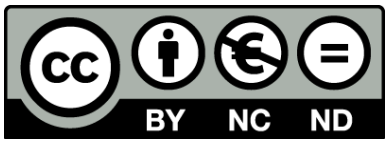
UNIVERSITAT DE BARCELONA

Essays on Empirical Evaluation of Development Programs

Anastasiya Yarygina Udovenko



2022

PhD in Economics | Anastasiya Yarygina Udovenko

ERSITAT DE
CELONA

PhD in Economics

Essays on Empirical Evaluation of Development Programs

Anastasiya Yarygina Udovenko

UNIVERSITAT DE BARCELONA

LIBERTAS PERFVNDET
OMNIA LVCE

UNIVE
BARC

PhD in Economics

Thesis title:

Essays on Empirical Evaluation of Development Programs

PhD candidate:
Anastasiya Yarygina Udovenko

Advisors:
Daniel Albalate del Sol
Julian Cristia

Date:
March 2022

UNIVERSITAT DE BARCELONA

To my mother, Laryssa, and my husband, Jaume.

Acknowledgements:

I am deeply grateful to my advisers, Daniel Albalate and Julian Cristia, for their infinite patience and unconditional support during these years. Their professional and personal advice was essential for moving forward and improving this dissertation. I consider them both outstanding professionals, excellent role models, and feel fortunate for having a chance of working together.

I am grateful to Julia Johannsen, Sebastian Martinez and Cecilia Vidal, for working together on different projects at the Inter-American Development Bank and for their valuable inputs to the first two studies included in this dissertation, particularly, at the intervention design, implementation, and data collection. I am thankful to Cecilia Vidal and Sebastian Martinez who gave me valuable advice throughout the process of data acquisition, data analysis and interpretation, producing of the results, and drafting of the manuscript. I am particularly thankful to Sebastian Martinez, who trusted in me and gave me the opportunity to grow as an impact evaluation professional.

I also want to thank my friends and colleagues, particularly Cecilia Vidal, who gave me her full support during challenging times. I am also thankful to my colleagues at the Fiscal Management Division of the Inter-American Development Bank for encouraging and motivating me during the final sprint of writing this dissertation.

Finally, I want to thank my husband and my mother for their unconditional support, patience, and love.

The studies presented in Chapter 2 and Chapter 3 of this dissertation are a result of collaboration between the Inter-American Development Bank and the Ministry of Health of the Plurinational State of Bolivia. These studies were funded by the Inter-American Development Bank Investment Loan BO-L1064 and Technical Cooperations BO-T1114, BO-T1159, BO-T1214. I thank the Early Childhood Development Unit and the Early Childhood Development Program at the MH, in particular, Mariana Ramirez and Marlene Calle, the Private Bolivian University survey team, in particular,

Carlos Foronda and Mauricio Chumacero, and Teresa Reinaga, for her support in the quality control of the survey; early childhood development professionals and families who contributed their time to the implementation of the project and its evaluation. The identification number of the randomized controlled trial is AEARCTR-0002261. The analysis, results and interpretations presented in this dissertation should not be attributed to the Ministry of Health of the Plurinational State of Bolivia, or to the Inter-American Development Bank, its Executive Directors, or countries it represents. The first two studies included in this dissertation received helpful comments from Marta Rubio, Samuel Berlinski, and other seminar participants at the Inter-American Development Bank in Washington DC and La Paz, Bolivia.

Table of Contents

Chapter 1: Introduction

1.1 Motivation and contribution: why evaluate?

Enormous sums of money are spent every year to try to improve outcomes. According to the Organisation for Economic Co-operation and Development (OECD), on average, OECD countries spend around 20% of Gross Domestic Product (GDP) on social policies and programs (OECD, 2021a). The expenditures on development aid in 2020 reached US$175 billion (OECD, 2021b), financing numerous development interventions across the globe. Through commitments to the Millennium Development Goals and subscriptions of global initiatives, donors have repeatedly committed to increasing their financial assistance to the developing countries. But do we know whether these expenditures improve well-being? For example, did a water sanitation project in rural indigenous communities increase the use of clean water and better hygiene practices? Did a new computer-based instruction of mathematics in primary schools of a large city increase learning? Was the training program for disadvantaged youth in shantytowns effective in fostering labor market outcomes? Can we say how the development assistance funds should be allocated across different areas of social spending? Program evaluation is the tool to understand whether programs and policies work and whether the resources spent on designing and implementing the interventions make a difference.

Generally, the objectives of program evaluations are to determine whether programs are effective, provide accountability for organizations running programs, and improve future programs based on the lessons learned from past experiences. The ultimate purpose of program evaluation is to identify what works so that we can improve people's outcomes in a more effective and efficient way.

