ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Behavioural and Health Economics

The effects of Project Upgrade's literacy interventions on children's language and literacy outcomes: heterogeneity by gender and home language

---

Name student: Anna Tameris

Student ID number: 622758

Supervisor: Ilse van der Voort

Second assessor: Aysenur Ahsen

Date final version: 10-07-2024

Abstract

Low literacy is a big problem in America, which is defined as having difficulties with reading, writing, and/or with performance of simple mathematical skills. To address this big problem, a two-year RCT intervention program called Project Upgrade, in Miami-Dade County, Florida, will be studied. The program tests the effectiveness of three different types of language and literacy interventions, implemented in child care centers, on improving teachers behavior and the environment in the classroom, to eventually improve children's outcomes based on their language and pre-literacy skills. While previous studies are focused on the outcomes in comparison with characteristics of the teachers, this paper will look at how the effect of Project Upgrade's pre-school literacy interventions on the TOPEL scores and later reading and math achievement scores, differ based on child's sex and home language. There will be focused on four year old children whose parents have a low-income. A sample of 999 children clustered in 151 centers of the existing dataset of Project Upgrade, from the Child & Family Data Archive (2011), will be analyzed. Normal ordinary Least Squares (OLS) regressions and regressions with interaction terms, including cluster-robust standard errors will be performed. The results show that the RSL and BELL intervention have different effects by gender on some of the scores, with women always having higher effects than men. There is also evidence of all the interventions having different effects by home language on some of the scores, where the language group with the highest effects differ per outcome variable.

Table of Contents

1. Introduction

Low literacy is a big problem in America (U.S. Department of Education, 2019). It is defined as having difficulties with reading, writing, and/or with performance of simple mathematical skills (Stichting Lezen en Schrijven, 2021). Result from the Nation's Report Card (U.S. Department of Education, 2019) show that more than 60% of the students in American schools are reading below grade level, and that 54% of the adults in America read below the sixth-grade level. The literacy rates have not been increased since 2000 (Schleicher, 2019) and the reading levels of children decreased even more since the COVID-19 pandemic (Curriculum Associates, 2021). Low literacy rates come with a high cost, as it is correlated with a higher unemployment rate, a reduction in income, a higher percentage of people in prison and bad health outcomes (World literacy foundation, 2018). It creates a cycle that can go from generation to generation, which keeps the inequalities a live (Barbara Bush Foundation for family literacy, 2021).

To address this big problem, I will study an intervention program called Project Upgrade, which is explicitly focused on developing language and emergent literacy skills among four year old children whose parents have a low-income (Layzer, Layzer, Goodson & Price, 2007, 2009; Layzer & Price, 2010). The program is a two-year RCT intervention program in Miami-Dade County, Florida, that tests the effectiveness of three different types of language and literacy interventions, implemented in child care centers, on improving teachers behavior and the environment in the classroom, to eventually improve children's outcomes based on their language and pre-literacy skills. Language and literacy are defined as more than just saying something or being able to read (Stichting Lezen en Schrijven, 2021; Gee, 1989). In Project Upgrade I use three domains to predict the development of literacy, with language being a part of it, which are; Definitional Vocabulary, Phonological Awareness and Print Knowledge (TOPEL scores: Lonigan, Wagner, Torgesen & Rashotte, 2002). Also reading and math scores are assessed, for a complete definition.

Focusing on this specific program is very relevant, since it addresses the big problem of low-literacy, with especially the focus on children of low-income families. Previous literature shows that it is argumentative to focus on children in disadvantaged families, because they are at a high risk for social and economic failure (Heckman, Holland, Makino, Pinto & Rosales-Rueda, 2017). They are more likely to commit a crime and drop out of school early, and they are with their language a year behind the national norms. Layzer et al. (2007) and Grimm (2008) show that starting with curricula that develop the child's language and literacy already at a young age, 4 years old in this experiment, will result in positive long term effects on reading, math and a lot of other aspect later in life. Besides this, the program has a professional development component where teachers receive three training session. This focus on training child care staff to improve children's language and emergent literacy is really important, because

requirements for becoming a teacher in Florida are easily met, which results in low quality teachers with limited experience (Teachers of Tomorrow, 2023). With training, the teachers will be better educated, resulting in better behavior and interaction with the children, leading to improvement in children's language and literacy skills (Layzer et al., 2007).

So, Project Upgrade is addressing the big and important problem of low literacy, with a relevant and effective program, according to the previous mentioned literature. However, previous literature about the program (Layzer et al. 2007, 2009; Layzer & Price, 2010) is really focused on the outcomes in comparison with characteristics of the teachers. Since the ultimate goal of the program is to create the best outcomes possible for the students, regarding language, literacy and a lot of other aspects in life, I will be specifically focused on the child in this paper, by looking at heterogeneous treatment effects regarding the sex and home language of the child. This results in the following research question:

"How does the effect of Project Upgrade's pre-school literacy interventions on the TOPEL scores and later reading and math achievement scores, differ based on child's sex and home language?"

For answering this question and the related hypotheses, given in section 3, I will make use of the existing dataset of Project Upgrade in Miami-Dade County, from the Child & Family Data Archive (2011) in the period of 2003 until 2009. In my analysis I will use variables of the child-level data set, class-level dataset of 2003 and the follow-up dataset, eventually resulting in a sample of 999 children clustered in 151 centers. This sample is a good representation of the real life population of four-year-old children going to a child care center in Miami-Dade county, where some of the children receive subsidies, and for centers elsewhere that serve children with low-income (Layzer et al., 2007). The used method for analyzing the data is Ordinary Least Squares (OLS) regression, with cluster-robust standard errors, clustered at class level. First, some normal OLS regression analyses will be performed, to compare my results with previous results and to fill the gap of some missing outcomes of previous research. After this, OLS regressions with interaction terms will be performed to look for heterogeneous treatment effects by the sex and home language of the children. The analyses will be performed in Stata, with the results being converted to effect sizes.

The results show that the RSL and BELL intervention have different effects by gender on some of the scores, with women always having higher effects than men. There is also evidence of all the interventions having different effects by home language on some of the scores, where the language group with the highest effects differ per outcome variable. For all the other combinations of interventions and outcome scores, there is no evidence of different treatment

effects by gender and home language of the child, on the scores. The differences between student with different characteristics was not studied before for the program, and thus with these results I will fill an important literature gap (Chin and Spector, 2019) and I will contribute to a growing literature about this topic (Heckman et al., 2017). Also with this focusing heterogeneous treatment effects, I show which group of students benefits the most of which intervention program, which gives policymakers the possibility to target the right interventions towards the right teachers and student. Eventually this will lead to children being better prepared for elementary school, having better grades and a gain for society as a whole. According to a study of the Organization for Economic Cooperation and Development (World literacy foundation, 2018.), there is a correlation between having a higher literacy and political efficacy, having a higher trust in others, participating more in voluntary activities and having a better health. While the challenges in different cities may differ in degree, the result of the study could help many other communities too (Layzer et al., 2007), because of the representativeness of the sample.

In the remainder of this paper, I will first give an in-dept overview of Project Upgrade in Section 2. After this, Section 3 will provide a theoretical framework of previous literature regarding research of Project Upgrade and regarding heterogeneity by gender and home language. Subsequently, in Section 4 an overview of the used data is given, including sample selection, randomization and attrition, the variables of interest and baseline balance tests. Also an explanation of the methods used for analyzing the data is given in this section, including the OLS assumptions, This is followed by the presentation and interpretation of the results in Section 5, with robustness checks included. Finally, in Section 6, the results will be critically and in more detail discussed, and a summary of the results will be given. Resulting in an answer to the research question, implications and suggestions for further research. In this paper, I will use the words intervention/treatment, class/center and sex/gender interchangeably.

2. Project Upgrade in Miami-Dade County, Florida: An In-Depth Overview

Project Upgrade is a two-year program, that attempts to improve the English language and pre-literacy skills of four-year-old children, whose parents have a low income. They do this by implementing three different language and literacy interventions at child care centers in Miami-Dade County, Florida (Layzer et al., 2007, 2009; Layzer & Price, 2010). The program is part of the multi-site, multi-year Evaluation of Child Care Subsidy Strategies, whose goal is to try to get as much information as possible about how to allocate child care subsidies in the most effective way, to improve the quality of child care. Before the start of the program, the Early Learning Coalition (ELS) assessed four-year-old children who were receiving subsidies, which

resulted in finding a big gap in their language development. As a response, they introduced Project Upgrade between fall 2003 and spring 2005.

The program is implemented in Miami-Dade County, since it is the largest and most populous country in Florida, with a lot of diversity in ethnicity and languages spoken (Layzer et al., 2007). There are some challenges in the child care system, like for example the lag in language development of the children. There are also high teacher turnover and low education achievements of teachers, which make providing high-quality education more difficult. Besides these problems, previous evidence about the importance of good language and pre-literacy skills at an early age on the success in reading and math later in life (Layzer et al., 2007; Grimm, 2008), gives a good reason for implementing the program.

Project Upgrade is only for child care centers that met a couple of criteria, which were selected by the ECL (see section 4, Methodology). Eventually, 165 centers were randomly assigned to the three different interventions or to the control group (see section 4, Methodology). All children in the classroom received the intervention, regardless of receiving subsidy or not.

The three intervention programs are Ready, Set, Leap (RSL), Breakthrough to Literacy (BTL) and Building Early Language and Literacy (B.E.L.L.), two nationally-known and one local developed curriculum respectively. They are all focused on providing support for the development of English knowledge and early literacy skills of the children (Layzer et al., 2009). However, they differ in instructional strategies, intensity, cost and the materials that are provided. RSL used throughout the day three interactive technology tools and activities around a thematical collection of trade books, for stimulating oral language, phonological and print knowledge. B.E.L.L focused more on adding a pre-kindergarten literacy component, with two 15 minute whole group sessions a day, to stimulate the general language, print awareness, phonological awareness, and shared reading skills of the child. At last, BTL implemented an integrated literacy and language curriculum, spread throughout the day. This included activities built around a Book of the Week, with the focus on reading aloud and knowing questions about the book, to stimulate the vocabulary. It also included 8 to 12 minute computer session of the Book of the Week, to stimulate print and phonological knowledge. Both BTL and RSL had additional math and science activities and their curricula existed of whole group, small groups, and individual sessions. All interventions had some materials in Spanish to stimulate reading.

Besides the curriculum component of the interventions, there is a professional development component where teachers receive three training sessions. The first training is about the implementation of the curriculum, supplemented by refresher sessions. They also receive bi-weekly visits of trained mentors over approximately 18 months and there is some supervision, including support and feedback about specific things to work on.

The whole program is formed with the goal to deliver good curricula, find good ways of training teachers, and to see the impact of the training and support on the behavior of the teachers, on the class environment and eventually on the English language development and pre-literacy skills of the child. They hoped to find evidence of effectiveness of the programs, so that they can implement one of more of the curricula to the system as a whole.

3. Theoretical Framework

Project Upgrade uses three different language and literacy interventions in child care centers to improve the language and pre-literacy skills of the children. Literacy is not only about reading, but it is also about writing and/or performing simple mathematic skills (Stichting Lezen en Schrijven, 2021). When someone has difficulties with this, he has low literacy, which makes it hard to fully participate in society. According to Gee (1989 language is, besides grammar and what you say, also about how you say things and about what a person does when he talks.

In this paper, I will use the abovementioned broad definition of language and literacy. Three domains will be used as predictors of the literacy development, which are; Definitional Vocabulary, Phonological Awareness and Print Knowledge (Layzer et al., 2007; Lonigan et al., 2002). Language will be looked at as being a part of literacy. Also reading and math scores are assessed, for having a more complete definition.

3.1 Previous research

There are already some articles discussing the effects of Project Upgrade. Layzer et al. (2007) focus on the impact of the interventions on the behavior of the teacher and the environment in the classroom, seen as intermediate outcomes, and on the impacts on early literacy skills and language development of the child. They found that all of the interventions had significant positive effects on most of the aspects of the Observation Measures of Language and Literacy Instruction (OMLIT: Goodson, Layzer, Smith, Rimdzius, 2004), which represents four aspects of behavior and interaction of the teacher with the children, that support literacy. Also some significant positive effects were found on literacy resources and activities that involve literacy. However, the interventions differ in the exact aspects they had a significant effect on. Besides this, the RSL and BTL curricula had significant positive effects on all of the four measures of the Test of Preschool Emergent Literacy score (TOPEL: Lonigan et al., 2002), which are aspects of language development and pre-literacy skills of the child predicting success in reading later in life. A more detailed explanation of the scores are given in the methodology part. In 2009, Layzer et al. (2009) published the final report of the previous research, including some more detailed information about the design, the implementation of the interventions and a cost-effectiveness analyses for the RSL and BTL intervention. They found that RSL is most

cost effective for every child outcome measure. The rest of the paper showed the same result as given before. Because of some strong and significant positive short-term effects, Layzer et al. (2010) did more research about long-term effects of the interventions of Project Upgrade. They looked for effects on first, second and third grade math and reading scores in elementary school. The research focused on the RSL and BTL intervention combined, because of no significant impacts of the BELL intervention on children's outcomes in the previous study. Children from previous study (Layzer et al., 2007), that were still present, were divided in two cohorts that enter elementary school a year apart, because of differences in age. Significant positive effects where found for the younger cohort on their first grade reading and math score and on their second grade math score.

Previous research about Project Upgrade was really focused on the outcomes in comparison with characteristics of the teachers, like teacher's educational background and their training language. First of all, they found that since the interventions include training and mentoring of teachers, teachers became almost the same in their behavior toward children in supporting their literacy (Layzer et al., 2007). This resulted in no interaction effects on the outcomes of the child. Chin and Spector (2019) also found that the interventions of Project Upgrade are most effective for teachers who have poor instruction qualities at baseline, which will eventually lead to equalization of the instructional quality of the teachers. Besides this, all of the outcomes were compared between teacher with primary language English and teacher with Spanish as language. They found that effect of the interventions on teachers behavior/classroom environment and on the child outcomes were stronger for Spanish speaking teachers than for English teachers, and for children in the classes of Spanish speaking teachers. Layer et al. (2009) also found that the impacts on the read and math scores were bigger for children with a teacher who speaks Spanish. However, Chin and Spector (2019) found that an underperforming teacher who speaks Spanish does not have different treatment effects on the outcomes.

The research already done about Project Upgrade mention some limitations. Although the program resulted in low-income children moving closer to the national norm on three of the four aspects of the TOPEL score for English language development and pre-literacy skills, they stayed very far behind the national norm with their vocabulary skills (Layzer et al. 2007). This is possibly because of the Spanish speaking children, who start with really low vocabulary skills in English. Eliminating this gap asks for the right intervention to be given to the right children (Layzer et al. 2010). Chin and Spector (2019) agreed on this by saying that we need to identify which group of teachers and student gain the most from which intervention. As already mentioned before, previous studies really focused on the teachers and not on the children. However, the ultimate goal of Project Upgrade is to create the best outcomes possible for the

students, regarding language, literacy and a lot of other aspects in life. By identifying which group of students benefit the most of which intervention program, I fill an important literature gap (Chin and Spector, 2019).

So, in this paper I will be focused on the children by looking at heterogeneous treatment effects on the TOPEL scores and the math and reading scores, regarding the sex and home language of the child. With heterogeneous treatment effects I mean the extent to which various interventions/treatments have different effects for specific groups (Imai & Ratkovic, 2013).

3.2 Heterogeneity by gender

Previous research suggests that men and women react differently on early childhood interventions that involve preschool education. The HighScope Perry Preschool Program, which provided preschool of high quality to a random group of disadvantaged African-American children in Michigan, found economically important significant effects for men and women (Heckman, Moon, Pinto & Savelyev, 2010). However, women have stronger effects for education attainments and employment early in life, and men will catch up later with stronger significant effects in later life outcomes, like employment at age 27-40. Elango, Garcia, Heckman and Hojman (2016) confirm the differences in treatment effects for men and women, by finding gender differences of the Perry Preschool Project, the Infant Health and Development Program, the Carolina Abecerdarian and the Early Training Project. Women develop earlier in life, and because of this they will have more benefits of the preschool interventions. Their literature also mention that because of the gender differences, it might be a good idea to introduce gender-specific curricula to the children in preschool.

Looking at reading skills, Buchmann, DiPrete and McDaniel (2007) found that the reading skills of women are better than that of men when they enter kindergarten, and that there will stay a gap between their reading skills during elementary school (Trzesniewski, Moffitt, Caspi, Taylor & Maughan, 2006). Anti-social behavior and emotional behavior of men are given as an explanation. Where women experience advantages of the classroom environment, men have difficulties with this (Zill & West, 2001), resulting in them being more disruptive in class, paying not much attention and being negative about learning activities. However, Legewie and DiPrete (2012) conclude that men can actually also have advantages of the classroom environment. They found that men are in general more sensitive to the classroom environment than women, which can result in them benefiting more in terms of their learning orientation, work habits, attitude toward school and eventually in their grade achievement, when they are in a classroom with children and teachers with a high socioeconomic status. Diette and Oyelere (2014) contribute to this, by finding that having a lot of student around you with Limited English will result in no effects on math and reading scores for women, but significant negative effects for

men. Also Millard (1997) state that men and women in the same class experience education very differently, with men being more sensitive to the environment of the class/school they are in. Legewie and DiPrete (2012) mention that further research needs to be done about how to create the best classroom environment for men and women, and about how/to what extent interventions have an effect on the performance of men and women separately.

Looking at literacy, there are a lot of differences between men and women in their use and experience of the literacy education in class (Millard, 1997). Their development in how they read also varies; they prefer other books and act differently in how they organize and share their readings with others. Millard (1997) also found that these differences in attitude towards reading and writing between men and women diverge even more over time. However, about the exact differences between men and women and the changes over time is still a lot of discussion. Some researchers have said that the differences in test scores of men and women have declined over time (Hyde, Fennema, Lamon, 1990; Feingold, 1988), while others said that the differences in writing, math and science results remained the same (Hedges & Nowell, 1995) or even become bigger (Millard, 1997). More research needs to be done about how big these differences between men and women is nowadays. Millard (1997) mentioned that a lot of research starts when children enter kindergarten, so more research about the differences by gender already during child care centers is an addition to the literature.

Following from the above mentioned research about the differences between men and women regarding their reaction on interventions, their reaction to the classroom environment and the differences in how they make use of literacy and reading tools, I formulate the following two hypotheses:

“Project Upgrade’s pre-school literacy interventions causes different effects on the TOPEL scores for men and women”

“Project Upgrade’s pre-school literacy interventions causes different effects on reading and math achievement scores for men and women”

For clarification, the TOPEL scores (Lonigan et al, 2002) are aspects of language development and pre-literacy skills, predicting success in reading later in life.

