS3**
Nommer le moment de la journée (le matin, l'après-midi, la soirée, la nuit) associé à deux de ces situations communes : ***des étoiles dans le ciel, prendre le petit déjeuner, retourner à la maison après l'école***	**A13**	**x**	**CS4**
Quel jour est-ce aujourd'hui? **Peux-tu me dire quel jour on sera demain?** **Hier, quel jour était-ce?** **Peux-tu me nommer un jour de la fin de semaine?**		**A17**	**CS5**
Avant de traverser la rue, que dois-tu faire? **Que dois-tu faire si un inconnu te demande d'aller avec lui?** **Dans la voiture, que portes-tu toujours pour être en sécurité?** **Si tu entends le détecteur de fumée, que dois-tu faire?**		**A20**	**CS6**

Table C3: Phonological Awareness Subscale

Items	Initial ÉPE–AD	Extended ÉPE–AD	Modified ÉPE–AD
Phonological Awareness			
Écoute-moi prononcer le mot jouet : jouet commence par le son /j/. **Par quel son commence le mot piano?** **Par quel son commence le mot tomate?** **Par quel son commence le mot boîte?** **Par quel son commence le mot rouge?**		**C16**	**CP1**
Dis-moi si les mots riment, s'ils finissent par le même son. **pomme/homme; boîte/lune; chien/chat; pain/main**	**B11**	**x**	**CP2**
Trouve deux mots qui commencent par /b/ comme dans « ballon ».	**B13**	**x**	**CP3**
Dis-moi si les mots suivants commencent par le même son : **mère/lait, balle/beau, plat/clou, peau/pas**	**B14**	**x**	**CP4**
Écoute pendant que je prononce le mot canif. Le mot canif se termine par le son /f/. **Quel est le son qui termine le mot cheval?** **Quel est le son qui termine le mot album?** **Quel est le son qui termine le mot autobus?** **Quel est le son qui termine le mot neuf (9)?**		**C19**	**CP5**
Voici un bateau (une pomme, une table, une fenêtre et du lait). Le mot bateau commence par un son /b/. Écoute /b/, bateau. **Quelle image commence par le son t?** **Quelle image commence par le son f?** **Quelle image commence par le son l?** **Quelle image commence par le son p?**		**B18**	**CP6**

Note: For item C16, we ask only for the sound at the start of a word and not the letter. The goal here was to measured phonological awareness specifically, and we wished to keep this measure pure. We added the words "piano" and "tomate" to keep the scale score at 4, the same as other version of the test.

Table C4: Numeracy Subscale

Items	Initial ÉPE–AD	Extended ÉPE–AD	Modified ÉPE–AD
Numeracy			
Dire le nombre de parties : **Un chat a combien de queues?** **Un chien a combien de pattes?** **Un oiseau a combien d'ailes?** **Ta main a combien de doigts?**	**A9**	**x**	**N1**
Compte toutes les étoiles et dis-moi combien il y en a.	**B9**	**x**	**N2**
Regarde chaque chiffre et dis-moi lequel est le plus grand. **6 ou 8, 12 ou 10, 9 ou 7, 11 ou 12.**	**B10**	**x**	**N3**
Mets ensemble des jetons pour faire un groupe de… **5, 7, 8, 9.**	**B12**	**x**	**N4**
Montre-moi le chiffre… **13, 20, 45, 112.**		**B19**	**N5**

Note: Pour l'item B19, l'ordre des chiffres a été changé sur l'image afin que les chiffres ne soient pas présentés dans le même ordre que demandés.

Table C5: Memory for Personal Information

Items	Initial ÉPE–AD	Extended ÉPE–AD	Modified ÉPE–AD
Memory for Personal Information			
Peux-tu me dire quel âge tu as et quels sont le jour et le mois de ton anniversaire (de ta fête)?	**A16**	**x**	**MP1**
Savoir où l'on habite : **Quel est le nom de la rue où tu habites?** **Quel est le nom de la ville où nous habitons?** **Quel est le nom de notre province?** **Quel est le pays où nous habitons?**		**A19**	**MP2**

Table C6: Alphabetic Knowledge Subscale

Items	Initial ÉPE–AD	Extended ÉPE–AD	Modified ÉPE–AD
Alphabetic Knowledge			
Dire le son de la lettre majuscule	**B16**	**B17**	**Livret**
Dire la lettre majuscule	**B15**		
Dire la lettre minuscule		**B20**	**Livret**

Note: The letters were presented in order of difficulty based on the performance of Canadian francophone children. A total of 26 letters were presented, including those with accents 'é, è, ê'. Each correct response contributed 1 to the total. The range of the scale is from 0 to 26 for B17 and B20.

Table C7: Oral Reading

Items	Initial ÉPE–AD	Extended ÉPE–AD	Modified ÉPE–AD
Oral Reading			
Lire huit mots perçus de façon globale à fréquence élevée. Dis-moi quel mot est écrit ici.	**B17**	**B16** **Chien** **Oui** **Non** **Balle** **Chat** **Le** **La** **Nez**	**L1** **Le** **La** **Oui** **Non** **Balle** **Maman** **Nez** **Papa**

Note: We changed the order of the words so that they are in decreasing frequency of usage for French. We changed two of the words for alternatives that are more commonly encountered by French children. The words “dog” and “cat” are very common is English-language alphabet books, but the French equivalents are more orthographically complex (i.e., “ch” is a multi-letter symbol for one sound). We substituted these words with “maman” [mom] and “papa” [dad], which are slightly longer but also more frequent in French-language children books.

SOCIAL RESEARCH AND DEMONSTRATION CORPORATION

OTTAWA OFFICE
55 Murray Street, Suite 400
Ottawa, Ontario K1N 5M3
Tel.: 613.237.4311
Fax: 613.237.5045

TORONTO OFFICE
481 University Avenue, Suite 705
Toronto, Ontario M5G 2E9
Tel.: 416.593.0445
Fax: 647.725.6293

VANCOUVER OFFICE
128 West Pender Street, Suite 301
Vancouver, British Columbia V6B 1R8
Tel.: 604.601.4070
Fax: 604.601.4080

www.srdc.org