Program evaluation is a very broad field, encompassing different types of evaluations, such as the assessment of whether the program is relevant (i.e., whether it is needed, and whether the program design has a logic and theory of change); assessment of program implementation, which focuses on the analysis of processes; assessment of program effectiveness (i.e., whether the

program objectives are achieved), the assessment of program impact (impact evaluation); assessment of program costs (i.e., efficiency).[1]

Impact evaluation is one of many different approaches to evaluating policies and programs. Being part of a large program evaluation toolkit, impact evaluation is the only one that provides rigorous evidence that the changes in the well-being of individuals can be attributed to a program or policy. In a nutshell, impact evaluation is a set of methods that can be used to identify a causal relationship between the intervention (program o policy) and the outcome (e.g., learning, employment, income) by comparing the results in the population with and without the intervention. If the evaluation is well-designed and well-implemented, it can provide convincing evidence on program effectiveness, the evidence that can be used to inform policy decisions, shape public opinion, and improve program implementation.[2]

In this dissertation, I present the results of impact evaluations of two development programs implemented in Bolivia and one social policy implemented in Brazil. The availability of rigorous evidence to inform policy decisions in Latin America and the Caribbean (LAC) is as relevant as elsewhere in the developing world. Nonetheless, the need for evidence-based decision making in LAC acquired particular importance in the past two years because of the urgency to develop a road map guiding an intelligent recovery from the economic downturn caused by the COVID-19 pandemic, which hit particularly hard LAC countries.[3] In addition, the region strives to solve the structural problem of ineffective and inefficient public spending, which requires producing rigorous evidence for better allocation of public resources (Izquierdo et al., 2018). The first two impact evaluations contribute to the literature by providing causal estimates of the effect of large randomized controlled trials which sought to improve children's development through better early childhood experiences and environment. The third impact evaluation provides quasi-experimental causal evidence answering an

[1] See, for example, Rossi et al. (2018) for a broad systematic approach to program evaluation.

[2] The extent to which impact evaluation can influence stakeholders and a broader policy agenda is discussed elsewhere. See, for example, Achie (2019).

[3] Latin America and the Caribbean has become the region hardest hit by the COVID-19 pandemic. On May 22 of 2020, the region was declared the "new epicenter" of the pandemic by the World Health Organization and remained in that condition until September 2020.

important question of whether an increase in spending on education makes a difference for student achievement. All three studies evaluate programs focused on fostering human capital, which has proved to be important in determining adult outcomes and intergenerational mobility at the individual level.[4] At the macro level, human capital is considered to be a key factor driving productivity, economic growth and development.[5]

Given that the empirical studies presented in this dissertation implement impact evaluation methods, in the following sections of this introductory chapter I lay out the methodological framework of impact evaluation. The purpose of these sections is to contextualize the methodologies that are used in the impact evaluations presented in chapters two, three and four. Because two impact evaluations presented in this dissertation use the experimental evaluation method and one uses the instrumental variables (IV) approach, in addition to presenting the potential outcome framework for causal inference, I elaborate on the experimental evaluation and the IV method in the context of imperfect compliance.[6] The last section of this introductory chapter presents the summaries of the chapters two through five.

1.2 Causal inference and potential outcome[7]

The focus on causality and attribution is the cornerstone and the main challenge of impact evaluation. Specifically, the challenge is that the estimation of the program impact with attribution requires estimating outcomes with and without the program. Formally, for an individual i, let D_i be a treatment indicator. When the individual is treated, $D_i = 1$, and when not treated, $D_i = 0$. Let Y_{1i} be the outcome if treated and Y_{0i} be the outcome if not treated. Then, the impact of treatment (program or policy) for the ith individual is:

[4] Attanasio et al. (2021) and studies cited there.
[5] See, e.g., Lucas (2015), Attanasio et al. (2020) and the studies cited there.
[6] The details on other impact evaluation methods can be found in dedicated textbooks and toolkits, including Duflo, et al. (2007), Banerjee and Duflo (2011), Angrist and Pischke (2008), Imbens and Rubin (2015), Gertler et al. (2016).
[7] Sections 1.2 -1.5 draw on the lecture notes from the Program Evaluation course by Ofer Malamud, Applied Econometrics course by Koichiro Ito (both of the University of Chicago Harris School of Public Policy), and the textbook Duflo et al. (2007).

$$\Delta_i = Y_{1i} - Y_{0i}$$

This equation illustrates the fundamental problem of evaluation: it is impossible to observe one individual in two states. That is, we only get to observe one of the two potential outcomes for the same individual. As a result, it is impossible to determine actual program impact for a specific individual. Instead, what impact evaluation methods do, they focus on estimating the average effect of the treatment for all individuals in the relevant population.

Formally, the Average Treatment Effect (ATE), or the impact of the treatment on the entire population, is an expectation of the ith individual ATE, or an expectation of the difference between the outcome with and without program for the ith individual:

$$\Delta^{ATE} = E(\Delta_i) = E(Y_{1i} - Y_{0i})$$

Another commonly estimated parameter is the Average Treatment Effect on the Treated (ATT), which shows the program effect in the treated population:

$$\Delta^{ATT} = E(\Delta_i | D = 1) = E(Y_{1i} - Y_{0i} | D = 1)$$

If treatment effects are constant or homogeneous, then $\Delta^{ATT} = \Delta^{ATE}$. A closer look at the ATT formula shows that it consists of difference between two means: $E(Y_{1i}|D = 1)$ and $E(Y_{0i}|D = 1)$. The first mean can be estimated in a straightforward way. Because of the law of large numbers, $E(Y_{1i}|D = 1) \cong \bar{Y}_{1,D=1}$. The estimation of the second mean, $E(Y_{0i}|D = 1)$, is the estimation of the "missing counterfactual."
By subtracting and adding $E(Y_{0i}|D = 0)$ to Δ^{ATT}, we obtain the following expression:

$$\begin{aligned}
\Delta^{ATT} + E(Y_{0i}|D = 1) - E(Y_{0i}|D = 1) \\
&= E(Y_{1i}|D = 1) - E(Y_{0i}|D = 1) + E(Y_{0i}|D = 0) \\
&- E(Y_{0i}|D = 0) \\
&= \bar{Y}_{1,D=1} - \bar{Y}_{0,D=0} + [E(Y_{0i}|D = 1) - E(Y_{0i}|D = 0)] \\
&= \Delta^{ATT} + [E(Y_{0i}|D = 1) - E(Y_{0i}|D = 0)]
\end{aligned}$$

Where $E(Y_{0i}|D=1) - E(Y_{0i}|D=0)$ is the selection term. The selection term captures differences in the outcome of the treatment group (the group that received the treatment), had it not been treated, and the control group (the group that does not receive the treatment).

Because $Y_i = Y_{1i}$ if $D_i = 1$ and $Y_i = Y_{0i}$ if $D_i = 0$, we can write the observed outcome Y_i in terms of the potential outcomes Y_{1i} and Y_{0i} as follows:

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$$

Under constant treatment effects, $\Delta_i = Y_{1i} - Y_{0i} = \delta$. Then:

$$\begin{aligned} Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i} = D_i Y_{1i} + Y_{0i} - D_i Y_{0i} = Y_{0i} + (Y_{1i} - Y_{0i}) D_i \\ = Y_{0i} + (Y_{1i} - Y_{0i}) D_i + [E(Y_{0i}) - E(Y_{0i})] = E(Y_{0i}) \\ + (Y_{1i} - Y_{0i}) D_i + Y_{0i} - E(Y_{0i}) = \beta + \delta D_i + \varepsilon_i \end{aligned}$$

Where β is the mean of expectation of Y_{0i}, δ is a constant treatment effect, $\Delta_i = Y_{1i} - Y_{0i} = \delta$, and ε_i is the random part of Y_{0i}. Now, if we take the expectation of Y_i conditional on $D_i = 1$ and $D_i = 0$ and substruct both expressions, we obtain:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \delta + E[\varepsilon_i|D_i = 1] - E[\varepsilon_i|D_i = 0]$$

or

$$\bar{Y}_{1,D=1} - \bar{Y}_{0,D=0} = \delta + E[\varepsilon_i|D_i = 1] - E[\varepsilon_i|D_i = 0]$$

Where $E[\varepsilon_i|D_i = 1] - E[\varepsilon_i|D_i = 0] = E(Y_{0i}|D=1) - E(Y_{0i}|D=0)$ is the selection term expressed in terms of the error of a linear regression. So, for estimation of the treatment effect without selection, we need to assume that the expectation of the error term does not depend on the treatment status or that the error term is uncorrelated with treatment.

Assuming linearity of potential outcomes as a function of observed characteristics X, assuming linearity in parameters and separability of unobserved characteristics captured by ε_i we get:

$$Y_{1i} = X_i\beta_1 + \varepsilon_{1i}$$
$$Y_{0i} = X_i\beta_0 + \varepsilon_{0i}$$

Assuming that these regression functions have identical parameters and only differ by a constant δ which vaires by treatment D, we get a standard Ordinary Least Squares (OLS) regression:

$$Y_i = X_i\,\beta + \delta D_i + \varepsilon_i$$

In which δ is an unbiased estimate of causal effect of treatment D on outcome Y if the following assumptions on the error term hold:

$$E(\varepsilon_i) = 0$$
$$E(\varepsilon_i|X_i) = E(\varepsilon_i)$$
$$E(\varepsilon_i|D_i) = E(\varepsilon_i) \Leftrightarrow E(\varepsilon_i|D_i = 1) = E(\varepsilon_i|D_i = 0)$$

Combining these assumptions, we can write:

$$E(\varepsilon_i|X_i, D_i) = 0$$

This expression indicates that under the assumption that the conditional mean of the error is zero, there may be selection on observable characteristics X, but no selection on any unobservable characteristics. Under this assumption, the OLS estimates of δ are unbiased such that $E(\hat{\delta}) = \delta$.

1.3 Experimental method as a "gold standard"

Experimental evaluation, in which program participants are assigned to treatment and control groups randomly, is considered to be a "gold standard"[8] of impact evaluation methods because it allows to obtain a rigorous and unbiased estimate of the causal impact of an intervention. This happens because, if the treatment status is randomly determined, the distribution of observable and unobservable characteristics for the treated and untreated populations is the same as the distribution for the whole population.

[8] Abdelghafour (2017), Donovan (2018).