3.3 Heterogeneity by home language

Besides looking at gender differences, it is important to look at the differences in treatment effects by the home language of the child. The National Center for Education Statistics (2011) found that in 2008 21% of the school-aged children in the US spoke at home another language than English, which is still a representative percentage nowadays (ChildStats Forum, 2023).

Miami-Dade County is ethically and linguistically really diverse, with a majority of Hispanics, having Spanish as their home language (Layzer et al., 2007). From the children that took part in Project Upgrade, 54% spoke Spanish as their home language, 41% spoke mainly English and 1% spoke Haitian Creole. There were classes where everyone, including the teacher, spoke English, Spanish and there were mixed classes. Mixed classes had a English teacher and always an aide who spoke Spanish or Haitian Creole. Previous literature of Hoff (2013) shows that children with another home language than English, will have a different language development than children who speak only English at home. While children with other languages have some unique linguistic strengths, many of them will start their school with lower skills of the English language. Also Layzer et al. (2007) and Scheele, Leseman and Mayo (2010) mention that children with other home languages will already have disadvantages in English language skills when they start with schooling, especially in their vocabulary. These differences in skills will have negative consequences for later academic achievement. So, the home language of the child is also an important factor to look at, since it can influence the academic performance of children to a great extent (Oller & Eilers, 2002).

Kramsch (2014) gives a more in depth explanation for the negative effects on academic achievement. She says that people who speak different languages think very different when they speak and they use other linguistic forms, which will have an influence on their cognitive processes. These differences in cognitive patterns, that form our thinking, can make it more difficult to understand what other people say, leading to smaller effects of treatments on achievement scores.

Previous research of Project upgrade (Layzer et al., 2007) already included a small analysis of the combined effect of the RSL and BTL treatments on the English language development and pre-literacy skills, for Spanish/Creole and English speaking children separately. They found that both language groups had significant effects on all four TOPEL scores, representing language development and pre-literacy skills. However, children with Spanish or Creole as home language had bigger effect sizes on each score, than the English speaking children. This is not in line with the literature mentioned in previous paragraphs, where smaller effects sizes for the non-English home speakers were expected. More in line with the previous research is that the children with a home language other than English will still have lower mean scores on language and pre-literacy skills than children speaking English only.

Some children that take part in Project Upgrade speak more than one language at home (Layzer et al. 2007). Scheele, Leseman and Mayo (2010), found that bilingual children will have a disadvantage in language skills in both of their languages, especially in their vocabulary. This can be explained by the fact that bilingual families have to divide their specific language

inputs of oral and literacy activities between the two languages, which will result in the bilingual children having less oral and literacy interactions for each language apart than the monolingual children have for their language. Eventually, this will result in lower language skills, which will again have negative effects on their academic achievements.

Following from the above mentioned research about the differences between children with different languages, I formulate the following two hypotheses:

"Project Upgrade's pre-school literacy interventions causes different effects on the TOPEL scores for children with just English as home language, than children speaking Spanish/Spanish and English or other languages at home"

"Project Upgrade's pre-school literacy interventions will causes different effects on reading and math achievement scores for children with just English as home language, than children speaking Spanish/Spanish and English or other languages at home"

4. Methodology

4.1 Data description and collection

For finding an answer to the research question and hypothesis, I will make use of the existing dataset of Project Upgrade in Miami-Dade County, from the Child & Family Data Archive (2011). This is an experimental dataset containing data from the early-childhood intervention program in the period of 2003 until 2009. The dataset of Project Upgrade contains five separate STATA datafiles existing of one child-level dataset, three class-level datasets and one follow-up dataset (at child-level), which can be merged with each other using the *Center_ID* and *Student_ID* variables, given in the datafiles.

In this paper, I will use variables of the child-level dataset, the class-level dataset of 2003 and of the follow-up dataset. The child-level dataset contains variables about children's demographic characteristics and about their pre-literacy and language skills, measured in 2005 with the TOPEL scores (Lonigan et al., 2002). The class-level dataset of Fall 2003 exists of baseline variables of the class, like the language and literacy environment and interaction in the classroom, measured with the OMLIT scores (Goodson et al., 2004). Also information about aspects of the interaction of the teacher with the children are given with the Arnett Caregiver Rating Scale (Arnett, 1989) and LapD scores (Hardin, Peisner-Feinberg & Weeks, 2005), representing the class mean score of cognitive, language and fine motor skills, are included at baseline. On top of this, variables that represents teachers preferred training language and their attained education level, self-administered via a questionnaire, and experimental design variables like the randomization block a class is in and treatment

indicators, are included. The dataset of the follow-up contains child-level 1st and 2nd grade reading and math achievement scores, for different type of students and different cohorts from the schoolyears 2007-2008 and 2008-2009, measured with the with the Stanford Achievement Test 10 (SAT-10: Pearson, 2022).

I will use all of the data mentioned in previous paragraph in my analyses. In the 'variables of interest' part, I will specify in more detail what the exact variables are that I will use from this data and in which way I will use them. There I will also give more information about some of the measurements of the variables.

4.2 Sample selection, randomization and attrition

The above mentioned data is available for a specific sample, generated by the ELC through a selection process before the start of the interventions (Layzer et al., 2007, 2009). The intervention program is only for child care centers in Miami-Dade County who met a couple of criteria. The first one is that the centers need to serve children who had their care subsidized or just children who come from low-income families. The reason for this is previous research that mention the importance for especially children with a high risk to have education programs early in life for the development of their literacy. All the interested centers that met this criteria needed to fill out an fact sheet, to see if they met all the other criteria too. One of these criteria is that the centers need to have at least one classroom with at least five four-year-old children, when they were recruiting. Only one classroom per center was eligible for the program, so if more than one class met the criteria, the one with the most subsidized children was chosen. If this was also equal, than the class with the most children in it was chosen, and otherwise the class was chosen randomly. The choice of preferring the larger classrooms, is because it is easier to detect a significant effect on the children's literacy and language development in big classes. Another reason is that when the interventions have positive effects, more children will benefit in this case. The final criteria is that the centers cannot already have a literacy curriculum, because I only want to test the effect of the curricula of Project Upgrade.

The selection process resulted in 300 eligible centers (Layzer et al., 2007). After a series of meetings, where some centers were eliminated, the 200 remained centers were randomly assigned. The randomization process started by strata-randomization, where the centers were sorted in 4 homogeneous groups based on their agency affiliation and the (preferred) training language of the teacher. Within these groups, the centers were sorted by the amount of four-year-olds in the classroom, in blocks/clusters of 12 centers. Eventually cluster-randomization took place where, within these 20 blocks, 3 centers were randomly assigned to the control group, 2 centers to each treatment group and 3 centers were reserve for when some center

declined to take part, before even knowing their assignment. This design is unbalanced, because of budget constraints that limited the amount of curricula that could be tested and the amount of centers that could receive a treatment. Eventually, 165 centers agreed to take part and received their assignments, resulting in 38 centers being assigned to RSL, 36 to BELL and BTL, and 55 centers to the control group. All the children in these classrooms were allowed to participate, not only the once who received subsidy, resulting in data of 1535 children that were at least two months in the classrooms and from who the parents agreed to them being assessed.

During the two years of intervention, five centers left because of closure or reselling the center to someone that wasn't interested in the program, and there left two centers because the directors didn't want to continue with the assigned curriculum anymore. The attrition was no problem, since it were just a few centers and it was quite evenly distributed across the groups. During the follow-up (Layzer et al., 2010), data about 1137 children of the original sample are used and 127 children that were already in the 165 centers but not present when the child assessments in the original study took place, were added to this.

Before doing the analyses of this paper, the data was checked and cleaned, resulting eventually in a sample of 999 children from 151 centers that is a good representation of the real life population of four-year-old children going to a child care center in Miami-Dade county, where some of the children receive subsidies. Layzer et al. (2007) also mentioned that the centers were representative for centers elsewhere that serve children with low-income. They verified this by comparing the LapD scores of the centers in the sample with other centers, showing no significant differences between the centers.

4.3 Variables of interest

Dependent variables

For answering the research question partly and looking at hypothesis 1 and 3, I will use the standardized *TOPEL scores* (Lonigan et al., 2002) as dependent variables. This is an assessment that consists of 3 aspects of language development and pre-literacy skills of which previous research has shown to be a predicter of success in reading later in life. The aspects are *Definitional Vocabulary, Phonological Awareness and Print Knowledge,* leading to one combined index called *Early Literacy Index.* The scores were first assessed from the children in the classes with an individual test of the pre-curser Pre-CTOPP (Lonigan et al., 2002), during spring 2005. These scores were in 2006 converted to the TOPEL standardized scores with a mean of 100 and a standard deviation of 15, where a higher value means a better achievement.

More information about the measuring, converting and quality of the scores is given in appendix A1.

For answering the other part of the research question and looking at hypothesis 2 and 4, I will use the *1st and 2nd grade reading and math scores* as dependent variables. I focus on the scores of grade 1 and 2, since these are available for all type of students in the measured schoolyears 2007/2008 and 2008/2009 (Layzer et al. 2010). This will result in a representative sample generating representative (heterogeneous) treatment effects. A more in detail explanation about this is given in appendix A1. The scores are obtained with the SAT-10 Test (Pearson, 2022). There are different versions of the SAT-10 test, with different subsets that also vary by grade. In Project Upgrade I look at the achievement of the children in their reading comprehension and how good they are in mathematical problem-solving, with asking multiple choice questions about different components of reading and math. For the math score these are; number sense and operations, patterns relationship, data statistics and probability, and geometry and measurement. The read score includes questions regarding; initial understanding, interpretation, critical analysis, reading strategies, literary, informational and regarding functional. The total scores are created by first adding up the amount of questions of the subsets answered right, and after this the score is rescaled. This is done so that the performance of the Florida students can be compared to students from other nations and so that it is possible to compare same grade scores, obtained in different years by different students (Pearson, 2022). Because of this, the *1st and 2nd grade reading and math scores* are represented as four outcome variables, where the scores from both schoolyears and all types of students are combined into one variable for each type of achievement of each grade. The higher the score, the better the achievement of the child.

Independent variable

The independent variable in the analyses is the group a child is assigned to. The groups exist of three treatments groups*;* Ready, Set, Leap! (*RSL*), Building Early Language and Literacy (*B.E.L.L.)* and Breakthrough to Literacy (*BTL*), already explained in section 2, and the control group, where someone does not receive any of the treatments. To which group a child is assigned, is given through a categorical variable, which gives 1 if the assigned treatment is *RSL*, 2 if the treatment is *B.E.L.L.*, 3 if it is *BTL* and 4 if the child is in the *control group*. The assigning is also given through the dummy variable *treatment*, which is 1 if the child receives any of the treatments and 0 if the child is in the control group. During a meeting with ECL before the start of the program, the randomly assignment took place and was reported for further analysis. In the analyses, both the categorical and dummy variable will be used.

Variables for heterogeneous treatment effects

For looking at heterogeneous treatment effects by gender and home language of the child, I need a variable for gender and home language that can be used in the regression with interaction terms. For the gender of the child, I will use the binary variable *male*, that represents a men when the variable is 1 and a women when it is a 0. The home language of the child is the primary language that is spoken at the child's home. It is given through a categorical variable *home language*, where the child has English only as home language when the value is 1, English and Spanish or Spanish only when the value is 2 and English and other, Spanish and other or just another language when the value of the variable is 3. I can assume that both variables stay the same during the period of the intervention and the assessments, since all the children have the exact same gender and home language in our child-level dataset, measured at the beginning of the program, as in our later measured follow-up dataset (Layzer et al. 2007, 2010).

Control variables

In the analyses, also some control variables are used for robustness checks and they are included in the baseline tests. I use only good control variables in our analyses, which are variables that are measured before the interventions took place, and they are thus not influenced by the treatment. The control variables included are; the continuous variable *age* of the child (given in years), a categorical variable of teacher's *highest attained education degree* and a dummy variable of the (preferred) *training language* of the teacher. Besides these child an teacher characteristics, some scores of the class the child is in are included. First, four *OMLIT scores* (Goodson et al., 2004), focusing mainly on the literacy and language environment of the class and also on the interactions within the class, are included. Also three *Arnett scores* (Arnett, 1989) that give a value to the discipline style, the emotional tone, encouragement of independence and supervision of and interest in the children by the teacher in the class, are added. At last, three *LapD scores* (Hardin et al., 2005) are included, which represents the class mean score of cognitive, language and fine motor skills.

Besides these control variables, the variable that gives information about which *center* each child is in, is necessary for the use of cluster-robust standard errors. This is given by the variable *center_id*, which represents the number of the center the child is in. For each center, there is a different, unique number and there are children with the same number, meaning that they are in the same classroom. This information is collected by mentors that visited the classrooms/centers, during the two-year program.

The statistical characteristics of all the variables mentioned above are presented in the descriptive statistic table in appendix A1.

4.4 Baseline Balance tests

In this paper baseline balance tests are done to check if the randomization is done correctly, so that the variables are approximately the same across the different groups, before the start of the interventions. The results of the tests will be given in detail appendix A3, from which I can conclude that random assignment was done successfully, which makes it possible to interpret the results of our further analyses as causal effects.

4.5 Analysis method

For answering my research question and the hypothesis, I will make use of Ordinary Least Squares (OLS) regressions with cluster-robust standard errors. I use OLS since this is a clear method for describing a relationship between variables, and it is especially suited for continuous dependent variables in combination with any kind of independent variable, which is the case in our analyses (Stock & Watson, 2020). Besides this, OLS exists of a flexible framework, that can be extended in various ways. The model can be extended to a multiple regression model by adding control variables, which makes it possible to check for robustness. It also allows for nonlinear relationships, like interaction effects, as long as the parameters are linear. This is a useful tool for looking at the heterogeneous treatment effects in our analyses.

In the analysis, first I will use OLS regressions to compare my result of general treatment effects on the TOPEL, reading and math scores with the results of previous papers (Layzer et al., 2007, 2009, 2010), and to fill the gap of some missing outcomes in previous research:

$$Y_{ic} = \beta_0 + \beta_1 T_c + \beta_2 X_{ic} + \varepsilon_{ic}$$

$$Y_{ic} = \beta_0 + \beta_1 T_{RSL,c} + \beta_2 T_{BELL,c} + \beta_3 T_{BTL,c} + \beta_4 X_{ic} + \varepsilon_{ic}$$

Y_{ic} is the continuous outcome variable for child i sitting in class c. It represents one of the four different aspects of the *TOPEL score or the 1st, 2nd grade math or reading* score. The variable T_c in the first model is the dummy variable *treatment* with value 1 if the child receives any of the treatments and 0 if the child is in the control group. In the second model, I have three dummy variables for treatment, where $T_{RSL,c}$ is equal to 1 when RSL is the treatment and 0 otherwise, $T_{BELL,c}$ is equal to 1 when BELL is the treatment and 0 otherwise, and $T_{BTL,c}$ is equal to 1 when BTL is the treatment and 0 otherwise. When the class is assigned to the control group, all dummies will be zero. As mentioned in the 'variables of interest' part, I will use the

categorical treatment variable in the STATA analysis, instead of the separate dummies for the treatment groups, since STATA will automatically see them as dummies and this will reduce the possibility of making mistakes. Both treatment variables vary at class level. X_{ic} is a vector of control variables. I will first do a regression without control variables. After this I will include the Arnett Not Detached subscale, the three LapD scores and the OMLIT Oral Language, Print Motivation and Literacy Resources score variables, because they were different between groups at baseline, to show the robustness of our results. Eventually, I will include all the baseline variables to show that our results are not sensitive when I add those variables as controls. ε_{ic} is a cluster-robust standard error, clustered at class level to account for the clustering in the program.

I cluster at class level, because the children in classes may not be independent, which could lead to dependent residuals within the clusters (UCLA, 2021). With adding cluster-robust standard errors, this non-independence within the class is taken into account. Since OLS can cluster only at one level, I will cluster at class level, because I am interested in the children, who are mostly affected by everything that happens in the classroom instead of the things that happen in the blocks. The correlation at class level is most relevant for my research and not clustering for this correlation may give problems. Clustering at block level is less relevant and the estimates might even be less precise, since the amount of blocks is way smaller than the amount of classes/centers (Michols & Schaffer, 2007).

After the abovementioned OLS regressions, I will do regressions with interaction terms included to look for heterogeneous treatment effects. Below, I give the most simple form with a categorical variable with just two categories, for example gender, and the dummy treatment variable:

$$Y_{ic} = \beta_0 + \beta_1 D_{ic} + \beta_2 T_c + \beta_3 T_c * D_{ic} + \beta_4 X_{ic} + \varepsilon_{ic}$$

Y_{ic} is again one of the continuous outcome variables for child i sitting in class c. The variable D_{ic} is a dummy variable(s) referring to some of the categories of the categorical variables gender or home language of the child. For the gender there is only one dummy included; D_{male}, with 1 the child being a man and 0 being a women. For home language, two dummies will be included, because of three categories of language. $D_{EngOth/SpanOth/Oth}$ is 1 if home language is English and other, Spanish and other or just another language and 0 otherwise, and $D_{engSpan}$ is 1 if the home language of the child is English and Spanish or Spanish only and 0 otherwise. When the child has only English as home language, all dummies will be zero. Besides this I will again have regressions where the variable T_c is the dummy treatment variable, with value

1 if the child receives any of the treatment and 0 if the child is in the control group. After this, regressions with the categorical treatment variable will be done, where three dummies will be included; T_{RSL} is 1 if the child received the RSL treatment and 0 otherwise, T_{BELL} is 1 if the child received the BELL treatment and 0 otherwise and T_{BTL} is 1 if the child received the BTL treatment and 0 otherwise. When the class is assigned to the control group, all dummies will be zero.

$\beta 3Tc * Dic$ represents one of the interaction terms between the gender/home language dummies and the dummies that refer to the treatment groups. With the categorical home language and treatment variable, more interaction terms needs to be included, since each variable contain more than one dummy. X_{ic} is a vector of control variables and ε_{ic} the cluster-robust standard error, clustered at class level to account for the clustering in the program. I will again do multiple regression with including several control variables, as given in previous simple OLS regressions.

The model specification above is the more detailed version of the one I will use in the Stata analysis. In Stata I will use the categorical variables for language and the treatment groups, instead of the separate dummies, since STATA will automatically see them as dummies and this will reduce the possibility of making mistakes.

The results of the analyses will be given in effect sizes. This is a standardized measure, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group (Layzer et al., 2007, 2009; Layzer & Price 2010). I make use of these effect sizes, to be able to compare the results with previous research of Layzer et al. (2007, 2009, 2010) and because it helps understanding the magnitude of the effects found (Sullivan & Feinn, 2012). For example, the results can be compared with the rule-of-thumb of Cohen (1988) to see how big the effects are. He says that an effect is small, medium and large when the effect sizes are 0.20, 0.50 and 0.80 respectively. It can also be compared with a national benchmark, where Hill, Black & Lipsey (2008) calculated the average expected annual gain of the transition to a higher class. For reading these effect sizes are 1.52, 0.97, 0.60 and 0.36 for transition of grade K to 1, 1 to 2, 2 to 3 and 3 to 4 respectively. For math these effect sizes are 1.14, 1.03, 0.89 and 0.52 for transition of grade K to 1, 1 to 2, 2 to 3 and 3 to 4

4.6 OLS assumptions

The OLS, for simple and multiple regression models, has some assumptions that need to be met (Stock & Watson, 2020). The zero conditional mean assumption holds, because the *treatment* is assigned randomly to the classes, before the start of the interventions. The second assumption is that the observations need to be independent and identically distributed (i.i.d.). With including cluster-robust standard errors, this assumption is relaxed a little bit. Now

correlation within the classroom/center is allowed, but no correlation between the centers is allowed. This last part is not fully satisfied, since centers are also clustered within blocks. There will be some correlation between the centers, which may result in the standard errors being underestimated. This is a limitation of the study, but not a big problem, since regressions with cluster-robust standard errors at block level showed no big differences in the standard errors. Also the effect sizes and result being significant or insignificant stayed the same when the non-independence within the blocks/between the centers is taken into account. The third OLS assumption, that large outliers in X and Y are unlikely, holds since I checked the data on outliers and removed the ones with wrong values, resulting in no large outliers that OLS is sensitive to. On top of this, for multiple regression models, the assumption of no perfect multicollinearity between any of the explanatory variables holds, since no variables that have a perfect linear relationship with each other are used in the same regression. For categorical and dummy variables, I always leave one category/dummy out of the model.

5 Results

5.1 Regression results of the general effect

First of all, I want to test if there is a general effect of the treatment group as a whole and of the treatment groups separately on the TOPEL scores and on the 1st and 2nd grade reading and math scores of the child, with our OLS model given before.

Table 5.1 gives the results of four linear regressions, where the separate TOPEL scores are regressed on the categorical treatment variable. The results show that there is evidence of an effect of the RSL and BTL intervention program on the Phonological Awareness, Print Knowledge and Early Literacy Index score ($p < 0.01$, $p < 0.05$). These effects are all positive, with effect sizes between 0.332 and 0.657 for the RSL program and between 0.386 and 0.570 for the BTL program. According to Cohen (1988), these are on average medium effects. I can interpret these results as that receiving for example treatment RSL, will result on average in an 7.578 (0.474x 15.986 (=SD) point higher Early Literacy Index score than being assigned to the control group. The BELL program does not show any significant results ($p > 0.1$), meaning that there is no evidence of an effect of the BELL program on any of the TOPEL scores. Table 5.1 also shows that there is no evidence of an effect of any of the treatments on the Definitional Vocabulary score ($p > 0.1$). All these insignificant effects show much smaller effect sizes than the significant ones.

Table A4.1, in appendix A4, shows that there is also evidence of a positive effect of the treatment as a whole on the Phonological Awareness, Print Knowledge and Early Literacy Index score, with smaller effect sizes between 0.266 and 0.438 ($p < 0.05$, $p < 0.01$). From table

5.1 I saw that these effects are driven by the RSL and BTL intervention program. Again, there is no evidence of an effect of the treatment on the Definitional Vocabulary score (p > 0,01).

So, from these results, I can conclude that the RSL and BTL program can significantly improve the outcomes of the children in terms of language development and pre-literacy skills, since this is what the TOPEL score measures.

Table 5.1 Linear regression results of the relationship between the treatment groups and the TOPEL scores

	TOPEL scores			
Variable	**Definitional Vocabulary**	**Phonological Awareness**	**Print Knowledge**	**Early Literacy Index**
RSL	0.189	0.332**	0.657***	0.474***
	(2.487)	(2.147)	(1.976)	(2.300)
BELL	-0.063	0.082	0.092	0.040
	(2.472)	(2.010)	(2.205)	(2.425)
BTL	0.141	0.386***	0.570***	0.442***
	(2.596)	(2.106)	(2.073)	(2.487)
Constant	79.132	88.353	96.146	84.227
	(1.425)	(1.363)	(1.451)	(1.522)
Observations	999	999	999	999
R^2	0.009	0.026	0.086	0.046

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the four different TOPEL scores regressed on the categorical treatment variable; the dependent variables (TOPEL scores) are standardizes scores with a mean of 100 and a standard deviation of 15; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

Table 5.2 gives the results of four linear regressions, where the 1st and 2nd reading and math score are regressed on the categorical treatment variable. There is evidence of an effect of the BTL intervention program on the 1st grade reading and math score (p < 0.05, 0.01). These effects are positive, with a size of 0.534 for reading and 0.660 for math, seen as a medium and already more towards a big sized effect respectively (Cohen, 1988). However, the scores are still smaller than the average national annual gain for grade transition, which are 0.97 for reading and 1.03 for math (Hill et al. 2008). The RSL and BELL program does not show any significant results (p > 0.1), meaning that there is no evidence of an effect of the programs on any of the math and reading scores. Table 5.2 also shows that there is no evidence of an effect of any of the treatment programs on the 2nd grade reading and math score (p > 0.1). Again all these insignificant effects show much smaller effect sizes than the significant ones.

Table A4.2, in appendix A4, shows that there is also evidence of a positive effect of the treatment as a whole on the 1st grade reading and math score (p < 0.10). The effect sizes are

smaller, with 0.276 for reading and 0.319 for math 1st grade. From table 5.2 I saw that these effects are driven by the BTL intervention program. Again, there is no evidence of an effect of the treatment on the 2nd grade scores ($p > 0.01$).

So, from these results, I can conclude that only the BTL program can significantly improve the outcomes of the children in terms of 1st grade reading and math achievement. Receiving treatment BTL will result for example on average in an 23.991 (0.534 x 44.970 (=SD) point higher 1st grade reading score than being assigned to the control group.

Table 5.2 Linear regression results of the relationship between the treatment groups and the 1st and 2nd grade reading and math scores

	1st and 2nd grade achievement			
Variable	**Reading Score 1st grade**	**Math score 1st grade**	**Reading Score 2nd grade**	**Math score 2nd grade**
RSL	0.291	0.372	0.109	0.144
	(8.482)	(9.537)	(4.751)	(5.379)
BELL	0.041	-0.024	0.109	0.125
	(8.707)	(7.163)	(4.311)	(4.692)
BTL	0.534**	0.660***	0.161	0.132
	(10.370)	(8.684)	(4.486)	(4.986)
Constant	558.813	545.021	605.606	587.135
	(5.807)	(5.239)	(2.907)	(3.442)
Observations	259	259	958	958
R^2	0.038	0.064	0.004	0.004

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the 1st and 2nd grade reading and math scores regressed on the categorical treatment variable; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

5.2 Regression results of heterogeneous treatment effects by a child's gender

I want to test hypothesis 1, that states that Project Upgrade's pre-school literacy interventions will cause different effects on the TOPEL scores for men and women.

Table 5.3 gives the results of four OLS regressions, where the separate TOPEL scores are regressed on the categorical treatment variable, the male variable and on the interaction terms between gender and treatment. First of all, it is interesting to note that I don't find any significant results of the male variable ($p > 0.1$), which means that there is no evidence of a difference in TOPEL scores between men and women in the control group. Besides this, there is evidence of positive main effects for the same treatments as in table 5.1 from previous regressions ($p < 0.01$). However, we now have to interpret these as the effect of the interventions for a women, since interaction terms are included in the regression.

Focusing on the interactions, the results in table 5.3 show that there is evidence of a negative interaction effect with the RSL program on the Phonological Awareness and Early Literacy Index score ($p < 0.05$), with effect sizes of -0.450 and -0.371 respectively. This means that the effect of for example the RSL treatment on the Phonological Awareness score is 0.450 effect size lower for men than for women, resulting in an effect size of 0.573 - 0.450 = 0.123 for men and 0.573 for women. Table 5.3 also shows evidence of a negative interaction effect with the BELL intervention program on the Print Knowledge score, with a size of -0.338 ($p < 0.1$). So, the effect of BELL on the Print Knowledge is 0.338 effect size lower for men than for women. There are no other significant interaction terms present ($p > 0.1$), meaning that there is no evidence of heterogeneous treatment effects by gender for the RSL and BELL program on the rest of the TOPEL scores, and also not for the BTL program on all the TOPEL scores. The effect sizes of the insignificant interaction terms are smaller than the significant ones.

Table A4.3, in appendix A4, shows that there is also evidence of negative interaction effects with the treatment as a whole on the Phonological Awareness and Early Literacy Index score ($p < 0.10$). From table 5.3 I saw that these interactions effects are driven by the interaction with the RSL intervention program.

So, from the results, I can conclude that there is evidence of different effects for men and women of the RSL treatment on the Phonological and Literacy score and also of the BELL treatment on the Print score. In all these cases, the effects for women are bigger than for men. Besides this, no evidence of different effects for men and women of the RSL and BELL program on the rest of the TOPEL scores, and of the BTL program on all the TOPEL score, was found.

Table 5.3 Linear regression results of the relationship between the treatment groups and the TOPEL scores, with interaction terms between the treatment groups and gender included

	TOPEL scores			
Variable	**Definitional Vocabulary**	**Phonological. Awareness**	**Print Knowledge**	**Early Literacy Index**
Male	0.147	0.072	-0.020	0.086
	(1.759)	(2.177)	(1.923)	(2.017)
RSL	0.302	0.573***	0.774***	0.669***
	(3.095)	(2.909)	(2.395)	(2.881)
BELL	0.009	0.145	0.256	0.160
	(3.233)	(2.807)	(2.657)	(3.177)
BTL	0.217	0.535***	0.645***	0.566***
	(2.950)	(2.620)	(2.481)	(2.935)
Male*RSL	-0.218	-0.450**	-0.221	-0.371**
	(2.554)	(2.922)	(2.288)	(2.591)
Male*BELL	-0.140	-0.127	-0.338*	-0.243
	(3.002)	(2.844)	(2.696)	(3.067)
Male*BTL	-0.149	-0.304	-0.155	-0.251

	(2.749)	(3.052)	(2.689)	(2.941)
Constant	77.896	87.778	96.299	83.521
	(1.738)	(1.954)	(1.851)	(1.992)
Observations	999	999	999	999
R^2	0.011	0.038	0.099	0.054

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the four different TOPEL scores regressed on the categorical treatment variable, the dummy variable gender and on the interaction terms between gender and treatment dummies/categories; the dependent variables (TOPEL scores) are standardizes scores with a mean of 100 and a standard deviation of 15; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

Besides the hypothesis for the TOPEL score, I also want to test if Project Upgrades pre-school literacy interventions will cause different effects on 1st and 2nd grade reading and math achievement scores for men and women, which is hypothesis 2 of this paper.

Table 5.4 gives the results of four OLS regressions, where the separate reading and math scores are regressed on the categorical treatment variable, the male variable and on the interaction terms between gender and treatment. First of all, it is interesting to note that there is evidence of a positive main effect of men on the 2nd grade math score ($p < 0.01$), which means that the 2nd grade math score is on average 0.395 effect size higher for men than for women in the control group. Besides this, I have evidence of more positive main effects of the treatments ($p < 0.05$), now interpreted as the effects for a women, than in table 5.2 from previous regression.

Focusing on the interactions, the results in table 5.4 shows that there is evidence of a negative interaction effect with the RSL program on the 1st grade math and 2nd grade reading and math score ($p < 0.05, 0.01$), with effect sizes of -0.725, -0.695 and -0.507 respectively. There is also evidence of a negative interaction effect with the BELL intervention program on the 1st grade reading and 2nd grade reading and math score ($p < 0.1, 0.05$), with effect sizes of -0.701, -0.367 and -0.340 respectively. Because of these big effect sizes, the effects of the RSL and BELL program on the abovementioned grades are so much lower for man than for women, that it is even negative for men. For example, the effect size of the RSL program on the 2nd grade reading score is for men 0.475 – 0.695 = - 0.220, while the effect size for women is 0.475. The BTL intervention doesn't show any significant interaction terms ($p > 0.01$), meaning that there is no evidence of heterogeneous treatment effects by gender for the BTL program on the math and reading scores, which was also the case in the previous results with the TOPEL scores as outcome (see table 5.3).

Table A4.4, in appendix A4, shows that there is evidence of negative interaction effects with the treatment as a whole on the 2nd grade reading and math scores ($p < 0.05$), which are driven by the interaction with the RSL and BELL program, as seen in table 5.4.

So, from the results, I can conclude that there is evidence of different effects for men and women of the RSL treatment on the 1st grade math and 2nd grade reading and math score, and also of the BELL treatment on the 1st grade reading and 2nd grade reading and math score. In all these cases, the effects for women are bigger than for men. Besides this, no evidence of different effects for men and women of the BTL program on all the math and reading scores and of the RSL and BELL program on the rest of the math and reading scores, was found.

Table 5.4 Linear regression results of the relationship between the treatment groups and the 1st and 2nd grade reading and math scores, with interaction terms between the treatment groups and gender included

	1st and 2nd grade achievement			
Variable	**Reading Score 1st grade**	**Math score 1st grade**	**Reading Score 2nd grade**	**Math score 2nd grade**
Male	0.072	0.288	0.145	0.395***
	(11.632)	(8.910)	(4.919)	(5.257)
RSL	0.575*	1.781**	0.475***	0.403***
	(14.920)	(14.521)	(6.006)	(5.777)
BELL	0.383	0.330	0.287**	0.297**
	(13.697)	(8.425)	(5.329)	(5.436)
BTL	0.699*	1.454**	0.197	0.219
	(16.427)	(9.607)	(5.850)	(5.973)
Male*RSL	-0.515	-0.725*	-0.695***	-0.507***
	(19.379)	(16.633)	(7.043)	(7.640)
Male*BELL	-0.701*	-0.351	-0.367**	-0.340*
	(16.895)	(12.509)	(6.791)	(7.388)
Male*BTL	-0.314	0.039	-0.069	-1.689
	(18.644)	(12.997)	(7.038)	(7.255)
Constant	557.327	539.865	602.729	578.771
	(8.331)	(6.291)	(4.131)	(4.387)
Observations	259	259	958	958
R^2	0.069	0.083	0.026	0.019

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the 1st and 2nd grade reading and math scores regressed on the categorical treatment variable, the dummy variable gender and on the interaction terms between gender and treatment dummies/categories; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

5.3 Regression results of heterogeneous treatment effects by a child's home language

I want to test hypothesis 3, that states that Project Upgrade's pre-school literacy interventions will cause different effects on the TOPEL scores for children with just English as home language, than children speaking Spanish/Spanish and English or other languages at home.

Table 5.5 gives the results of four OLS regressions, where the separate TOPEL scores are regressed on the categorical treatment variable, the categorical language variable and on the

interaction terms between language and treatment. First of all, there is evidence of negative main effects of English and Spanish or Spanish only as home language (EngSpan/Span only) on the Definitional Vocabulary and Early Literacy index score ($p < 0.01$), and of a negative main effect of English and other, Spanish and other or just another home language (EngOth/SpanOth/Oth) on the Definitional Vocabulary score ($p < 0.05$). This means that on these scores children with EngSpan/Span only or EngOth/SpanOth/Oth as home language have lower effect sizes than children with English only (Eng only) as home language, in the control group. Besides this, we now interpret the main effects of the treatments as the effect of the interventions for children with Eng only as home language, since interaction terms are included in the regression.

Focusing on the interactions, table A4.5, in appendix A4, shows no evidence of heterogeneous treatment effects by home language for the treatment as a whole on all the TOPEL scores. However, the results in table 5.5 show that there is evidence of a positive interaction effect of the EngSpan/Span only home language with the RSL program on the Print Knowledge score ($p < 0.10$). The effect size is 0.443, which means that the effect of the RSL treatment on the Print score is 0.443 effect size higher for children with EngSpan/Span only as home language, than for children with Eng only as home language. Table 5.5 also shows evidence of a negative interaction effect of the EngOth/SpanOth/Oth home language with the BELL program on the Phonological Awareness score ($p < 0.05$), with an effect size of -0.905. This effect size is so big that the effect of the BELL program on the Phonological score is so much lower for the children with EngOth/SpanOth/Oth as home language, than for children with Eng only as home language, that it is even negative with an effect size of $0.514 – 0.905 = - 0.391$. There are no other significant interaction terms present ($p > 0.1$), meaning that there is no evidence of heterogeneous treatment effects by home language for the BTL program on all the TOPEL scores, and also not for the RSL and BELL program on the rest of the TOPEL scores.

From the results, I can conclude that there is evidence of different effects for children with home language EngSpan/Span only and Eng Only of the RSL treatment on the Print Knowledge score. Also different effects for children with home language EngOth/SpanOth/Oth and Eng Only for the BELL treatment on the Phonological Awareness score are present. Besides this, no evidence of different effects for people with different home languages of the BTL program on all the TOPEL scores and of the RSL and BELL program on the rest of the TOPEL scores, was found.

Table 5.5 Linear regression results of the relationship between the treatment groups and the TOPEL scores, with interaction terms between the treatment groups and home languages included

	TOPEL scores			
Variable	**Definitional Vocabulary**	**Phonological. Awareness**	**Print Knowledge**	**Early Literacy Index**
EngSpan/Span only	-0.707***	-0.142	-0.248	-0.467***
	(2.413)	(2.009)	(2.677)	(2.520)
EngOth/SpanOth/Oth	-0.367**	0.064	0.054	-0.115
	(2.932)	(3.060)	(2.389)	(2.675)
RSL	0.171	0.321**	0.331	0.331*
	(2.741)	(2.486)	(3.117)	(3.035)
BELL	0.294	0.514*	0.098	0.367
	(4.528)	(4.435)	(3.665)	(4.777)
BTL	0.369**	0.422*	0.695***	0.601***
	(2.558)	(3.512)	(3.598)	(3.563)
EngSpan/Span only	0.057	0.017	0.443*	0.209
*RSL	(3.650)	(3.169)	(3.382)	(3.529)
EngSpan/Span only	-0.293	-0.448	0.047	-0.283
*BELL	(5.186)	(4.802)	(4.336)	(5.346)
EngSpan/Span only	-0.259	-0.042	-0.119	-0.172
*BTL	(3.444)	(4.036)	(4.075)	(4.208)
EngOth/SpanOth/Oth	0.171	0.254	0.325	0.316
*RSL	(5.993)	(5.317)	(4.581)	(4.749)
EngOth/SpanOth/Oth	-0.307	-0.905**	-0.147	-0.559
*BELL	(7.751)	(5.916)	(4.610)	(5.992)
EngOth/SpanOth/Oth	0.182	0.106	-0.299	0.006
*BTL	(4.745)	(4.991)	(4.709)	(5.207)
Constant	87.542	89.806	98.639	89.486
	(2.130)	(1.713)	(2.422)	(2.240)
Observations	999	999	999	999
R^2	0.106	0.042	0.099	0.090

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the four different TOPEL scores regressed on the categorical treatment variable, the categorical home language variable and on the interaction terms between home language and treatment categories; the dependent variables (TOPEL scores) are standardizes scores with a mean of 100 and a standard deviation of 15; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

At last, I want to test hypothesis 4, that states that Project Upgrade's pre-school literacy interventions will cause different effects on reading and math achievement scores for children with just English as home language, than children speaking Spanish/Spanish and English or other languages at home.