Formally:

$$F(X, \varepsilon, |D = 1) = F(X, \varepsilon, |D = 0) = F(X, \varepsilon)$$

This implies that there is no systematic difference between treated and not treated, meaning that there is *no selection problem by design*. This allows to estimate the ATE by a simple difference between treatment and control group means:

$$\hat{\Delta}^{ATE} = E(Y_{1i}) - E(Y_{0i}) = E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) \cong \bar{Y}_1 - \bar{Y}_0$$

In the OLS regression framework:

$$Y_i = \beta + \delta D_i + \varepsilon_i$$

Where $\delta = \hat{\Delta}^{ATE} = \bar{Y}_1 - \bar{Y}_0$ and $\hat{\Delta}^{ATE}$ is unbiased under random assignment. If we include observed characteristics X, the regression to estimates is the following:

$$Y_i = X_i \beta + \delta D_i + \varepsilon_i$$

If the treatment is randomly assigned, we add X to improve precision of the estimate $\hat{\Delta}^{ATE}$. Adding X should not change the estimate itself, so long as D and X are statistically independent.

In summary, an experimental evaluation, through randomized assignment in treatment and control groups, achieves the best possible "missing counterfactual". The result is that the estimation of $\hat{\Delta}^{ATE}$ is unbiased because the conditional independence assumption is satisfied by construction.

1.4 Imperfect compliance and instrumental variables

Despite being a "gold standard" of impact evaluation methods, experimental evaluations present challenges. One of them is imperfect compliance with the assignment to the treatment or control group. Without complete compliance, the possibility of self-selection arises because, on the one hand, treatment group members who chose not to get treated may not be a random subset of

the experimental sample. On the other hand, the control group members who receive treatment (or a substitute of treatment) may not be a random subset of the experimental sample. In this situation, the difference in the means of the treated group and the control group is no longer an unbiased estimate of the ATE. In this case, the experimental estimate is referred to as the intent-to-treatment (ITT) parameter and is interpreted as the mean effect of offering the treatment. While ITT is identified by the random assignment, the identification of ATE requires more assumptions.

For a situation with some treatment units not complying with the treatment (dropout or incomplete take-up) and all control units complying (no control substitution or control contamination), let R be an indicator for randomization status with $R = 1$ for those who were randomly assigned to get treated and $R = 0$ otherwise. Let $D = 1$ be an indicator for actual receipt of the treatment for those assigned to treatment and $D = 0$ otherwise. Let D^* be a latent variable for those in the control group with $D^* = 1$ indicating that the individual would have received the treatment if they were in the treatment group, and $D^* = 0$ otherwise. An unbiased estimate of the impact of the actual receipt of the treatment requires an additional assumption that the mean outcome of those who did not take up treatment in the treated group is the same as that of the control group analog:

$$E(Y_i|R = 1, D = 0) = E(Y_i|R = 0, D^* = 0)$$

Under this assumption we can estimate the Bloom estimator. It scales up the experimental difference estimate by the fraction of the treatment group which receives treatment Pr $(D = 1|R = 1)$:

$$\Delta^{Bloom} = \frac{E(Y_i|R = 1) - E(Y_i|R = 0)}{\Pr(D = 1|R = 1)}$$

When everyone in the treatment group receives the treatment, the Bloom estimator equals the experimental estimate.

In the context of control substitution, we need an additional assumption. Let S be an indicator for control group members receiving treatment, with $S = 1$ for those who received the substitute treatment and $S = 0$ for those who did

not receive any treatment. To get an unbiased estimate, we need to assume that, for those who are getting treated, the impact for the program being evaluated is the same as the impact for the substitute programs. Formally:

$$E(Y_{1i} - Y_{0i}|D = 1, R = 1) = E(Y_{1i} - Y_{0i}|S = 1, R = 0)$$

In the case of cross-over, or control contamination, where some control group gets treated through the experimental program being evaluated and $S = D$, this assumption is easier to hold. Under this assumption, we can estimate a Wald estimator, which identifies the mean impact on those receiving the treatment in either treatment of control groups:

$$\Delta^{Wald} = \frac{E(Y_i|R = 1) - E(Y_i|R = 0)}{\Pr(D = 1|R = 1) - \Pr(D = 1|R = 0)}$$

With heterogeneous treatment effects, the Wald estimator gives the impact of receiving treatment for those who comply with treatment (compliers), also known as the Local Average Treatment Effect, or LATE.

The Wald estimate is the Instrumental Variable (IV) estimate which can be obtained in a two-stage estimation procedure, where, in the first stage, the actual treatment indicator D is regressed on the assignment to treatment R, and in the second stage, the outcome indicator is regressed on the prediction of D from the first stage:

$$D_i = \alpha R_i + X_i\gamma + v_i$$
$$Y_i = \delta \widehat{D}_i + X_i\beta + \varepsilon_i$$

The reduced form estimation of the outcome on the assignment to treatment is the ITT estimate:

$$Y_i = \lambda R_i + X_i\theta + \mu_i$$

Considering that R and D are binary, and there are no covariates, Angrist, Imbens, and Rubin (1996) showed that under (i) Stable Unit Treatment Value Assumption (SUTVA), (ii) Random treatment assignment, (iii) exclusion

restriction, (iv) instrument condition, and (v) monotonicity assumption, the IV estimator corresponds to the following ratio:

$$\delta = \frac{E(Y|R=1) - E(Y|R=0)}{\Pr(D|R=1) - \Pr(D|R=0)} = \frac{\bar{Y}_{R=1} - \bar{Y}_{R=0}}{\bar{D}_{R=1} - \bar{D}_{R=0}} = \Delta^{Wald}$$

Which is the Wald estimator in case of constant treatment effects, and LATE in case of heterogeneous treatment effects.