Table 5.6 gives the results of four OLS regressions, where the separate reading and math scores are regressed on the categorical treatment variable, the categorical language variable and on the interaction terms between language and treatment. First of all, there is evidence of positive main effects of EngSpan/Span only on the 1st grade reading and math score and on the 2nd grade math score ($p < 0.05, 0.01$), which means that on these scores children in the control group with EngSpan/Span only as home language have higher effect sizes than those with Eng only as home language. Again the main effects of the treatments are now the effects of the interventions for children with Eng only as home language.

Focusing on the interactions, the results in table 5.6 show that there is evidence of a negative interaction effect of the BELL treatment with the EngSpan/Span only home language on the 1st grade reading score ($p < 0.10$), with an effect size of -0.655. This means that the effect of the BELL treatment on the 1st grade reading score is 0,655 effect size lower for children with EngSpan/Span only as home language, than for children with Eng only as home language, resulting in a negative overall effect of BELL on the 1st grade reading score of 0.465 – 0.655 = -0.190 effect size for the EngSpan/Span only children. For the BELL program, there is also evidence of positive interaction effects with the EngOth/SpanOth/Oth home language on the 1st and 2nd grade math score ($p < 0.05, 0.10$), with effect sizes of 0.542 and 0.823 respectively. So, now the effects of BELL on the 1st and 2nd grade math score are higher for children speaking EngOth/SpanOth/Oth languages at home than for the Eng only children. At last, there is evidence of a big negative interaction effect of the BTL treatment with EngOth/SpanOth/Oth on the 1st grade math score ($p < 0.10$), where the effect size is -1.135. This results also in a negative overall effect of the BELL program on the 1st grade math score for the children with EngOth/SpanOth/Oth as home language, with an effect size of 1.128 – 1.135 = -0.007.

Interestingly to notice from table 5.6, is that the interaction term of the RSL program with EngOth/SpanOth/Oth as home language has no observations for the effects on the 1st grade scores. There is no combination of children speaking EngOth/SpanOth/Oth at home, being in the RSL program and having 1st grade score available, in the data available.

Table A4.6, in appendix A4, shows that there is evidence of a positive interaction effect of the treatment as a whole with EngOth/SpanOth/Oth on the 2nd grade reading score ($p < 0.10$), which is quite surprisingly since no significant interactions effects of separate treatment groups on this score are found in table 5.6.

From the results, I can conclude that there is evidence of different effects for children with home language EngSpan/Span only and Eng only of the BELL treatment on the 1st grade math score. Also different effects for children with home language EngOth/SpanOth/Oth and Eng only of the BELL program on the 1st and 2nd grade math score, and for these children of the

BTL treatment on the 1st grade math score, are found. Besides this, no evidence of different effects for people with different home languages of the RSL program on all the math and reading scores and of the BTL and BELL program on the rest of the scores, was found.

Table 5.6 Linear regression results of the relationship between the treatment groups and the 1st and 2nd grade reading and math scores, with interaction terms between the treatment groups and home languages included

	1st and 2nd grade achievement			
Variable	**Reading Score 1st grade**	**Math score 1st grade**	**Reading Score 2nd grade**	**Math score 2nd grade**
EngSpan/Span only	0.554**	0.677***	0.193	0.535***
	(10.950)	(9.706)	(6.532)	(7.141)
EngOth/SpanOth/Oth	0.407	-0.047	-0.218	0.030
	(17.440)	(0.156)	(7.436)	(9.045)
RSL	0.271	0.279	-0.173	0.035
	(14.720)	(16.670)	(9.388)	(11.780)
BELL	0.465*	0.070	0.151	0.093
	(11.632)	(8.437)	(8.987)	(9.094)
BTL	0.888**	1.128***	-0.033	0.186
	(18.087)	(14.463)	(7.790)	(8.210)
EngSpan/Span only *RSL	-0.003	0.023	0.329	0.091
	(16.884)	(17.939)	(10.139)	(12.874)
EngSpan/Span only *BELL	-0.655*	-0.349	-0.103	-0.089
	(16.701)	(12.681)	(10.255)	(10.005)
EngSpan/Span only *BTL	-0.527	-0.747	0.195	-0.110
	(21.206)	(17.895)	(8.918)	(9.717)
EngOth/SpanOth/Oth *RSL	-	-	0.593	0.458
	-	-	(14.921)	(17.792)
EngOth/SpanOth/Oth *BELL	-0.465	0.542**	0.199	0.823*
	(26.567)	(10.349)	(15.430)	(18.430)
EngOth/SpanOth/Oth *BTL	-1.180	-1.135*	0.559	-0.048
	(44.729)	(24.156)	(13.695)	(16.205)
Constant	543.353	530.618	600.940	571.409
	(8.823)	(7.136)	(5.830)	(6.560)
Observations	259	259	958	958
R^2	0.076	0.133	0.025	0.051

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the 1st and 2nd grade reading and math scores regressed on the categorical treatment variable, the categorical variable home language and on the interaction terms between home language and treatment categories; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

5.4 Robustness checks

Robustness checks are done for the analyses with the categorical treatment variables, to see if the results stay approximately the same when including control variables to our OLS models. With staying the same, I mean that the sign and the significant/not significant state of the coefficient don't change. The results of these checks are shown in more detail in appendix A5, from which we can conclude that most of the results of our analyses are reliable and valid.

6 Discussion

6.1 Main findings

General effect

From our OLS regression results of the general effect, I found evidence of effects of the RSL and BTL intervention on the Phonological Awareness, Print Knowledge and Early Literacy Index score, with effect sizes between 0.332 and 0.657 for the RSL program and between 0.386 and 0.570 for the BTL program (see table 5.1), seen as a medium effect regarding to Cohen (1988). I find no evidence of an effect of the BELL program on any of the TOPEL scores, with effect sizes between -0.063 and 0.092. These result are in accordance with previous literature of Layzer et al. (2007), who also found evidence of effects of RSL and BTL on the same TOPEL scores, with effect sizes between 0.350 and 0.650 for RSL and between 0.440 and 0.600 for the BTL program, and no evidence of effects of BELL. The smaller effect sizes for BELL can be explained by the fact that the teachers have to create their own materials for this intervention (Layzer et al. 2009). According to another research of Layzer, Layzer and Wolf (2010) about the implementation of the curricula, the implementation rate of factors like vocabulary activities, oral language activities and comprehension when reading aloud, that will have an effect on the TOPEL scores, was way less in BELL than in the RSL and BTL program. This can be a result of the fact that teachers have to come up with their own material in the BELL program, which will eventually cause lower effect sizes on the TOPEL scores.

In the analysis for the math and reading scores, I found evidence of a positive effect of the BTL intervention program on the 1st grade reading and math score, with effect sizes of 0.534 and 0.660 (see table 5.2), seen as medium and big respectively (Cohen, 1988). The RSL and BELL program does not show any evidence of an effect of the programs on any of the math and reading scores, whit also way smaller effect sizes than BTL. The higher effects of BTL on reading can be explained by the research of Layzer and Wolf (2010), which shows that BTL implemented more reading activities that will help them answering the questions of the reading score. The higher effect on math can be explained by the fact that the children in BTL were engaged in independent activities were almost one-third of the class spent more than half of their time on high-value activities including math.

When comparing both the results of the TOPEL scores, math and reading grades, I can conclude that the BTL program has on average the biggest effect size and also most evidence of an effect. RSL shows quite similar results for the TOPEL scores, however especially their 1st grade math and reading scores are way lower. At last, the BELL program has the smallest effects on the child outcomes. Reasons for the differences in size can be found in the implementation of specific exercises that contribute to the understanding of questions used to assess the scores. I can also conclude that, since the comparable results from the read and math scores and the TOPEL scores are almost the same as previous research, the OLS method used in my analyses is a good method for analyzing the effects of Project Upgrade.

Heterogeneous treatment effects by a child's gender

From our OLS regression results with interaction terms of gender included, I found evidence of a negative interaction effect with the RSL program on the Phonological Awareness and Early Literacy Index score, with effect sizes of -0.450 and -0.371 respectively (see table 5.3). I also found evidence of a negative interaction effect with the BELL intervention program on the Print Knowledge score, with a size of -0.338. These effects are all towards medium size, according to Cohen (1988). The negative sign means that the interventions will have lower effects for men than for women. No evidence of heterogeneous treatment effects by gender for the RSL and BELL program on the rest of the TOPEL scores, and also not for the BTL program on all the TOPEL scores, are found, with smaller effect sizes than the significant scores. The differences in the size of the effects can be explained by the literature of Millard (1997), that says that there are a lot of differences between men and women in their use and experience of the literacy education in class. Interventions that have literacy components where women and men react differently on, will create bigger interaction effects, than interventions with literacy that men and women experience the same. From our results, I can conclude that the RSL and BELL program have specific things in their program that stimulate this differences, while the BTL program doesn't has this.

In the analysis for the math and reading scores, I found evidence of a negative interaction effect with the RSL program on the 1st grade math and 2nd grade reading and math, with effect sizes of -0.725, -0.695 and -0.507 respectively (see table 5.4). Also evidence of a negative interaction effect with the BELL intervention program on the 1st grade reading and 2nd grade reading and math score, with effect sizes of -0.701, -0.367 and -0.340 respectively, were found. At last, there is no evidence of heterogeneous treatment effects by gender for the BTL program on the math and reading scores, with also smaller effect sizes. The differences in effect sizes of reading can be explained by Millard (1997), who said that boys and girls differ in the development of how they read, with preferring other books and acting differently in reading

assignments. Men's attention during lessons in class is poorer than that of women (Zill & West, 2001) which results in men doing better in reading activities were they need less focus and when there is less distraction. The BTL program includes individual computer sessions were children are just doing one task by themselves, with not so much distraction, while the other programs have more broad and whole class activities around reading. This will result in men having less difference in BTL treatment effects with women, meaning a smaller negative effect size.

In the findings mentioned above, all interaction terms are negative, meaning that for every treatment group the effect on the TOPEL, math and reading score is lower for men than for women. This is in accordance with the literature of Heckman et al. (2010), which said that women have stronger treatment effects for education attainment early in life. Also Elango et al. (2016) agreed with this and said that women develop earlier in life and consequently will have more benefits of the interventions early in life.

So, while all the interaction effects show that men will have lower treatment effects than women, these differences in effects are for all scores the smallest for the BTL program. The RSL and BELL program show evidence of interaction effects, with bigger effect sizes.

Heterogeneous treatment effects by a child's home language

From our OLS regression results with interaction terms of home language included, I found evidence of a positive interaction effect of the EngSpan/Span only home language with the RSL program on the Print Knowledge score, with an effect size of 0.443. This result is in accordance with the previous findings of Layzer et al. (2007), who found that RSL and BTL together will have bigger effect sizes on the TOPEL scores for children with Spanish or Creole as home language, than for children speaking English at home. Also evidence of a negative interaction effect of the EngOth/SpanOth/Oth home language with the BELL program on the Phonological Awareness score, with an effect size of -0.905, was found. Scheele et al. (2010) gave as a reason that bilingual children will have disadvantages in both language skills, which will result in less effect on their academic achievements. At last, there was no evidence of heterogeneous treatment effects by home language for the BTL program on all the TOPEL scores, and also not for the RSL and BELL program on the rest of the TOPEL scores.

In the analysis for the math and reading scores, I found evidence of a negative interaction effect of the BELL treatment with the EngSpan/Span only home language on the 1st grade reading score, with an effect size of -0.655. For the BELL program, there is also evidence of positive interaction effects with the EngOth/SpanOth/Oth home language on the 1st and 2nd grade math score, with effect sizes of 0.542 and 0.823 respectively. At last, there is evidence

of a big negative interaction effect of the BTL treatment with EngOth/SpanOth/Oth on the 1st grade math score, where the effect size is -1.135. Kramsch (2014) gave as explanation for the negative interaction effects that people who speak different languages think very different and they use other linguistic forms, which will influence their cognitive process and makes it harder to understand things, leading to smaller treatment effects.

Unlike the results of the other parts, there is not a clear pattern of which treatments have heterogeneous treatment effects and of how big these effects are. Also the sign of the interaction effects is constantly different, so it is not the case that always the children with English only as home language or always one of the other home languages groups has bigger effect sizes of the treatment. It really depends on the combination of the intervention, the home language and the score as outcome.

6.2 Limitations

This study has also some limitations. One of these limitations is that with OLS only clustering at one level can be taken into account. So, clustering at block level is not done, which resulted in a partly violation of the second OLS assumption of independence, that may result in standard errors being underestimated. This is a limitation of the study, but not a big problem, since regressions with cluster-robust standard errors at block level showed no big differences in the standard errors, effect sizes and significance state. There is also a limitation of some of the coefficients of the effects being not completely robust when adding control variables. Because of this, the evidence of the EngSpan/Span only*BELL interaction effect on the 1st grade reading score and for the EngOth/SpanOth/Oth*BELL and the EngOth/SpanOth/Oth*BTL interaction effects on the 1st grade math scores needs to be interpreted cautiously, since these significant effects are not significant anymore when adding control variable.

7. Conclusion

In this paper I did research about how the effect of Project Upgrade's pre-school literacy interventions on the TOPEL scores and later reading and math achievement scores, differ based on child's sex and home language. With first looking at the sex of the child, I found evidence of different effects for men and women of the RSL program on the Phonological Awareness and Early Literacy Index score and also of the BELL treatment on the Print Knowledge score. There is also evidence of different effects for men and women of the RSL treatment on the 1st grade math and 2nd grade reading and math score, and also of the BELL treatment on the 1st grade reading and 2nd grade reading and math score. In all cases, the effects of the interventions are negative, which means bigger effects for women than for men. No evidence of different effects for men and women of the BTL program on all TOPEL scores,

and on the 1st and 2nd grade reading and math scores, are found. The difference in effect sizes can be explained by the theory of Millard (1997), which says that men and women use and experience literacy education in class very differently. He also explains that boys and girls differ in the development of how they read, in their preferences of books and they acting differently in reading assignments. The components of the intervention programs, and to which degree men a women react differently on them, will determine the size of the interaction effects.

When looking at the home language of the child, I found evidence of a positive interaction effect of the EngSpan/Span only home language with the RSL program on the Print Knowledge score. Also evidence of a negative interaction effect of the EngOth/SpanOth/Oth home language with the BELL program on the Phonological Awareness score were found. Besides this, there is evidence of a negative interaction effect of the BELL treatment with the EngSpan/Span only home language on the 1st grade reading score. At last, there is evidence of a big negative interaction effect of the BTL treatment with EngOth/SpanOth/Oth on the 1st grade math score. However, the evidence of the EngSpan/Span only*BELL interaction effect on the 1st grade reading score and for the EngOth/SpanOth/Oth*BELL and the EngOth/SpanOth/Oth*BTL interaction effects on the 1st grade math scores needs to be interpreted cautiously, since these significant effects are not significant anymore when adding control variable. There is no evidence of different effects for people with different home languages of the BTL program on all the TOPEL scores and of the RSL program on all the math and reading scores. There is no clear pattern of which treatments have heterogeneous treatment effects and of how big and in which direction these effects are. The negative interaction effects are in accordance with literature of Scheele et al. (2010) and Kramsch (2014

From all these results, I can conclude that there is evidence of the RSL and BELL intervention program having different effects for men and women on some of the TOPEL scores, 1st and 2nd grade reading and math scores mentioned before, with women always having higher effects than men. This is in accordance with hypothesis 1 and 2. I can also conclude that there is evidence of all the intervention programs having different effects, for children with just English as home language and children speaking Spanish/Spanish and English or other languages at home, on some of the TOPEL scores and 1st and 2nd grade reading and math scores mentioned before. There is not one home language group that has always higher effects, but this differs per combination of variables. These results are in accordance with hypothesis 3 and 4. For all the other combinations of interventions and outcome scores, there is no evidence of different treatment effects by gender and home language of the child, on the scores, which is not in accordance with our four hypotheses.

These results are important, because previous literature of Project Upgrade focused on the teachers and not on the children. With focusing on the difference in treatment effects by gender and home language of the child, I show which group of students benefits the most of which intervention program. The differences between student with different characteristics was not studied for the program, and thus with the results I will fill an important literature gap (Chin and Spector, 2019) and I will contribute to a growing literature about this topic (Heckman et al., 2017). With my results, policymakers will have the possibility to target the right interventions towards the right teachers and student, that will get the most benefit out of it. The result of the study will also help communities and states to allocate child care subsidy as effectively as possible. While the challenges in different cities may differ in degree, the result of the study could help many other communities too (Layzer et al., 2007), because of the representativeness of the sample.

Further research need to be done about the implementation of the intervention programs, and especially about what will be the effects on the implementation on the results, since I now only measure the intention to treat. In further research, there can also be looked for another model that has the benefits of the OLS model, but is able to cluster at more than one level. With this, I make sure that the standard errors are not underestimated. This al will result in knowing even better which intervention will be best for which type of person.

Reference list

Alwan, L.C., Craig, B.A. & McCabe, G. P. (2020), *The Practice of statistics for Business and Economics*. New York: W.H. Freeman and Company.

Arnett, J. (1989). Caregivers in day care centers: Does training matter? *Developmental Psychology*, 10, 541-552

Barbara Bush Foundation for family literacy. (2021). *National action plan for adult literacy.* Retrieved from https://www.barbarabush.org/wp-content/uploads/2021/11/BBF-National-Action-Plan-for-Adult-Literacy-2021.pdf

Buchmann, C., DiPrete, T.A., McDaniel, A. (2007) Gender Inequalities in Education. *Annual review of sociology. 34*, 319-337.

Child & Family Data Archive. (2011). *Project Upgrade in Miami-Dade County, Florida, 2003-2009 (ICPSR 31061)*. Retrieved from https://www.childandfamilydataarchive.org/cfda/archives/cfda/studies/31061/datadocumentation

ChildStats Forum. (2023). *Language Spoken at Home and Difficulty Speaking English.* Retrieved from https://www.childstats.gov/americaschildren/family5.asp

Chin, A., & Spector, A. (2019). *Heterogenous Treatment Effects in Early Language Literacy Interventions.* Retrieved from https://amspector100.github.io/assets/pdf/heterogenous_ell_treatments.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum.

Curriculum Associates. (2021). *i-Ready: Understanding Student Learning: Insights from Fall 2021.* Retrieved from https://www.curriculumassociates.com/-/media/mainsite/files/i-ready/iready-understanding-student-learning-paper-fall-results-2021.pdf

Dickinson, D.K., & Tabors, P.O. (2001). *Beginning literacy with language: Young children learning at home and school.* Baltimore, MD: Paul H. Brookes.

Diette, T. M., & Oyelere, R. U. (2014). Gender and race heterogeneity: The impact of students with limited english on native students' performance. *American Economic Review*, *104*(5), 412-417.

Elango, S., Garcia, J.L., Heckman, J.J. & Hojman, A. (2016). Early Childhood Education. *National Bureau of Economic Research, 2,* 235-297.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, *43*(2), 95-103.

Gee, J.P. (1989). Literacy, Discourse, and Linguistics: Introduction. *Journal of Education*, *171*(1), 1-176.

Goodson, B.D., Layzer, C., Smith, W.C., and Rimdzius, T. (2004). *Observation Measures of Language and Literacy Instruction (OMLIT)*. Cambridge, MA: Abt Associates Inc.

Grimm, K.J. (2008). Longitudinal Associations Between Reading and Mathematics Achievement. *Developmental Neuropsychology. 33*(3), 410-426.

Hardin, B.J., Peisner-Feinberg, E.S. & Weeks, S.W. (2005). *The Learning Accomplishment Profile Diagnostic (LAP-D): Examiner's Manual & Technical Report*. Retrieved from LAPD_Manual.pdf (kaplanco.com)

Heckman, J. J., Holland, M. L., Makino, K. K., Pinto, R., & Rosales-Rueda, M. (2017*).* An analysis of the memphis nurse-family partnership program. *National Bureau of Economic Research.* (No. w23610). Retrieved from https://www.nber.org/papers/w23610

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P. & Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative economics*, *1*(1), 1-46.

Hedges, L.V. & Nowell, A. (1995). Sex Differences in Mental Test Scores, Variability, and Numbers of High-Scoring Individuals. *Science*, *269*(5220), 41-45. doi: 10.1126/science.7604277

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology, 49*(1), 4–14.

Hill, C.J., Bloom, H.S., Black, A.R., Lipsey, M.W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives, 2*(3), 172-177.

Hyde, J. S., Fennema, E. & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological bulletin*, *107*(2), 139-155.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annuals of Applied Statistics., 7*(1), 443-470.

Kramsch, C. (2014). Language and culture. *AILA review*, *27*(1), 30-55.

Layzer, J.I., Price, C.S. (2010). Evaluation of Child Care Subsidy Strategies: Follow-up Study of the Participants in Project Upgrade in Miami-Dade. Retrieved from https://www.childandfamilydataarchive.org/cfda/archives/cfda/studies/31061/datadocumentation

Layzer J.I., Layzer J.C. & Wolf, A. (2010). Evaluation of Child Care Subsidy Strategies: Implementation of Three Language and Literacy Interventions in Project Upgrade. Retrieved from https://www.acf.hhs.gov/sites/default/files/documents/opre/pu_intervention.pdf

Layzer, J.I., Layzer J.C., Goodson B.D., Price C.S. (2007). Evaluation of Child Care Subsidy Strategies: Findings from Project Upgrade in Miami-Dade county. Retrieved from https://www.childandfamilydataarchive.org/cfda/archives/cfda/studies/31061/datadocumentation

Layzer, J.I., Layzer J.C., Goodson B.D., Price C.S. (2009). Evaluation of Child Care Subsidy Strategies: Findings from an Experimental Test of Three Language/ Literacy Interventions in Child Care Centers in Miami-Dade County. Retrieved from https://www.acf.hhs.gov/opre/report/evaluation-child-care-subsidy-strategies-findings-experimental-test-three-language

Legewie, J., & DiPrete, T. A. (2012). School context and the gender gap in educational achievement. *American sociological review*, *77*(3), 463-485.

Lonigan, C.J., Wagner, R.K., Torgesen, J.K., and Rashotte, C.A. (2002). *The Preschool Comprehensive Test of Phonological and Print Processing.* Tallahassee, FL: Florida State University.

Michols, A. & Schaffer, M. (2007). *Clustered Errors in Stata.* Retrieved from https://www.stata.com/meeting/13uk/nichols_crse.pdf

Millard, E. (1997). *Differently Literate: Boys, Girls and the Schooling of Literacy*. London: Routledge.

National Center for Education Statistics. (2011). *The Condition of Education 2011*. Retrieved from https://nces.ed.gov/pubs2011/2011033.pdf

Oller, D.K. & Eilers, R.E. (2002). *Language and Literacy in Bilingual Children.* Bristol: Multilingual Matters.

Pearson (2022). *Standford Achievement Test Series – Tenth Edition*. Retrieved from https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Academic-Learning/Stanford-Achievement-Test-Series-%7C-Tenth-Edition/p/100000415.html

Scheele, A.F., Leseman P.P.M. & Mayo A.Y. (2010). The home language environment of monolingual and bilingual children and their language proficiency. *Applied Psycholinguistics, 31*(1), 117-140.

Schleicher, A. (2019). *PISA 2018: Insights and Interpretations.* Retrieved from https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf

Stichting Lezen en Schrijven (2021). *Kennisblad Laaggeletterdheid in Nederland.* Retrieved from https://www.lezenenschrijven.nl/informatie-over-laaggeletterdheid-nederland

Stock, J.H. & Watson, M.W. (2020). *Introduction to Econometrics.* Harlow: Pearson Education Limited.

Sullivan, G. M. & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. *Journal of graduate medical education*, *4*(3), 279-282

Teachers of Tomorrow. (2023). *How to Become a Teacher In Florida in 2023.* Retrieved from https://www.teachersoftomorrow.org/blog/insights/how-to-become-a-teacher-in-florida/

Trzesniewski, K. H., Moffitt, T. E., Caspi, A., Taylor, A., & Maughan, B. (2006). Revisiting the Association Between Reading Achievement and Antisocial Behavior: New Evidence of an Environmental Explanation From a Twin Study. *Child development*, *77*(1), 72-88.

UCLA. (2021). *Regression with Stata Chapter 4 – Beyond OLS.* Retrieved from https://stats.oarc.ucla.edu/stata/webbooks/reg/chapter4/regressionwith-statachapter-4-beyond-ols-2/

U.S. Department of Education. (2019). *NAEP Report Card:2019 NAEP Reading Assessment*. Retrieved from https://www.nationsreportcard.gov/highlights/reading/2019/g12/

World literacy foundation. (2018). *The economic & social cost of illiteracy.* Retrieved from https://worldliteracyfoundation.org/wp-content/uploads/2021/07/TheEconomicSocialCostofIlliteracy-2.pdf

Zill, N. & West, J. (2001). *Entering Kindergarten: A Portrait of American Children When They Begin School. Findings from the Condition of Education, 2000*. Retrieved from https://eric.ed.gov/?id=ED448899

Appendix

A1. Detailed information of the dependent variables

TOPEL scores

Since the TOPEL measure was not finalized during the program, a pre-curser was used, called the Pre-CTOPP (Lonigan et al., 2002). In a seven-week period all children in the classes got tested with a 25 to 30 minutes test, by assessors who could provide instructions in both Spanish and English. The test was in English and consisted of specific questions regarding the different aspects, which were almost the same as the ones from the TOPEL scores. In 2006, the results were converted to TOPEL raw scores, and later also to standardized TOPEL scores. This was done so that the development status of the children can be compared to a national sample of children with the same age.

The quality of the TOPEL scores is good, since Abt. trainers were at the centers during the data collection, where they met with the assessors and observed the child assessment sessions. Besides this, the scores of the children in the different treatment and control groups were assessed at the same time.

1st and 2nd grade math and reading scores

The 999 children that were part of the sample were between 3 and 6 years old during spring 2005. So, part of the children were 5 years or older at September 1 in 2005, and thus eligible to enter kindergarten, and the other part was still too young to enter kindergarten (Layzer et al. 2010). This resulted in the children entering kindergarten and also elementary school at different years, and thus having math and reading scores of different grades in the measured schoolyears 2007/2008 and 2008/2009. Since I want to use a representative sample and generate representative average (heterogeneous) treatment effects, I decided to only use the 1st and 2nd grade reading and math scores in my analysis. This is because in the schoolyears 2007/2008 and 2008/2009, the 1st and 2nd grade reading and math scores are measured for all types of children; the children who do their school in normal pace, children who retained for one year and the children who are a year ahead on their peers. The 3rd grade scores were in these years only available for children at normal speed and for the children who are running ahead, which could give misleading results.

A2. Descriptive statistics

Table A2.1 Descriptive statistics of variables of interest of Project Upgrade

Variable	Obs. (1)	Mean (2)	Std. Dev. (3)	Min. (4)	Max. (5)
TOPEL scores:					
Definitional Vocabulary	999	80.155	17.307	55	115
Phonological Awareness	999	91.267	15.799	55	129
Print Knowledge	999	100.756	14.726	63	136
Early Literacy Index	999	87.800	16.463	47	125
1st grade math score	259	552.853	42.093	459	671
1st grade reading score	259	566.637	48.392	459	667
2nd grade math score	958	591.092	41.939	484	716
2nd grade reading score	958	609.108	39.251	476	729
Dummy treatment	999	0.705	0.456	0	1
Categorical treatment	999	2.596	1.138	1	4
Gender (child)	999	0.503	0.500	0	1
Home Language (child)	999	1.854	0.472	1	3
Age (child)	999	5.090	0.441	3.4	6.13
Highest education degree (teacher)	999	3.289	0.978	1	4
Training language (teacher)	999	0.506	0.500	0	1
Omlit scores:					
Support for oral language	999	53.932	9.980	39.43	71.29
Support for print knowledge	999	53.347	2.865	50.21	65.7
Support for print motivation	999	53.596	7.879	45.52	77.2
Literacy resources in classroom	999	50.667	4.863	38.28	60.63
Arnett scores:					
Positive subscale	999	50.293	8.862	27.55	62.46
Not Punitive subcale	999	47.231	6.428	13.99	54.23
Not Detached subscale	999	48.020	12.928	8.89	56.27
LapD scores (Class mean):					
Fine motor	999	39.673	4.190	24	56
Language Total	999	29.508	4.080	19.5	41
Cognitive Total	999	31.063	3.693	18.75	47

Notes: This table gives descriptive statistics of the variables of interest, of the whole sample. Column 1 shows the amount of observations and column 2 and 3 show the mean and standard deviation. Column 4 and 5 show the minimum and maximum value of each variable.

A3. Baseline Balance tests of separate groups

The balance tests are done by OLS regressions for the continuous variables and logistic regressions for the categorical variables, where clustering is taken into account by adding cluster-robust standard errors. For categorical variables with more than two categories, dummies were made to do the logistic regressions. Each test with a specific variable is done twice. One time with the dummy treatment variable as independent variable, to see if there is difference of a variable between the treatment group as a whole and the control group, and one time with the categorical treatment variable as independent variable, to see if there is difference of a variable between each treatment group separately and the control group. Also a joint test is done, to see if there is no significant difference of a variable between all the groups (treatment and control groups). The variables tested for balance are the ones mentioned in part 4.3. Table A3.1 gives the results of the tests done for the whole sample. The other five tables give results of balance tests done for parts of the sample I am looking at, which are; women only, men only, English only, English and Spanish or Spanish only, and English and other, Spanish and other, or other. All the categories of the categorical and dummy variables are presented separately and given in proportions.

For the whole sample, in table A3.1, there were no significant differences between the treatment and control groups, except from the *English & Spanish or Spanish only* category of home language and from the *Not Detached subscale*, when taking 0.5 as significant p-value. For the joint test, only the Not detached subscale show significant differences. Most of the results of parts of the sample show no significant differences between treatments and control groups. Only for the *LapD scores* and the *Not Detached subscale* are significant differences visible for women only and men only, respectively. When looking at the languages, for English only a joint difference is present for the *support for oral language* and *support for print motivation,* for the second home language category of table A3.5 this is *High school without CDA*, and for the third language category of table A3.6 this is *literacy resources*.

The variables that have significant differences for the joint test, except from the High school without CDA since this is just one category of a categorical variable, are added as control variables to the later presented regression, to check for robustness. As given in the tables in appendix A5, I can state that the differences at baseline of these few variables are not a big problem, since the signs and significance states of the coefficients mostly don't change when adding these variables as controls. I can conclude that random assignment was done successfully, which makes it possible to interpret the results of our further analyses as causal effects.

Table A3.1 Balance test of the whole sample at baseline

Variable	Control mean(1)	T-C diff.(2)	T1-C diff. (3)	T2-C diff.(4)	T3-C diff.(5)	P-value T1=0,T2= 0,T3=0 (6)
Child characteristics:						
Gender						
Male	0.512	-0.133	0.016	-0.027	-0.029	0.670
		(0.032)	(0.038)	(0.044)	(0.041)	
Female	0.488	0.133	-0.016	0.027	0.029	0.670
		(0.032)	(0.038)	(0.044)	(0.041)	
Home Language						
English only	0.244	-0.069	-0.028*	-0.122	-0.058	0.385
		(0.061)	(0.775)	(0.687)	(0.718)	
English & Spanish/ Spanish only	0.688	0.096	0.061	0.160**	0.066	0.260
		(0.700)	(0.085)	(0.078)	(0.870)	
English & other/ Spanish & other/ other	0.068	-0.027	-0.033	-0.038	-0.009	0.405
		(0.030)	(0.031)	(0.031)	(0.036)	
Age (in years)	5.053	0.053	0.015	0.049	0.094	0.387
		(0.049)	(0.057)	(0.065)	(0.059)	
Teacher characteristics:						
Highest education degree						
High School without CDA	0.058	0.012	-0.001	-0.028	0.065	0.500
		(0.048)	(0.053)	(0.050)	(0.070)	
High School with CDA	0.092	0.122	0.112	0.133	0.133	0.394
		(0.060)	(0.083)	(0.086)	(0.086)	
Some College with CDA	0.180	-0.031	-0.028	-0.053	-0.010	0.937
		(0.074)	(0.090)	(0.085)	(0.100)	
College Graduate	0.671	-0.103	-0.082	-0.038	-0.188	0.447
		(0.092)	(0.116)	(0.115)	(0.118)	
Training language						
Spanish	0.475	0.044	0.058	0.082	-0.008	0.865
		(0.096)	(0.120)	(0.122)	(0.121)	
English	0.525	-0.044	-0.058	-0.082	0.008	0.865
		(0.096)	(0.120)	(0.122)	(0.121)	
OMLIT scores:						
Support for oral language	54.702	-1.093	-2.242	-2.949	1.896	0.216
		(1.600)	(2.273)	(2.491)	(2.422)	
Support for print knowledge	52.969	0.537	0.772	-0.110	0.956	0.286
		(0.532)	(0.703)	(0.536)	(0.762)	
Support for print motivation	55.344	-2.480	-2.568	-3.626*	-1.243	0.267
		(1.654)	(2.035)	(1.959)	(1.886)	
Literacy resources in classroom	50.425	0.342	0.463	-0.625	1.197	0.458
		(1.028)	(1.242)	(1.275)	(1.176)	
Arnett scores:						
Positive subscale	51.953	-2.356	-1.416	-3.602	-2.024	0.439
		(1.624)	(2.022)	(2.264)	(2.056)	
Not Punitive subscale	47.804	-0.813	-0.721	-0.839	-0.878	0.906
		(1.120)	(1.504)	(1.454)	(1.327)	
Not Detached subscale	51.450	-4.868**	-1.760	-9.277***	-4.488	0.042**

		(2.007)	(2.555)	(3.029)	(3.134)	
LapD scores (Class mean):						
Fine motor	38.875	1.132	1.524*	1.697*	0.180	0.143
		(0.697)	(0.836)	(0.909)	(1.017)	
Language Total	28.934	0.814	1.046	1.844*	-0.445	0.108
		(0.754)	(0.933)	(0.980)	(0.937)	
Cognitive Total	30.762	0.427	1.032	1.189	-0.930	0.063
		(0.499)	(0.803)	(0.805)	(0.807)	
Sample size (N)						
Students	295	409	-64	-58	-59	
Classes/centers	47	57	-11	-14	-12	

Note: This table gives results of OLS and logistic regressions to test for balance at baseline in the full sample. Standard errors are between brackets. Column 1 shows the mean of the control group, column 2 shows the difference in mean between the treatment as a whole and the control group, with standard errors adjusted for clustering at class level. Column 3, 4 and 5 show the differences in mean between the treatments and the control group, with standard errors adjusted for clustering at class level. Column 6 shows the p-value of the joint test. T1 = RSL, T2 = BELL, T3 = BTL. All the categories of the categorical and dummy variables are presented separately and given in proportions. *p < 0.1, **p < 0.05, ***p < 0.01.