1.5 Limitations of experimental evaluations

The experimental evaluations have numerous advantages. First, they "solve" mathematically the selection problem to produce unbiased estimates of program effects. Second, they allow to observe the results of new interventions that were not previously implemented.[9] Finally, experimental evaluation results are easy to convey to larger audiences and policymakers because of the simplicity of the estimator, which is just a difference in means.

However, experimental evaluations also present limitations. The problems with experiments can be classified into three groups: (i) implementation problems, (ii) threats to internal validity, (iii) threats to external validity.

The implementation problems are usually closely related to ethical considerations. Ethical issues in experimental evaluations may arise, for instance, in situations when the experimental treatment can be potentially harmful. On the other hand, if the program is perceived as beneficial, the administrators or policymakers might find it difficult to deny the treatment to a group of potential beneficiaries selected at random. To overcome this problem, experimental evaluations can be implemented at the pilot stage of program implementation when the exclusion of the control group is temporary, and everyone is offered the treatment once the pilot is evaluated. It is also worth noting that in the context of over-subscription, i.e., a situation in which the number of potential beneficiaries is larger than the number of

[9] Sometimes nonexperimental evaluations are also called "observational studies."

possible treated individuals, randomizing treatment assignment is probably the fairest way to allocate the program across potential beneficiaries.[10]

Regarding internal validity, the most common threats include imperfect compliance and substitution. This is usually addressed by estimating the LATE parameter, as discussed in the previous section. Another problem that may arise is attrition. It occurs when not all program participants report on the outcomes and the attrition rates differ by treatment status. Inverse probability weighting and estimation of treatment effect bounds are some methods to address non-random attrition.[11] Other threats are related to participants' behavioral responses, which happen because people may change their behavior if they perceive that they are being observed. These effects are known as the Hawthorne and John Henry effects.[12]

Recently, more emphasis is being placed on the external validity of experimental evaluation estimates (Achie, 2019). The main criticism is that the results obtained in a specific population group would not necessarily generalize to other populations or other contexts. In addition, when replicating the program in a different country, in a different population group, or scaling up, it is usually necessary to introduce changes in the program design to address the specificities of the new context. Also, the estimated program effect is a partial equilibrium effect and does not reflect the general equilibrium effect of a policy change.[13] Arguably, larger experiments may displace non-participants and full-scale programs may change relative prices.

In sum, experimental evaluations have limitations and recently have received criticism.[14] However, as of today, experimental evaluation is still the best we have for identification of causal program effects. As it was shown in the seminal research by LaLonde (1986), it may be challenging to find non-experimental evaluations able to produce unbiased estimates of program effect.[15] While it is true that experiments cannot answer all questions, it is

[10] This is the context of the experimental evaluations presented in this dissertation in Chapter 2 and Chapter 3.
[11] See, e.g., Anderson et al. (2021).
[12] See, for example, Sedgwick and Greenwood (2015), and Saretsky (1972).
[13] Heckman, Lochner, and Taber (1998).
[14] See, e.g., Teele (2014).
[15] For a recent review see the footnote nine in Harding et al. (2021).

also the fact that many other impact evaluation techniques are even not able to produce an ATE estimate. Moreover, the limitations receiving criticism are not only circumvented to experiments, but they are also a problem in other impact evaluation methods. For that reason, regardless of the impact evaluation method at use, researchers should be aware of these limitations, try to anticipate them and mitigate their consequences to the extent possible.

1.6 Chapter summaries

This dissertation is organized as follows. The first introductory chapter presents the argument for rigorous evidence necessary for better decision-making. It also lays out the methodological framework for the empirical work presented in this dissertation.

Chapter 2 presents the results of the large-scale experimental evaluation of improvements to center-based childcare in Bolivia through the "Grow Well to Live Well" (GWLW) program, implemented by the Bolivian Ministry of Health. The main contribution of this study is that it provides evidence from a randomized controlled trial implemented at a large scale in disadvantaged communities in a developing country with limited institutional capacity. The main results of the study show that the program was highly effective in improving the quality of childcare centers, with investments in process quality explaining most of the improvements and at a substantially lower cost when compared to the investments in infrastructure. The departures from the program protocol in the assignment of the infrastructure investment treatment resulted in the impossibility of estimating causal effects of these investments. The study leaves a promising opportunity for future research, specifically for evaluating the program impact on child development. Given the large changes in the childcare quality indicators, there is a potential for the program also to generate impacts on child development.

Chapter 3 presents the results of the experimental evaluation of the second operational arm of the GWLW, which consisted of a home-visiting program for children under three years of age. This study is an impact evaluation that uses an experimental method exploiting the random assignment of neighborhoods and communities into treatment and control groups. The main finding of the evaluation is that the program achieved significant and large

impacts on child development in rural areas measured by a comprehensive set of child development scales. These results were most likely realized through improved caregiver-child interactions and better cognitive stimulation practices attributable to the program. A very low program take-up in urban areas resulted in no observed effects on children in urban households. This evaluation provides an important case for a better understanding of the intervention context and implementational challenges so that they can be accounted for when scaling up or designing new programs.