Table A3.2 Balance test of the women only sample at baseline

Variable	**Control mean(1)**	**T-C diff.(2)**	**T1-C diff. (3)**	**T2-C diff.(4)**	**T3-C diff.(5)**	**P-value T1=0,T2=0, T3=0 (6)**
Child characteristics:						
Home Language						
English only	0.250	-0.077	-0.048	-0.102	-0.078	0.614
		(0.066)	(0.081)	(0.080)	(0.080)	
English & Spanish/ Spanish only	0.653	0.126*	0.109	0.159*	0.110	0.366
		(0.079)	(0.093)	(0.092)	(0.096)	
English & other/ Spanish & other/ other	0.097	-0.049	-0.061	-0.056	-0.032	0.348
		(0.041)	(0.043)	(0.043)	(0.048)	
Age (in years)	5.037	0.064	0.050	0.051	0.089	0.507
		(0.055)	(0.070)	(0.072)	(0.059)	
Teacher characteristics:						
Highest education degree						
High School without CDA	0.063	0.028	0.020	-0.030	0.093	0.442
		(0.055)	(0.066)	(0.055)	(0.084)	
High School with CDA	0.104	0.104	0.068	0.124	0.116	0.544
		(0.066)	(0.083)	(0.097)	(0.093)	
Some College with CDA	0.215	-0.085	-0.777	-0.092	-0.084	0.732
		(0.082)	(0.097)	(0.093)	(0.098)	
College Graduate	0.625	-0.047	-0.010	-0.002	-0.125	0.734
		(0.100)	(0.123)	(0.127)	(0.126)	
Training language						
Spanish	0.472	0.024	0.078	0.052	-0.054	0.766
		(0.100)	(0.126)	(0.130)	(0.125)	
English	0.528	-0.078	-0.078	-0.052	0.054	0.766
		(0.126)	(0.126)	(0.130)	(0.125)	

OMLIT scores:						
Support for oral language	53.980	-0.087	-0.529	-2.123	2.344	0.479
		(2.005)	(2.667)	(2.667)	(2.583)	
Support for print knowledge	53.173	0.395	0.380	-0.230	1.035	0.381
		(0.578)	(0.690)	(0.600)	(0.845)	
Support for print motivation	55.003	-1.973	-1.647	-3.377*	-0.860	0.381
		(1.692)	(2.124)	(2.035)	(1.961)	
Literacy resources in classroom	50.050	0.856	0.891	0.109	1.573	0.190
		(0.984)	(1.227)	(1.285)	(1.076)	
Arnett scores:						
Positive subscale	52.201	-2.426	-2.157	-3.272	-1.821	0.515
		(1.676)	(2.091)	(2.322)	(2.019)	
Not Punitive subscale	47.276	-0.261	-0.529	0.235	-0.516	0.953
		(1.193)	(1.603)	(1.534)	(1.443)	
Not Detached subscale	51.357	-4.426**	-2.601	-7.293**	-3.188	0.135
		(2.087)	(2.959)	(3.260)	(3.134)	
LapD scores (Class mean):						
Fine motor	38.694	1.392*	2.088**	1.935**	0.228	0.044**
		(0.710)	(0.915)	(0.852)	(0.535)	
Language Total	28.628	1.546*	2.071**	2.450**	0.172	0.030**
		(0.788)	(0.991)	(0.979)	(1.032)	
Cognitive Total	30.474	0.982	1.862**	1.692**	-0.515	0.020**
		(0.651)	(0.887)	(0.825)	(0.822)	
Sample size (N)						
Students	144	209	-35	-22	-22	
Classes/centers	44	56	-10	-12	-10	

Note: This table gives results of OLS and logistic regressions to test for balance at baseline for only the women of the full sample. Standard errors are between brackets. Column 1 shows the mean of the control group, column 2 shows the difference in mean between the treatment as a whole and the control group, with standard errors adjusted for clustering at class level. Column 3, 4 and 5 show the differences in mean between the treatments and the control group, with standard errors adjusted for clustering at class level. Column 6 shows the p-value of the joint test. T1 = RSL, T2 = BELL, T3 = BTL. All the categories of the categorical and dummy variables are presented separately and given in proportions. *p < 0.1, **p < 0.05, ***p < 0.01.

Table A3.3 Balance test of the men only sample at baseline

Variable	**Control mean(1)**	**T-C diff.(2)**	**T1-C diff. (3)**	**T2-C diff.(4)**	**T3-C diff. (5)**	**P-value T1=0,T2= 0, T3=0(6)**
Child characteristics:						
Home Language						
English only	0.238	-0.062	-0.009	-0.143*	-0.037	0.297
		(0.067)	(0.085)	(0.072)	(0.082)	
English & Spanish/ Spanish only	0.722	0.067	0.016	0.165**	0.024	0.200
		(0.072)	(0.091)	(0.076)	(0.092)	
English & other/ Spanish & other/ other	0.040	-0.006	-0.007	-0.022	0.013	0.610
		(0.023)	(0.026)	(0.024)	(0.031)	
Age (in years)	5.068	0.043	-0.016	0.049	0.101	0.449
		(0.060)	(0.065)	(0.078)	(0.083)	

Teacher characteristics:						
Highest education degree						
High School without CDA	0.053	-0.005	-0.020	-0.027	0.035	0.556
		(0.042)	(0.043)	(0.046)	(0.058)	
High School with CDA	0.086	0.139*	0.152*	0.114	0.151*	0.264
		(0.060)	(0.091)	(0.082)	(0.088)	
Some College with CDA	0.146	0.022	0.018	-0.015	0.065	0.901
		(0.077)	(0.093)	(0.088)	(0.115)	
College Graduate	0.715	-0.157	-0.150	-0.072	-0.250**	0.235
		(0.094)	(0.120)	(0.118)	(0.123)	
Training language						
Spanish	0.477	0.064	0.040	0.114	0.041	0.849
		(0.102)	(0.126)	(0.128)	(0.129)	
English	0.523	-0.064	-0.040	-0.114	-0.041	0.850
		(0.102)	(0.126)	(0.128)	(0.129)	
OMLIT scores:						
Support for oral language	55.391	-2.067	-3.815	-3.749	1.501	0.078*
		(1.994)	(2.318)	(2.554)	(2.484)	
Support for print knowledge	52.774	0.668	1.135	-0.005	0.848	0.306
		(0.561)	(0.791)	(0.548)	(0.799)	
Support for print motivation	55.669	-2.973	-3.411	-3.855*	-1.614	0.230
		(1.754)	(2.117)	(2.092)	(1.968)	
Literacy resources in classroom	50.783	-0.155	0.057	-1.363	0.837	0.422
		(1.163)	(1.352)	(1.395)	(1.394)	
Arnett scores:						
Positive subscale	51.716	-2.298	-0.738	-3.980	-2.271	0.408
		(1.792)	(2.193)	(2.503)	(2.331)	
Not Punitive subscale	48.307	-1.342	-0.924	-1.922	-1.204	0.641
		(1.171)	(1.532)	(1.539)	(1.331)	
Not Detached subscale	51.540	-5.308**	-1.013	-9.309***	-5.869*	0.022**
		(2.173)	(2.401)	(3.288)	(3.326)	
LapD scores (Class mean):						
Fine motor	39.048	0.879	1.010	1.464	0.151	0.441
		(0.762)	(0.845)	(1.050)	(1.104)	
Language Total	29.226	0.095	0.111	1.233	-1.071	0.177
		(0.782)	(0.945)	(1.062)	(0.911)	
Cognitive Total	31.036	-0.116	0.274	0.685	-1.341	0.166
		(0.648)	(0.794)	(0.859)	(0.852)	
Sample size (N)						
Students	151	200	-29	-36	-37	
Classes/centers	43	58	-8	-12	-8	

Note: This table gives results of OLS and logistic regressions to test for balance at baseline for only the men of the full sample. Standard errors are between brackets. Column 1 shows the mean of the control group, column 2 shows the difference in mean between the treatment as a whole and the control group, with standard errors adjusted for clustering at class level. Column 3, 4 and 5 show the differences in mean between the treatments and the control group, with standard errors adjusted for clustering at class level. Column 6 shows the p-value of the joint test. T1 = RSL, T2 = BELL, T3 = BTL. All the categories of the categorical and dummy variables are presented separately and given in proportions. *p < 0.1, **p < 0.05, ***p < 0.01.

Table A3.4 Balance test of the English only sample at baseline

Variable	Control mean(1)	T-C diff.(2)	T1-C diff. (3)	T2-C diff.(4)	T3-C diff.(5)	P-value T1=0,T2= 0, T3=0(6)
Child characteristics:						
Gender						
Male	0.500	0.004	0.060	-0.121	0.023	0.523
		(0.070)	(0.074)	(0.120)	(0.102)	
Female	0.500	-0.004	-0.060	0.121	-0.023	0.523
		(0.070)	(0.074)	(0.120)	(0.102)	
Age (in years)	4.896	0.136	0.105	0.078	0.209*	0.336
		(0.090)	(0.106)	(0.106)	(0.114)	
Teacher characteristics:						
Highest education degree						
High School without CDA	0.167	-0.102	-0.047	-0.132	-0.144	0.311
		(0.118)	(0.155)	(0.113)	(0.110)	
High School with CDA	0.083	0.388***	0.337**	0.399	0.439**	0.062*
		(0.113)	(0.170)	(0.204)	(0.163)	
Some College with CDA	0.278	-0.131	-0.078	-0.243*	-0.119	0.320
		(0.155)	(0.185)	(0.143)	(0.186)	
College Graduate	0.472	-0.155	-0.212	-0.024	-0.177	0.594
		(0.158)	(0.173)	(0.230)	(0.189)	
Training language						
Spanish	0.125	-0.011	0.015	0.013	-0.057	0.780
		(0.086)	(0.110)	(0.110)	(0.092)	
English	0.875	0.011	-0.015	-0.013	0.057	0.870
		(0.086)	(0.110)	(0.110)	(0.092)	
OMLIT scores:						
Support for oral language	55.416	-2.661	-7.458**	-6.063	5.034	0.007***
		(3.097)	(3.117)	(4.303)	(3.985)	
Support for print knowledge	52.240	1.389*	0.622	0.253	3.011*	0.260
		(0.830)	(0.843)	(0.857)	(1.519)	
Support for print motivation	58.473	-6.870**	-9.694***	-6.718*	-3.760	0.014**
		(2.953)	(3.146)	(3.628)	(3.349)	
Literacy resources in classroom	49.705	1.019	2.628	1.617	-1.202	0.276
		(2.414)	(2.868)	(2.576)	(2.520)	
Arnett scores:						
Positive subscale	50.747	-1.315	-0.754	3.327	-5.011	0.069*
		(2.904)	(3.114)	(2.595)	(4.438)	
Not Punitive subscale	46.956	-0.669	-0.337	2.605	-3.205	0.177
		(1.916)	(2.867)	(2.351)	(2.161)	
Not Detached subscale	52.884	-6.824**	-2.979	-3.970*	-13.076**	0.065*
		(2.733)	2.000	(2.232)	(6.150)	
LapD scores (Class mean):						
Fine motor	37.549	2.097*	3.002***	1.188	1.667	0.074*
		(1.148)	(1.117)	(0.929)	(2.339)	
Language Total	29.190	0.562	0.663	0.515	0.478	0.964
		(1.189)	(1.384)	(1.187)	(1.924)	

Cognitive Total	30.329	0.199	1.760	-0.605	-1.044	0.399
		(1.241)	(1.485)	(1.114)	(2.151)	
Sample size (N)						
Students	72	51	-22	-43	-28	
Classes/centers	21	24	-3	-9	-6	

Note: This table gives results of OLS and logistic regressions to test for balance at baseline for only the children with English as home language of the full sample. Standard errors are between brackets. Column 1 shows the mean of the control group, column 2 shows the difference in mean between the treatment as a whole and the control group, with standard errors adjusted for clustering at class level. Column 3, 4 and 5 show the differences in mean between the treatments and the control group, with standard errors adjusted for clustering at class level. Column 6 shows the p-value of the joint test. T1 = RSL, T2 = BELL, T3 = BTL. All the categories of the categorical and dummy variables are presented separately and given in proportions. *p < 0.1, **p < 0.05, ***p < 0.01.

Table A3.5 Balance test of the English and Spanish or Spanish only sample at baseline

Variable	**Control mean(1)**	**T-C diff.(2)**	**T1-C diff.(3)**	**T2-C diff.(4)**	**T3-C diff. (5)**	**P-value T1=0,T2 =0, T3=0 (6)**
Child characteristics:						
Gender						
Male	0.537	-0.035	-0.017	-0.029	-0.059	0.689
		(0.038)	(0.048)	(0.048)	(0.050)	
Female	0.463	0.035	0.017	0.029	0.059	0.689
		(0.038)	(0.048)	(0.048)	(0.050)	
Age (in years)	5.119	0.008	-0.038	0.016	0.045	0.595
		(0.049)	(0.057)	(0.068)	(0.061)	
Teacher characteristics:						
Highest education degree						
High School without CDA	0.005	0.064**	0.030	0.025	0.141***	0.017**
		(0.028)	(0.025)	(0.030)	(0.071)	
High School with CDA	0.074	0.071	0.065	0.095	0.050	0.660
		(0.054)	(0.076)	(0.077)	(0.068)	
Some College with CDA	0.128	0.022	0.005	0.006	0.057	0.950
		(0.073)	(0.089)	(0.087)	(0.109)	
College Graduate	0.793	-0.157	-0.099	-0.126	-0.248**	0.264
		(0.088)	(0.112)	(0.113)	(0.123)	
Training language						
Spanish	0.640	-0.012	0.019	-0.009	-0.045	0.971
		(0.101)	(0.124)	(0.127)	(0.131)	
English	0.360	0.012	-0.019	0.009	0.045	0.971
		(0.101)	(0.124)	(0.127)	(0.131)	
OMLIT scores:						
Support for oral language	54.768	-0.973	-0.872	-2.701	0.881	0.643
		(2.136)	(2.498)	(2.767)	(2.698)	
Support for print knowledge	53.260	0.197	0.727	-0.365	0.317	0.375
		(0.655)	(0.818)	(0.655)	(0.888)	
Support for print motivation	54.568	-1.575	-0.664	-3.011	-0.838	0.467
		(1.728)	(2.181)	(2.043)	(2.003)	

Literacy resources in classroom	50.986	-0.191	-0.653	-1.404	1.629	0.055
		(0.958)	(-1.181)	(1.254)	(1.074)	
Arnett scores:						
Positive subscale	52.611	-2.808	-1.728	-4.979*	-1.405	0.272
		(1.817)	(2.334)	(2.514)	(2.105)	
Not Punitive subscale	48.285	-1.018	-0.891	-1.602	-0.481	0.781
		(1.292)	(1.603)	(1.606)	(1.499)	
Not Detached subscale	51.234	-4.437*	-1.538	-8.991**	-2.114	0.081*
		(2.323)	(3.174)	(3.435)	(3.230)	
LapD scores (Class mean):						
Fine motor	39.407	0.750	1.000	1.516	-0.359	0.296
		(0.773)	(0.959)	(1.021)	1.039	
Language Total	29.043	0.740	0.970	1.973*	-0.878	0.036
		(0.844)	(1.078)	(1.097)	(0.908)	
Cognitive Total	31.048	0.363	0.659	1.390*	-1.083	0.135
		(0.642)	(0.865)	(0.838)	(0.720)	
Sample size (N)						
Students	203	322	-30	-2	-25	
Classes/centers	40	55	-8	-9	-8	

Note: This table gives results of OLS and logistic regressions to test for balance at baseline for only the children with English and Spanish or Spanish only as home language of the full sample. Standard errors are between brackets. Column 1 shows the mean of the control group, column 2 shows the difference in mean between the treatment as a whole and the control group, with standard errors adjusted for clustering at class level. Column 3, 4 and 5 show the differences in mean between the treatments and the control group, with standard errors adjusted for clustering at class level. Column 6 shows the p-value of the joint test. T1 = RSL, T2 = BELL, T3 = BTL. All the categories of the categorical and dummy variables are presented separately and given in proportions. *p < 0.1, **p < 0.05, ***p < 0.01.

Table A3.5 shows no results for the differences between BELL (=T2) and the control group for the education degree category *High School without CDA*, since there are in this specific sample no teacher in de BELL group that have this degree as highest education degree. This is also the case with the *Some College with CDA* category for the teachers in the BTL program (=T3). This probably happens because of the small sample of this subgroup.

Table A3.6 Balance test of the English and other, Spanish and other, or other sample at baseline

Variable	**Control mean(1)**	**T-C diff.(2)**	**T1-C diff (3).**	**T2-C diff.(4)**	**T3-C diff.(5)**	**P-value T1=0,T2= 0, T3=0(6)**
Child characteristics:						
Gender						
Male	0.300	0.114	0.200	-0.014	0.129	0.513
		(0.101)	(0.170)	(0.132)	(0.131)	
Female	0.700	-0.114	-0.200	0.014	-0.129	0.513
		(0.101)	(0.170)	(0.132)	(0.131)	
Age (in years)	4.952	0.070	0.264	-0.234	0.111	0.330
		(0.167)	(0.179)	(0.332)	(0.161)	
Teacher characteristics:						

Highest education degree						
High School without CDA	0.200	-0.097	-0.075	-	-0.057	0.936
		(0.192)	(0.220)	-	(0.211)	
High School with CDA	0.300	0.114	-0.050	-0.014	0.271	0.640
		(0.271)	(0.314)	(0.300)	(0.300)	
Some College with CDA	0.350	-0.212	-0.100	-0.064	-	0.934
		(0.231)	(0.273)	(0.322)	-	
College Graduate	0.150	0.195	0.255	0.279	0.136	0.629
		(0.149)	(0.212)	(0.261)	(0.179)	
Training language						
Spanish	0.050	0.088	0.200	0.093	0.021	0.556
		(0.087)	(0.171)	(0.151)	(0.091)	
English	0.950	-0.088	-0.200	-0.093	-0.021	0.556
		(0.087)	(0.171)	(0.151)	(0.091)	
OMLIT scores:						
Support for oral language	51.472	2.231	-1.890	1.209	5.098	0.507
		(5.895)	(1.209)	(7.547)	(6.327)	
Support for print knowledge	52.633	1.261**	1.283	0.678	1.540**	0.151
		(0.599)	(1.452)	(0.577)	(0.675)	
Support for print motivation	51.951	3.788	1.406	4.217	4.936	0.642
		(4.193)	(5.015)	(8.308)	(4.291)	
Literacy resources in classroom	47.329	3.099	6.541***	2.437	1.462	0.021**
		(2.231)	(2.092)	(2.953)	(2.924)	
Arnett scores:						
Positive subscale	49.615	-3.238	-3.151	-4.327	-2.743	0.772
		(3.791)	(4.447)	(4.167)	(5.330)	
Not Punitive subscale	45.972	-1.268	-2.719	-1.672	-0.238	0.913
		(2.034)	(4.222)	(4.549)	(2.062)	
Not Detached subscale	48.485	-3.770	-0.253	-2.370	-6.480	0.807
		(5.838)	(6.312)	(7.166)	(7.499)	
LapD scores (Class mean):						
Fine motor	38.248	0.441	1.043	-0.160	0.398	0.829
		(1.345)	(1.135)	(2.559)	(2.200)	
Language Total	26.908	2.177	3.130*	1.471	1.986	0.331
		(1.819)	(1.672)	(2.516)	(2.662)	
Cognitive Total	29.415	0.341	2.426*	-2.246	0.443	0.158
		(1.525)	(1.314)	(3.011)	(1.880)	
Sample size (N)						
Students	20	9	-12	-13	-6	
Classes/centers	8	13	-1	-3	1	

Note: This table gives results of OLS and logistic regressions to test for balance at baseline for only the children with English and another, Spanish and another or just another language as home language of the full sample. Standard errors are between brackets. Column 1 shows the mean of the control group, column 2 shows the difference in mean between the treatment as a whole and the control group, with standard errors adjusted for clustering at class level. Column 3, 4 and 5 show the differences in mean between the treatments and the control group, with standard errors adjusted for clustering at class level. Column 6 shows the p-value of the joint test. T1 = RSL, T2 = BELL, T3 = BTL. All the categories of the categorical and dummy variables are presented separately and given in proportions. *p < 0.1, **p < 0.05, ***p < 0.01.

A4. Results for the dummy treatment variable

The tables below show the results of the analyses done with the dummy treatment variable, instead of with the categorical treatment variable of the result tables shown in the main text.