Chapter 4 presents the results of a quasi-experimental evaluation of an important educational policy: an increase in educational spending. The policy impact is estimated using administrative data from Brazil. The main contribution of this study is that it provides causal IV estimates for identification of the effect of unrestricted or unconditional educational spending on student achievement. The identification strategy in this evaluation exploits the allocation mechanism of the federal transfer in which the amounts transferred to Brazilian municipalities depend on their population size, generating exogenous jumps in transferred funds at specific population thresholds. The main finding of this study is that the increase in discretional educational spending translates into an increase in student scores. The study did not find evidence that improvements in test scores were achieved through improvements in traditional school inputs, such as class size or teacher level of education. A comparison of costs with other interventions aiming at increasing student achievement shows that the evaluated policy is a relatively cost-effective way to improve student outcomes but not the most efficient.

Chapter 5 presents the general conclusions drawn from the studies included in this dissertation. It also outlines the policy implications stemming from the studies' results, discusses some practical challenges of impact evaluation implementation, and identifies potential areas for future research. The chapter also reflects on the need for evaluating already implemented policies and programs, making a point that quasi-experimental evidence generated by observational studies is as needed as experimental evaluations of the interventions amenable to randomization. The chapter also emphasizes developing rigorous evidence on the program costs for informing policy decisions. The main message of the chapter is that finding solutions to make

skills development interventions more effective, reduce costs, or both, would help developing countries channel resources in the right directions to improve human capital and foster development.

Chapter 2: The Effects of Improving Child Care Centers: Evidence from Bolivia

2.1 Introduction[16]

High quality center-based childcare and other forms of early childhood interventions are central to children's psycho-social and cognitive development and have proved to generate important lifelong gains (Garcia et al., 2021; Gertler et al., 2021). However, public policies that effectively improve the quality of interactions between caregivers and children in daycare centers have proven elusive. This challenge is particularly acute in low-income settings where qualified personnel are scarce. And while much of the literature has focused on the effects of expanding childcare or preschool coverage, there is less evidence on cost-effective ways to boost the quality of existing services.

This study presents the results of a large scale randomized controlled trial of improvements to center-based childcare in Bolivia through the "Grow Well to Live Well" (GWLW) program, implemented by the Bolivian Ministry of Health. Public childcare services in Bolivia are implemented and managed by municipal governments. Prior to the intervention, the average quality of care in the program centers was 1.25[17] according to the Infant/Toddler Environment Rating Scale®-Revised Edition (ITERS-R), amongst the lowest in Latin America (Araujo et al., 2015). Many centers had precarious infrastructure and were staffed by community caregivers (CCs). CCs were typically mothers of children attending the center, had no formal training in child development, and had an average education of less than primary school.

The GWLW program targeted existing childcare centers in poor rural and suburban communities. The program sought to improve the quality of services in existing centers, rather than financing construction of new ones.

[16] This chapter is based on the impact evaluation of the "Grow Well to Live Well" early childhood intervention in Bolivia funded by the Inter-American Development Bank (IDB). The impact evaluation report is published as an IDB Technical Note 1792. This program was conducted at the IDB under the direction of Julia Johannsen and Sebastian Martinez, Cecilia Vidal supported the tasks of data collection and acquisition.

[17] Pre-program baseline report (Bedregal et al., 2016).

The program provided new furniture and teaching materials for classrooms, and CCs were given a structured curriculum as well as training and supervision. Some centers also received infrastructure upgrades, which consisted in refurbishing existing buildings or building new facilities when existing infrastructure was considered inoperable.

Of 158 childcare centers identified in the study areas, the program randomly assigned 79 centers to treatment and 79 to control. Due to tenancy requirements unknown at the time of random assignment, only 48 treatment centers were legally entitled to receive the infrastructure upgrades. These 48 centers received all program components: (i) the infrastructure upgrades, (ii) personnel support (structure curriculum, training and supervision) and equipment (furniture, materials). The remaining 31 treatment centers received only personnel support and equipment.

This study contributes to the literature by providing the evidence from an experimental evaluation of a center-based childcare program implemented at scale in vulnerable areas of a developing Latin American country. The variation stemming from the randomized assignment in treatment and control groups allows rigorous estimation of the program treatment effects and identification of the impacts of the program on the quality of childcare. In addition, this study also provides quasi-experimental evidence and analyzes the cost-effectiveness of the infrastructure program component compared to the training and coaching program component, adding to few existing studies in developing countries evaluating each of these components separately (Bernal, 2015; Yoshikawa et al., 2015; Ozler et al., 2018; Bernal et al., 2019). This analysis uses a quasi-experimental variation in intervention components from the building tenancy requirements to estimate the marginal cost-effectiveness of infrastructure investments compared to the basic package of program components consisting of equipment and personnel improvements.[18] The study also uses various outcome measures of quality, including the ITERS-R, Arnett Caregiver Interaction Scale (CIS), Knowledge of Infant Development Inventory (KIDI), and tailored center and personnel questionnaires. Finally, the study also contributes to the discussion in the

[18] Throughout this study, we refer to the infrastructure component effect and costs as "marginal", because the infrastructure upgrades were implemented in addition to the basic program package.

literature on the scalability of early childhood interventions (Araujo et al., 2021), showing that the "cascade" training and supervision model used by the program can be effective in achieving improvements in childcare quality in an intervention implemented by community mothers without formal training in early childhood education.

The main fundings show that the GWLW program was highly effective in improving center quality with an impact of two standard deviation improvement in the main quality indicator. Investments in process quality explain most of the improvements and do so at approximately one-sixth the cost of marginal investments in infrastructure. These results suggest that, conditional on minimal infrastructure standards, process-related investments are highly cost-effective for improving childcare quality. The relevance of these findings for public policy decisions is highlighted in Egert et al. (2018), who finds that impacts of process quality are significant predictors of impacts on child development.

The reminder of this chapter is structure as follows. Section 2.2 discusses the relevant literature. Next, Section 2.3 presents the intervention and Section 2.4 describes the data and statistical analysis. Finally, Section 2.5 reports the results and Section 2.6 concludes. Some additional results and information are presented in Section 2.7 that acts as an Appendix.

2.2 Literature review

A growing body of evidence indicates that foundations for healthy and productive lives are formed at the very early age. Inadequate health and nutrition, parenting practices with limited interactions between parents and children, home environments with few books, toys and lack of other learning opportunities, can negatively affect cognitive and socioemotional development of children. Early developmental deficits can have lifelong consequences, including lower levels of school attendance and performance, lower income in adulthood, greater dependence on the health care system and higher crime rates (Naudeau et al., 2011; Walker et al., 2011). On the other hand, appropriately designed early childhood interventions can generate positive and sustainable development results (Engle et al., 2011; Gertler et al., 2014; Hoddinott et al., 2008). Providing quality center-based childcare is

particularly relevant for development of vulnerable children born into disadvantaged families, given the evidence that socioeconomic level of the household limits child's possibilities of physical and mental development from birth (Lozoff et al., 2006; Rubio-Codina et al., 2015; Schady et al., 2015).

The evidence on the effectiveness of center-based childcare suggests that the results vary according to the quality of the offered services. In the United States, evaluations of the high-quality demonstration childcare programs focused on vulnerable children found positive long-term impacts on health, education and employment (Campbell et al., 2008; Campbell et al., 2014; Heckman et al., 2010). On the other hand, evaluations of large-scale programs where high-quality standards are difficult to sustain found negative effects, specifically on socioemotional development (Baker et al., 2015; Gupta and Simonsen, 2010; Herbst and Tekin, 2010). Studies in Asia and Africa found that center-based interventions can produce small or negligible results if quality of care is not adequate (Bouguen et al., 2013; Özler et al., 2018). The evidence in Latin America is consistent with the international findings, including mixed results on the impacts on child development of programs implemented at scale, and also on the health and nutrition status of children (Berlinski and Schady, 2015; Leroy et al., 2012; Noboa-Hidalgo and Urzúa, 2012; Rosero and Oosterbeek, 2011).

This study contributes to the literature by bringing new evidence on the effectiveness of improving center-based early childhood care quality. The results are generated from a rigorous experimental evaluation of a program implemented at scale in disadvantaged areas of a developing country, which makes them relevant for informing policy decision central for social-emotional and cognitive development of children in disadvantaged settings with low institutional capacity.

2.3 Intervention

The GWLW program was established in 2012 with the objective of promoting child development in low-income populations of the departments of Chuquisaca and Potosí in Bolivia. The program was implemented by the Ministry of Health with the support from the Inter-American Development

bank. The GWLW included three operational arms (1) child stimulation centers for the treatment of children with developmental delays, (2) a home-visiting program that promoted caregiver-child interactions through home visits in communities without alternative childcare services and (3) a program to improve the quality of existing center-based childcare in communities with childcare centers. This evaluation focuses on the third operational arm. In particular, the evaluated intervention financed two main components: (i) facility infrastructure improvements of existing centers through refurbishments of buildings or new construction and (ii) equipment and personnel improvements, which consisted of providing learning materials; providing or replacing childcare center equipment, such as appliances, cookware, furniture; training of new and existing personnel; hiring of early childhood specialist coaches and nutritionists, and guidelines for caregiver-child interactions and nutrition.