Table A4.1 Linear regression results of the relationship between the treatment group as a whole and the TOPEL scores

	TOPEL scores			
Variable	**Definitional Vocabulary**	**Phonological. Awareness**	**Print Knowledge**	**Early Literacy Index**
Treatment	0.088	0.266**	0.438***	0.317***
	(1.872)	(1.654)	(1.726)	(1.890)
Constant	79.132	88.353	96.146	84.227
	(1.423)	(1.363)	(1.449)	(1.521)
Observations	999	999	999	999
R^2	0.002	0.014	0.041	0.020

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the four different TOPEL scores regressed on the dummy treatment variable; the dependent variables (TOPEL scores) are standardizes scores with a mean of 100 and a standard deviation of 15; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A4.2 Linear regression results of the relationship between the treatment group as a whole and the 1st and 2nd grade reading and math scores.

	1st and 2nd grade achievement			
Variable	**Reading Score 1st grade**	**Math score 1st grade**	**Reading Score 2nd grade**	**Math score 2nd grade**
Treatment	0.276*	0.319*	0.126	0.133
	(7.280)	(6.747)	(3.524)	(4.032)
Constant	558.813	545.021	605.606	587.135
	(5.784)	(5.219)	(2.904)	(3.439)
Observations	259	259	958	958
R^2	0.016	0.021	0.004	0.004

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the 1st and 2nd grade reading and math scores regressed on the dummy treatment variable; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A4.3 Linear regression results of the relationship between the treatment group as a whole and the TOPEL scores, with an interaction term between the treatment group and gender included

	TOPEL scores			
Variable	**Definitional Vocabulary**	**Phonological. Awareness**	**Print Knowledge**	**Early Literacy Index**
Male	0.147	0.072	-0.020	0.086
	(1.755)	(2.172)	(1.919)	(2.013)
Treatment	0.171	0.412***	0.550***	0.457***

	(2.288)	(2.285)	(2.126)	(2.408)
Male*Treatment	-0.163	-0.292*	-0.226	-0.279*
	(2.152)	(2.474)	(2.198)	(2.362)
Constant	77.896	87.778	96.299	83.521
	(1.734)	(1.950)	(1.848)	(1.988)
Observations	999	999	999	999
R^2	0.003	0.023	0.052	0.026

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the four different TOPEL scores regressed on the dummy treatment variable, the dummy variable gender and on the interaction terms between gender and treatment dummies; the dependent variables (TOPEL scores) are standardizes scores with a mean of 100 and a standard deviation of 15; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A4.4 Linear regression results of the relationship between the treatment group as a whole and the 1st and 2nd grade reading and math scores, with an interaction term between the treatment group and gender included

	1st and 2nd grade achievement			
Variable	**Reading Score 1st grade**	**Math score 1st grade**	**Reading Score 2nd grade**	**Math score 2nd grade**
Male	0.072	0.288	0.145	0.395***
	(11.541)	(8.840)	(4.909)	(5.246)
Treatment	0.537**	0.468**	0.314**	0.303**
	(11.057)	(8.414)	(4.734)	(4.865)
Male*Treatment	-0.498	-0.320	-0.375**	-0.337**
	(14.181)	(11.194)	(5.774)	(6.115)
Constant	557.327	539.865	602.729	578.771
	(8.265)	(6.242)	(4.123)	(4.378)
Observations	259	259	985	958
R^2	0.041	0.027	0.014	0.016

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the 1st and 2nd grade reading and math scores regressed on the dummy treatment variable, the dummy variable gender and on the interaction terms between gender and treatment dummies; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A4.5 Linear regression results of the relationship between the treatment group as a whole and the TOPEL scores, with interaction terms between the treatment group and home languages included

	TOPEL scores			
Variable	**Definitional Vocabulary**	**Phonological. Awareness**	**Print Knowledge**	**Early Literacy Index**
EngSpan/Span only	-0.707***	-0.142	-0.238	-0.467
	(2.506)	(2.003)	(2.669)	(2.512)
EngOth/SpanOth/Oth	-0.367**	0.064	0.052	-0.115
	(2.923)	(3.051)	(2.381)	(2.667)
Treatment	0.271*	0.402***	0.390**	0.436
	(2.476)	(2.377)	(2.872)	(2.785)

EngSpan/Span only *Treatment	-0.164 (2.976)	-0.150 (2.757)	0.072 (3.217)	-0.098 (3.186)
EngOth/SpanOth/Oth *Treatment	0.086 (4.157)	-0.084 (4.282)	-0.044 (3.515)	-0.011 (3.993)
Constant	87.542 (2.123)	89.806 (1.708)	98.639 (2.414)	89.486 (2.233)
Observations	999	999	999	999
R^2	0.100	0.026	0.049	0.065

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the four different TOPEL scores regressed on the dummy treatment variable, the categorical home language variable and on the interaction terms between home language and treatment categories/dummies; the dependent variables (TOPEL scores) are standardizes scores with a mean of 100 and a standard deviation of 15; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A4.6 Linear regression results of the relationship between the treatment group as a whole and the 1st and 2nd grade reading and math scores, with interaction terms between the treatment group and home languages included

	1st and 2nd grade achievement			
Variable	**Reading Score 1st grade**	**Math score 1st grade**	**Reading Score 2nd grade**	**Math score 2nd grade**
EngSpan/Span only	0.554** (10.841)	0.677*** (9.610)	0.121 (6.512)	0.535*** (7.119)
EngOth/SpanOth/Oth	0.407 (17.267)	-0.047 (9.065)	-0.218 (7.412)	0.030 (9.016)
Treatment	0.494* (11.654)	0.445 (11.521)	-0.049 (7.068)	0.102 (8.122)
EngSpan/Span only *Treatment	-0.371 (14.386)	-0.349 (13.832)	0.167 (7.871)	-0.037 (8.890)
EngOth/SpanOth/Oth *Treatment	-0.611 (26.520)	-0.091 (16.390)	0.501* (10.616)	0.332 (13.604)
Constant	543.353 (8.735)	530.618 (7.065)	600.939 (5.812)	571.409 (6.539)
Observations	259	259	958	958
R^2	0.042	0.071	0.021	0.046

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of an OLS regression of the 1st and 2nd grade reading and math scores regressed on the dummy treatment variable, the categorical variable home language and on the interaction terms between home language and treatment categories/dummies; results are given in effect sizes, calculated by dividing the estimated impact of the independent variable through the standard deviation of the control group; *p < 0.1, **p < 0.05, ***p < 0.01.

A5. Robustness checks

Two separate tables of each outcome variable are given; one of the normal regression and one of the regressions with interaction terms included. The tables first shows the results without control variables, as seen before in the result part, than some control variables that differ between groups at baseline are added, and at last all baseline variables are included as controls. The results are not given in effect sizes, since the check is about the changes of the coefficients and not about interpreting the values. For the regressions, except from table A5.1 that is given as an example, only the relevant variables without the controls are included in the tables.

When looking at table A5.1 until A5.8, with the results of the normal OLS regressions, I see that for most of the outcome variables no changes of sign and significance of the coefficients are visible, when control variables are included. Sometimes there is a change when the first control variables are added, but these changes are reversed when the other control variables are added. Only a sign change of the coefficient of the effect of BELL on the Definitional Vocabulary score and a significance change of the effect size of BTL on the 2nd grade reading score are visible. When looking at table A5.9 until A5.16, with the results of the OLS regression including the interaction terms, I see again that for most of the outcome variables no changes of sign and significance of the coefficients are visible, when control variables are added. For the interactions with gender, there is only a significance change of the main effect of RSL on the Definitional Vocabulary score, of the main effect of BELL and BTL on the 1st grade reading score and at last a significance change of the main effect of BTL on the 1st grade math score. For the interactions with home language, there are some more changes of the coefficients' sign and significance state, for some main effects and also for some of the interaction effects. However, also here most of the effects show no changes, when control variables are included.

Since, some of the coefficients of the effects do change when control variables are added, you have to be cautious with these results. This holds especially for the main effect of BTL on the 1st grade reading and math score in table 5.4, since adding control variables result in no longer significant findings. This also holds for the main effect of EngOth/SpanOth/Oth on the Definitional Vocabulary and for the main effect of RSL on the Early Literacy Index in table 5.5, for the BTL main effect on the 1st grade reading and math score in table 5.6, for the EngSpan/Span only*BELL interaction effect on the 1st grade reading score and for the EngOth/SpanOth/Oth*BELL and the EngOth/SpanOth/Oth*BTL interaction effects on the 1st grade math scores. All of these were significant before, but not anymore when control variables were added.

However, besides some changes in the coefficients, most of the results show no changes when control variables are included. From this I can conclude that most of the results of our analyses are reliable and valid.

Table A5.1 Linear regression results for the relationship between the treatment groups and the Definitional Vocabulary TOPEL score, including control variables

	Definitional Vocabulary		
Variable	**(1)**	**(2)**	**(3)**
RSL	3.115	2.877	3.256
	(2.487)	(2.426)	(2.095)
BELL	-1.035	-1.115	0.405
	(2.472)	(2.617)	(2.286)
BTL	2.321	3.439	3.870
	(2.596)	(2.555)	(2.386)
Fine motor		-0.119	-0.508
		(0.365)	(0.307)
Language Total		0.383	0.307
		(0.343)	(0.279)
Cognitive Total		0.085	-0.036
		(0.497)	(0.337)
Support for oral language		-0.148	-0.154
		(0.164)	(0.146)
Support for print motivation		0.087	0.090
		(0.205)	(0.168)
Literacy resources in classroom		-0.068	-0.339
		(0.214)	(0.186)
Not Detached subscale		0.083	0.165**
		(0.087)	(0.081)
Positive subscale			-0.276**
			(0.135)
Not Punitive subcale			0.170
			(0.110)
Support for print knowledge			0.368
			(0.263)
Gender			0.761
			(0.939)
Child Home Language			-5.523***
			(1.214)
Age			-3.600***
			(1.352)
Highest education degree			1.017
			(0.884)
Training language			-8.657***
			(1.717)
Constant	79.132	72.507	85.233
	(1.425)	(12.641)	(17.942)
Observations	999	999	999
R^2	0.009	0.020	0.154

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the Definitional Vocabulary TOPEL score regressed on the categorical treatment variable, without control variables, as in table 5.1. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.2 Linear regression results for the relationship between the treatment groups and the Phonological Awareness TOPEL score, including control variables

	Phonological Awareness		
Variable	**(1)**	**(2)**	**(3)**
RSL	5.163**	4.621**	4.864**
	(2.147)	(2.075)	(1.933)
BELL	1.276	1.010	1.830
	(2.020)	(2.078)	(1.797)
BTL	6.003***	7.219***	7.132***
	(2.106)	(2.071)	(1.932)
Constant	88.353	70.435	46.216
	(1.364)	(10.373)	(14.537)
Observations	999	999	999
R^2	0.026	0.047	0.121

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the Phonological Awareness TOPEL score regressed on the categorical treatment variable, without control variables, as in table 5.1. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.3 Linear regression results for the relationship between the treatment groups and the Print Knowledge TOPEL score, including control variables

	Print Knowledge		
Variable	**(1)**	**(2)**	**(3)**
RSL	9.815***	9.089***	9.037***
	(1.976)	(1.961)	(2.001)
BELL	1.377	0.776	1.167
	(2.205)	(2.081)	(2.013)
BTL	8.524***	9.033***	8.676***
	(2.073)	(2.044)	(2.044)
Constant	96.146	74.167	48.396
	(1.451)	(11.358)	(15.273)
Observations	999	999	999
R^2	0.086	0.107	0.159

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the Print Knowledge TOPEL score regressed on the categorical treatment variable, without control variables, as in table 5.1. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.4 Linear regression results for the relationship between the treatment groups and the Early Literacy Index TOPEL score, including control variables

	Early Literacy Index		
Variable	**(1)**	**(2)**	**(3)**
RSL	7.578***	6.950***	7.193***
	(2.300)	(2.203)	(2.060)
BELL	0.638	0.252	1.403
	(2.425)	(2.454)	(2.194)
BTL	7.065***	8.270***	8.273***
	(2.487)	(2.411)	(2.313)
Constant	84.227	64.691	49.324
	(1.522)	(12.416)	(16.127)
Observations	999	999	999
R^2	0.046	0.068	0.178

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the Early Literacy Index TOPEL score regressed on the categorical treatment variable, without control variables, as in table 5.1. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.5 Linear regression results for the relationship between the treatment groups and the 1st grade reading score, including control variables

	Reading score 1st grade		
Variable	**(1)**	**(2)**	**(3)**
RSL	13.093	12.235	12.777
	(8.482)	(8.921)	(8.415)
BELL	1.849	3.875	4.008
	(8.707)	(7.485)	(7.098)
BTL	23.991**	21.194*	19.136*
	(10.370)	(11.350)	(11.481)
Constant	558.813	491.015	443.010
	(5.807)	(54.781)	(93.642)
Observations	259	259	259
R^2	0.038	0.068	0.130

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the 1st grade reading score regressed on the categorical treatment variable, without control variables, as in table 5.2. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.6 Linear regression results for the relationship between the treatment groups and the 1st grade math score, including control variables

	Math score 1st grade		
Variable	**(1)**	**(2)**	**(3)**
RSL	14.507	10.474	12.202
	(9.537)	(9.025)	(8.030)
BELL	-0.919	-2.058	-2.443
	(7.163)	(7.736)	(7.411)

BTL	25.763***	19.864**	19.804**
	(8.684)	(9.053)	(9.820)
Constant	545.021	536.394	509.959
	(5.239)	(40.873)	(86.258)
Observations	259	259	259
R^2	0.064	0.089	0.135

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the 1st grade math score regressed on the categorical treatment variable, without control variables, as in table 5.2. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.7 Linear regression results for the relationship between the treatment groups and the 2nd grade reading score, including control variables

	Reading score 2nd grade		
Variable	**(1)**	**(2)**	**(3)**
RSL	4.280	2.670	3.455
	(4.751)	(4.297)	(4.160)
BELL	4.272	2.487	2.521
	(4.311)	(4.045)	(3.846)
BTL	6.345	7.613	8.516*
	(4.486)	(4.609)	(4.570)
Constant	605.606	542.179	495.298
	(2.907)	(16.903)	(35.037)
Observations	958	958	958
R^2	0.004	0.031	0.057

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the 2nd grade reading score regressed on the categorical treatment variable, without control variables, as in table 5.2. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.8 Linear regression results for the relationship between the treatment groups and the 2nd grade math score, including control variables

	Math score 2nd grade		
Variable	**(1)**	**(2)**	**(3)**
RSL	6.051	5.113	5.861
	(5.379)	(5.105)	(4.719)
BELL	5.238	4.144	4.166
	(4.692)	(4.552)	(4.285)
BTL	5.553	6.184	5.861
	(4.986)	(4.925)	(4.719)
Constant	587.135	538.355	452.953
	(3.442)	(22.940)	(34.385)
Observations	958	958	958
R^2	0.004	0.016	0.063

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the 2nd grade math score regressed on the categorical treatment variable, without control variables, as in table 5.2. Column 2 includes control variables that

were different between groups at baseline and column 3 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.9 Linear regression results of heterogeneous treatment effects by gender and home language of the child on the Definitional Vocabulary TOPEL score, including control variables

	Definitional Vocabulary					
Variable	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Male	2.415	2.385	2.064			
	(1.759)	(1.790)	(1.598)			
RSL	4.967	4.638	5.345**	2.818	1.661	3.202
	(3.095)	(3.053)	(2.495)	(2.741)	(2.808)	(2.939)
BELL	0.145	-0.076	0.897	4.838	4.931	6.103
	(3.233)	(3.346)	(2.929)	(4.528)	(4.828)	(4.627)
BTL	3.571	4.404	4.296	6.072**	8.211***	8.263***
	(2.950)	(2.948)	(2.779)	(2.558)	(2.717)	(3.138)
Male*RSL	-3.581	-3.391	-3.975			
	(2.554)	(2.659)	(2.644)			
Male*BELL	-2.300	-1.964	-0.851			
	(3.002)	(2.974)	(2.568)			
Male*BTL	-2.444	-1.807	-0.768			
	(2.749)	(2.780)	(2.480)			
EngSpan/Span only				-11.625***	-11.786***	-8.133***
				(2.413)	(2.310)	(2.300)
EngOth/SpanOth/Oth				-6.042**	-5.353*	-4.327
				(2.932)	(2.968)	(3.032)
EngSpan/Span only *RSL				0.942	1.858	0.249
				(3.650)	(3.537)	(3.341)
EngSpan/Span only *BELL				-4.829	-5.101	-5.848
				(5.186)	(5.607)	(5.257)
EngSpan/Span only *BTL				-4.258	-5.657*	-5.371*
				(3.444)	(3.284)	(3.214)
EngOth/SpanOth/Oth *RSL				2.807	2.314	1.631
				(5.993)	(6.202)	(6.153)
EngOth/SpanOth/Oth *BELL				-5.052	-3.972	-8.247
				(7.751)	(8.829)	(7.545)
EngOth/SpanOth/Oth *BTL				2.999	1.436	0.934
				(4.745)	(4.959)	(5.169)
Constant	77.896	71.781	84.602	87.542	72.792	71.690
	(1.738)	(12.666)	(18.061)	(2.130)	(10.932)	(17.201)
Observations	999	999	999	999	999	999
R^2	0.011	0.022	0.156	0.106	0.120	0.180

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the Definitional Vocabulary TOPEL score regressed on the categorical treatment variable, the dummy variable gender and on the interaction terms

between gender and treatment dummies/categories, without control variables, as in table 5.3. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline. Column 4 shows results of the OLS regression of the Definitional Vocabulary TOPEL score regressed on the categorical treatment variable, the categorical home language variable and on the interaction terms between home language and treatment categories, without control variables, as in table 5.5. Column 5 includes again control variables that were different between groups at baseline and column 6 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.10 Linear regression results of heterogeneous treatment effects by gender and home language of the child on the Phonological Awareness TOPEL score, including control variables

	Phonological Awareness					
Variable	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Male	1.123	8.220	1.236			
	(2.177)	(2.806)	(2.107)			
RSL	8.919***	1.899***	8.929***	4.994**	3.728	5.068*
	(2.909)	(2.797)	(2.605)	(2.486)	(2.673)	(2.609)
BELL	2.263	9.324	2.648	7.988*	8.452*	8.606**
	(2.807)	(2.633)	(2.449)	(4.435)	(4.690)	(4.182)
BTL	8.321***	9.324***	8.926***	6.558*	8.810	6.726*
	(2.620)	(2.633)	(2.520)	(3.512)	(3.432)	(3.598)
Male*RSL	-7.148**	-6.821**	-7.764***			
	(2.922)	(2.921)	(2.934)			
Male*BELL	-1.973	-1.674	-1.380			
	(2.843)	(2.921)	(2.659)			
Male*BTL	-4.730	-1.674	-3.508			
	(3.052)	(2.854)	(2.930)			
EngSpan/Span only				-2.209	-2.521	-0.339
				(2.009)	(2.061)	(2.035)
EngOth/SpanOth/Oth				0.994	1.506	1.545
				(3.060)	(3.180)	(3.501)
EngSpan/Span only *RSL				0.265	1.151	-0.410
				(3.169)	(3.454)	(3.215)
EngSpan/Span only *BELL				-6.962	-7.897	-7.527*
				(4.802)	(5.123)	(4.469)
EngSpan/Span only *BTL				-0.660	-1.953	0.379
				(4.036)	(3.817)	(3.593)
EngOth/SpanOth/Oth *RSL				3.956	3.615	2.884
				(5.317)	(5.371)	(5.790)
EngOth/SpanOth/Oth *BELL				-14.073**	-12.816*	-15.451**
				(5.916)	(6.673)	(6.606)
EngOth/SpanOth/Oth *BTL				1.642	-0.153	1.051
				(4.991)	(4.932)	(5.216)
Constant	87.778	-4.341	44.891	89.806	69.597	43.506
	(1.954)	(3.035)	(14.492)	(1.713)	(10.848)	(14.453)

Observations	999	999	999	999	999	999
R^2	0.038	0.058	0.130	0.042	0.066	0.130

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the Phonological Awareness TOPEL score regressed on the categorical treatment variable, the dummy variable gender and on the interaction terms between gender and treatment dummies/categories, without control variables, as in table 5.3. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline. Column 4 shows results of the OLS regression of the Phonological Awareness TOPEL score regressed on the categorical treatment variable, the categorical home language variable and on the interaction terms between home language and treatment categories, without control variables, as in table 5.5. Column 5 includes again control variables that were different between groups at baseline and column 6 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.11 Linear regression results of heterogeneous treatment effects by gender and home language of the child on the Print Knowledge TOPEL score, including control variables

	Print Knowledge					
Variable	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Male	-0.299	-0.484	-0.339			
	(1.923)	(1.862)	(1.755)			
RSL	11.564***	10.585***	10.856***	4.941	2.877	3.619
	(2.395)	(2.429)	(2.349)	(3.117)	(2.804)	(3.021)
BELL	3.824	3.049	3.401	1.465	0.920	0.616
	(2.657)	(2.594)	(2.455)	(3.665)	(3.743)	(3.504)
BTL	9.636***	9.952***	9.439***	10.384***	11.274***	9.527***
	(2.481)	(2.527)	(2.434)	(3.597)	(3.152)	(3.246)
Male*RSL	-3.301	-2.793	-3.500			
	(2.288)	(2.257)	(2.185)			
Male*BELL	-5.059*	-4.665*	-4.434*			
	(2.288)	(2.627)	(2.469)			
Male*BTL	-2.320	-1.989	-1.435			
	(2.689)	(2.708)	(2.625)			
EngSpan/Span only				-3.703	-4.565*	-3.719
				(2.677)	(2.462)	(2.266)
EngOth/SpanOth/Oth				0.811	1.128	0.787
				(2.389)	(2.342)	(2.412)
EngSpan/Span only				6.620*	8.246***	7.149**
*RSL				(3.382)	(3.127)	(3.101)
EngSpan/Span only				0.709	0.685	1.266
*BELL				(4.336)	(4.332)	(4.058)
EngSpan/Span only				-1.780	-2.236	-0.534
*BTL				(4.075)	(3.649)	(3.454)
EngOth/SpanOth/Oth				4.859	4.950	4.763
*RSL				(4.581)	(4.546)	(4.616)
EngOth/SpanOth/Oth				-2.200	-1.032	-2.655
*BELL				(4.610)	(5.175)	(4.829)

EngOth/SpanOth/Oth *BTL				-4.477	-5.240	-4.248
				(4.709)	(4.845)	(4.713)
Constant	96.299	75.268	47.574	98.639	73.327	46.878
	(1.851)	(11.293)	(15.191)	(2.422)	(11.864)	(15.666)
Observations	999	999	999	999	999	999
R^2	0.099	0.118	0.163	0.099	0.125	0.168

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the Print Knowledge TOPEL score regressed on the categorical treatment variable, the dummy variable gender and on the interaction terms between gender and treatment dummies/categories, without control variables, as in table 5.3. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline. Column 4 shows results of the OLS regression of the Print Knowledge TOPEL score regressed on the categorical treatment variable, the categorical home language variable and on the interaction terms between home language and treatment categories, without control variables, as in table 5.5. Column 5 includes again control variables that were different between groups at baseline and column 6 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.12 Linear regression results of heterogeneous treatment effects by gender and home language of the child on the Early Literacy Index TOPEL score, including control variables

	Early Literacy Index					
Variable	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Male	1.380	1.246	1.263			
	(2.017)	(1.990)	(1.812)			
RSL	10.690***	9.856***	10.568***	5.294*	3.408	4.933
	(2.881)	(2.797)	(2.447)	(3.035)	(2.944)	(3.092)
BELL	2.561	2.004	2.877	5.859	5.872	6.296
	(3.177)	(3.183)	(2.802)	(4.777)	(5.091)	(4.694)
BTL	9.045***	9.960***	9.540***	9.605***	11.861***	10.299***
	(2.935)	(2.912)	(2.755)	(3.563)	(3.250)	(3.589)
Male*RSL	-5.935**	-5.512**	-6.451**			
	(2.591)	(2.607)	(2.573)			
Male*BELL	-3.888	-3.459	-2.766			
	(3.067)	(3.011)	(2.677)			
Male*BTL	-4.016	-3.447	-2.428			
	(2.941)	(2.941)	(2.750)			
EngSpan/Span only				-7.461***	-8.019***	-5.211**
				(2.520)	(2.378)	(2.272)
EngOth/SpanOth/Oth				-1.836	-1.198	-0.889
				(2.675)	(2.754)	(3.022)
EngSpan/Span only *RSL				3.340	4.786	2.994
				(3.529)	(3.524)	(3.321)
EngSpan/Span only *BELL				-4.520	-5.032	-4.938
				(5.346)	(5.707)	(5.181)
EngSpan/Span only *BTL				-2.742	-4.100	-2.305
				(4.208)	(3.750)	(3.578)
EngOth/SpanOth/Oth				5.056	4.736	4.079

*RSL				(4.749)	(4.804)	(4.879)
EngOth/SpanOth/Oth *BELL				-8.937 (5.992)	-7.461 (7.326)	-11.034* (6.574)
EngOth/SpanOth/Oth *BTL				0.102 (5.207)	-1.655 (5.355)	-0.957 (5.546)
Constant	83.521 (1.992)	-3.447 (2.941)	48.152 (16.125)	89.486 (2.240)	64.148 (12.300)	41.837 (16.129)
Observations	999	999	999	999	999	999
R^2	0.054	0.074	0.183	0.090	0.119	0.192

Notes: Standard errors between brackets, adjusted for clustering at class level;this table shows results of a robustness check; column 1 shows the results of the OLS regression of the Early Literacy Index TOPEL score regressed on the categorical treatment variable, the dummy variable gender and on the interaction terms between gender and treatment dummies/categories, without control variables, as in table 5.3. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline. Column 4 shows results of the OLS regression of the Early Literacy Index TOPEL score regressed on the categorical treatment variable, the categorical home language variable and on the interaction terms between home language and treatment categories, without control variables, as in table 5.5. Column 5 includes again control variables that were different between groups at baseline and column 6 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.13 Linear regression results of heterogeneous treatment effects by gender and home language of the child on the 1st grade reading score, including control variables

	Reading score 1st grade					
Variable	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Male	3.241	1.716	0.947			
	(11.632)	(11.033)	(10.171)			
RSL	25.847*	26.100*	28.175*	12.203	12.202	10.740
	(14.920)	(14.886)	(14.198)	(14.719)	(10.667)	(15.139)
BELL	17.240	19.623	18.842*	20.920	18.084	18.858
	(13.697)	(10.544)	(10.233)	(11.632)	(11.039)	(13.535)
BTL	31.456*	27.691	23.580	39.920**	33.421*	31.908
	(16.427)	(17.541)	(16.476)	(18.087)	(18.613)	(22.957)
Male*RSL	-23.149	-24.404	-29.385			
	(19.379)	(19.411)	(18.694)			
Male*BELL	-31.532*	-30.925*	-30.577*			
	(16.895)	(15.688)	(15.000)			
Male*BTL	-14.131	-10.742	-10.859			
	(18.645)	(19.231)	(17.812)			
EngSpan/Span only				24.892**	24.855**	32.596**
				(10.950)	(10.667)	(13.243)
EngOth/SpanOth/Oth				18.314	14.000	12.853
				(17.440)	(15.834)	(18.609)
EngSpan/Span only				-0.134	1.262	-1.113
*RSL				(16.884)	(17.765)	(18.489)
EngSpan/Span only				-29.454*	-23.347	-23.718
*BELL				(16.701)	(16.248)	(23.701)

EngSpan/Span only *BTL				-23.691 (21.206)	-17.234 (20.524)	-19.284 (23.701)
EngOth/SpanOth/Oth *RSL				- -	- -	- -
EngOth/SpanOth/Oth *BELL				-20.920 (26.567)	-20.662 (22.422)	-26.874 (23.085)
EngOth/SpanOth/Oth *BTL				-53.086 (44.729)	-34.282 (45.617)	-40.101 (44.785)
Constant	-14.131 (18.645)	507.725 (55.246)	422.078 (96.248)	543.353 (8.823)	482.008 (56.301)	483.988 (91.325)
Observations	259	259	259	259	259	259
R^2	0.069	0.100	0.149	0.076	0.103	0.160

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the 1st grade reading score regressed on the categorical treatment variable, the dummy variable gender and on the interaction terms between gender and treatment dummies/categories, without control variables, as in table 5.4. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline. Column 4 shows results of the OLS regression of the 1st grade reading score regressed on the categorical treatment variable, the categorical home language variable and on the interaction terms between home language and treatment categories, without control variables, as in table 5.6. Column 5 includes again control variables that were different between groups at baseline and column 6 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.14 Linear regression results of heterogeneous treatment effects by gender and home language of the child on the 1st grade math score, including control variables

	Math score 1st grade					
Variable	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Male	11.248 (8.910)	9.640 (8.492)	9.837 (7.962)			
RSL	29.309** (14.521)	24.717* (13.638)	27.155* (12.995)	10.882 (16.670)	9.368 (16.886)	9.608 (16.944)
BELL	5.435 (8.425)	4.744 (8.214)	4.600 (8.874)	2.746 (8.437)	-0.879 (8.900)	2.340 (10.450)
BTL	23.917** (9.607)	17.532* (9.400)	17.044 (10.527)	44.019*** (14.463)	33.609* (17.011)	29.972 (19.325)
Male*RSL	-28.289* (16.633)	-26.486 (16.190)	-27.882* (14.682)			
Male*BELL	-13.686 (12.509)	-13.439 (12.943)	-13.964 (12.912)			
Male*BTL	1.505 (12.997)	3.907 (13.606)	2.881 (13.646)			
EngSpan/Span only				26.401*** (9.706)	26.044** (9.964)	28.518** (11.811)
EngOth/SpanOth/Oth				-1.840 (9.156)	-4.118 (9.074)	-1.070 (10.934)

EngSpan/Span only *RSL				0.899 (17.940)	1.213 (18.902)	-1.565 (19.227)
EngSpan/Span only *BELL				-13.609 (12.681)	-10.008 (14,317)	-13.339 (15.228)
EngSpan/Span only *BTL				-29.143 (17.895)	-20.751 (18.992)	-17.154 (20.217)
EngOth/SpanOth/Oth *RSL				- -	- -	- -
EngOth/SpanOth/Oth *BELL				21.143** (10.349)	23.492** (9.535)	9.522 (16.826)
EngOth/SpanOth/Oth *BTL				-44.296* (24.156)	-28.087 (29.673)	-27.302 (30.349)
Constant	539.865 (6.291)	530.483 (42.685)	486.588 (83.332)	530.618 (7.136)	532.170 (41.957)	548.099 (87.108)
Observations	259	259	259	259	259	259
R^2	0.083	0.107	0.153	0.133	0.151	0.170

Notes: Standard errors between brackets, adjusted for clustering at class level;this table shows results of a robustness check; column 1 shows the results of the OLS regression of the 1st grade math score regressed on the categorical treatment variable, the dummy variable gender and on the interaction terms between gender and treatment dummies/categories, without control variables, as in table 5.4. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline. Column 4 shows results of the OLS regression of the 1st grade math score regressed on the categorical treatment variable, the categorical home language variable and on the interaction terms between home language and treatment categories, without control variables, as in table 5.6. Column 5 includes again control variables that were different between groups at baseline and column 6 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.15 Linear regression results of heterogeneous treatment effects by gender and home language of the child on the 2nd grade reading score, including control variables

	Reading score 2nd grade					
Variable	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Male	5.715	4.441	4.291			
	(4.919)	(4.866)	(4.823)			
RSL	18.718***	15.640***	16.457***	-6.817	-10.717	-8.267
	(6.006)	(5.666)	(5.667)	(9.388)	(8.622)	(9.095)
BELL	11.313**	8.824*	8.950*	5.949	6.539	6.939
	(5.329)	(4.859)	(4.596)	(8.987)	(8.962)	(9.352)
BTL	7.740	8.293	8.864	-1.297	0.239	-0.475
	(5.850)	(5.943)	(6.090)	(7.790)	(7.815)	(8.168)
Male*RSL	-27.365***	-24.395***	-24.644***			
	(7.043)	(6.950)	(6.718)			
Male*BELL	-14.456**	-12.541*	-0.186*			
	(6.791)	(6.952)	(6.835)			
Male*BTL	-2.725	-1.089	-0.186			

	(7.038)	(6.973)	(6.835)			
EngSpan/Span only				7.591	5.937	6.948
				(6.533)	(6.018)	(6.415)
EngOth/SpanOth/Oth				-8.589	-6.558	-5.668
				(7.436)	(6.794)	(6.901)
EngSpan/Span only *RSL				12.933	16.053*	13.565
				(10.139)	(9.563)	(9.859)
EngSpan/Span only				-4.059	-7.182	-6.843
*BELL				(10.254)	(10.335)	(10.609)
EngSpan/Span only *BTL				7.689	7.390	8.986
				(8.918)	(8.560)	(8.648)
EngOth/SpanOth/Oth				23.342	21.700	19.158
*RSL				(14.921)	(14.107)	(13.512)
EngOth/SpanOth/Oth				7.843	11.425	6.079
*BELL				(15.430)	(12.812)	(14.005)
EngOth/SpanOth/Oth				22.023	19.014	19.280
*BTL				(13.695)	(13.414)	(13.769)
Constant	602.729	543.759	491.602	600.939	544.94	514.057
	(4.131)	(16.776)	(34.343)	(5.830)	(17.787)	(32.734)
Observations	958	958	958	958	958	958
R^2	0.026	0.049	0.073	0.025	0.050	0.068

Notes: Standard errors between brackets, adjusted for clustering at class level; this table shows results of a robustness check; column 1 shows the results of the OLS regression of the 2nd grade reading score regressed on the categorical treatment variable, the dummy variable gender and on the interaction terms between gender and treatment dummies/categories, without control variables, as in table 5.4. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline. Column 4 shows results of the OLS regression of the 2nd grade reading score regressed on the categorical treatment variable, the categorical home language variable and on the interaction terms between home language and treatment categories, without control variables, as in table 5.6. Column 5 includes again control variables that were different between groups at baseline and column 6 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.

Table A5.16 Linear regression results of heterogeneous treatment effects by gender and home language of the child on the 2nd grade math score, including control variables

	Math score 2nd grade					
Variable	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Male	16.609***	15.610***	15.824***			
	(5.257)	(5.277)	(5.081)			
RSL	16.918***	14.861***	16.209***	1.489	0.545	2.506
	(5.777)	(5.601)	(5.550)	(11.780)	(11.170)	(11.143)
BELL	12.501**	10.901**	10.871**	3.887	5.157	6.189
	(5.436)	(5.330)	(4.872)	(9.094)	(8.818)	(8.181)
BTL	9.202	9.488	9.902	7.829	9.798	7.758
	(5.973)	(5.954)	(6.186)	(8.210)	(8.774)	(9.442)
Male*RSL	-21.298***	-19.028**	-19.764***			
	(7.640)	(7.622)	(7.352)			

Male*BELL	-14.300*	-12.923*	-13.127*			
	(7.388)	(7.490)	(7.350)			
Male*BTL	-7.096	-5.933	-4.701			
	(7.255)	(7.217)	(7.052)			
EngSpan/Span only				22.499***	22.375***	23.343***
				(7.142)	(11.170)	(7.397)
EngOth/SpanOth/Oth				1.241	3.768	6.461
				(9.045)	(8.697)	(8.845)
EngSpan/Span only *RSL				3.841	4.387	2.540
				(12.874)	(12.521)	(12.521)
EngSpan/Span only *BELL				-3.750	-6.472	-6.016
				(10.005)	(9.949)	(9.617)
EngSpan/Span only *BTL				-4.613	-5.588	-1.546
				(9.717)	(10.102)	(10.481)
EngOth/SpanOth/Oth *RSL				19.236	15.950	13.025
				(17.792)	(17.576)	(18.075)
EngOth/SpanOth/Oth *BELL				34.606*	35.121**	30.169*
				(18.430)	(17.304)	(18.995)
EngOth/SpanOth/Oth *BTL				-2.017	-5.745	-5.981
				(16.205)	(16.343)	(16.686)
Constant	578.771	534.035	449.651	571.409	532.010	483.321
	(4.387)	(23.242)	(34.169)	(6.560)	(22.743)	(32.005)
Observations	958	958	958	958	958	958
R^2	0.019	0.068	0.071	0.051	0.061	0.084

Notes: Standard errors between brackets, adjusted for clustering at class level;this table shows results of a robustness check; column 1 shows the results of the OLS regression of the 2nd grade math score regressed on the categorical treatment variable, the dummy variable gender and on the interaction terms between gender and treatment dummies/categories, without control variables, as in table 5.4. Column 2 includes control variables that were different between groups at baseline and column 3 includes all the variables at baseline. Column 4 shows results of the OLS regression of the 2nd grade math score regressed on the categorical treatment variable, the categorical home language variable and on the interaction terms between home language and treatment categories, without control variables, as in table 5.6. Column 5 includes again control variables that were different between groups at baseline and column 6 includes all the variables at baseline; results are not presented in effect sizes; *p < 0.1, **p < 0.05, ***p < 0.01.