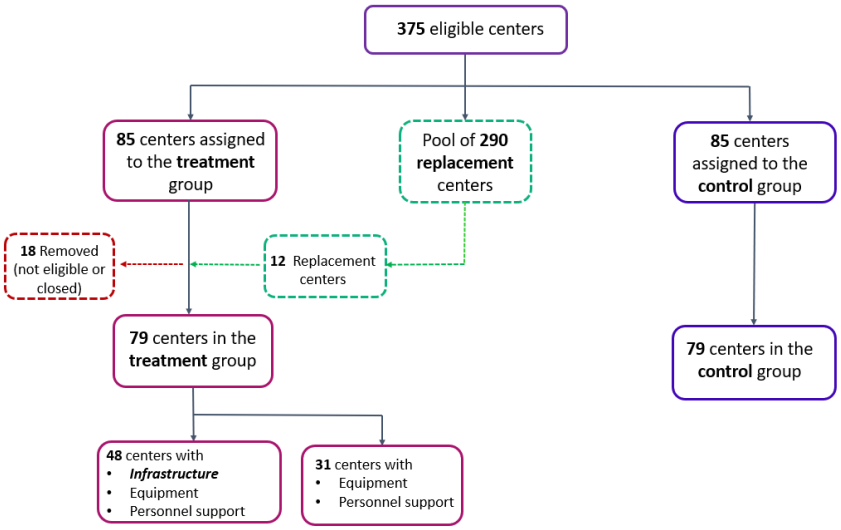
Because the resources of the program were limited, the intervention prioritized municipalities with higher level of poverty and health needs, and with eligible health networks for other GWLW program operational arms. Within the eligible municipalities, a lottery assignment mechanism was implemented to determine the childcare centers that would receive the intervention. The randomization was performed in 2012, but the interventions were implemented between 2015-2018.[19] The implementation of the personnel training and support intervention started in 2015. In this year, about half of the centers received facilitator and training of caregivers. The number of centers that received these activities increased in 2016 and reached 100% in 2017. From the beginning of the program, the support of nutritionist was not provided in all centers and was discontinued in 2018. The infrastructure upgrades were delivered between May 2017 and June 2018, and most of them were completed by August 2017. The equipment upgrades were implemented between July and December of 2017.

The program was implemented as a Randomized Controlled Trial. In total, 375 childcare centers from 36 municipalities were eligible to participate in the study. Within each program municipality, eligible centers were randomly

[19] The first years of the program were focused on establishing cooperation agreements with local communities and developing program guidelines and manuals based on the Reach Up methodology (Gertler et al., 2014; Walker et al., 2011).

assigned the order of entrance to the program and listed accordingly. Initially, based on available program resources, 85 childcare centers on top of the list were assigned to the treatment group and 85 centers from the bottom of the list were assigned to the control group. The centers listed in between treatment and control group formed the pool of replacement centers. In municipalities with only one childcare center the centers were directly assigned to receive the intervention and were excluded from the study sample. Randomization was carried out through public lotteries witnessed by the local authorities and the representatives of the Bolivian Ministry of Health to ensure transparency and legitimacy. Before program implementation, 18 noneligible centers were removed from the treatment group and substituted by 12 centers from the replacement pool. The final study sample comprised 79 centers in the control group, which did not receive any intervention, and 79 centers in the treatment group, which received either only personnel and equipment improvements (31 centers), or personnel, equipment and infrastructure upgrades (48 centers). The flow of centers' assignment to treatment is shown on Figure 2.1.

Figure 2.1: Flow of participating centers included in the study



Source: Authors' calculations

All 79 treated centers received the personnel and training improvements. Due to tenancy requirements unknown at the time of random assignment, only 48 treatment centers were legally eligible to receive the infrastructure upgrades. These 48 centers received all program components: (i) the infrastructure

upgrades, (ii) personnel support (structure curriculum, training and supervision) and equipment (furniture, materials). The remaining 31 treatment centers did not comply with the legal tenancy requirements and received only personnel support and equipment. In this study, we refer to the centers that received infrastructure upgrades as beneficiaries of the "infrastructure" component of the program. The remaining centers are referred as beneficiaries of the "soft" program component.

2.4 Data and method

2.4.1 Data

The data used in this evaluation was retrieved from the program administration records, and the baseline and the endline surveys. The program administration data includes the information on the eligible and beneficiary childcare centers, including their names, center municipality, assignment to treatment or control group, the program take-up. The baseline survey was administered between February and May 2014 in a random sample of 100 childcare centers (50 treatment and 50 control centers), with the purpose of providing baseline information of the targeted population and checking the validity of the experimental design.[20] To measure childcare quality, the survey included three main instruments: ITERS-R (Harms et al., 2006), CIS (Arnett, 1989), and the Caregiver Observation Form and Scale (COFAS) (Fiene, 1984). In addition, the survey included a tailored center and personnel questionnaire inquiring about center's infrastructure, availability of equipment and learning materials, and characteristics of the center's operations and personnel. For a random sample of children in each center, the survey also measured child development through the development screening instrument Ages and Stages Questionnaire (ASQ-III).[21]

The follow-up survey was carried out between April and May 2018 with the objective of evaluating GWLW program effects on childcare quality. Because the time span after implementation of some of the program interventions was too short, the follow-up survey did not evaluate child

[20] The findings from the baseline survey are presented in Bedregal et al. (2016).
[21] https://agesandstages.com/products-pricing/asq3/ (accessed on December 23, 2021).

development outcomes.[22] The follow-up survey was intended to be implemented in all centers assigned to the treatment group and the equivalent number of control centers, identified using the random ordering of the experimental design. Centers from two municipalities were excluded from the follow-up survey due to lack of variation in treatment within municipality.[23] The final survey was aimed to reach 69 treatment centers and 69 control centers. 97 centers (68% of 138) were surveyed. The non-response was due to center closure (6 centers in the treatment group and 31 centers in the control group) and misreporting of preschools as childcare centers (4 centers in the control group). Valid replacement centers from the replacement pool were found for 14 non-response control centers. The final follow-up evaluation sample consisted of 111 centers.

The follow-up survey was similar in its content to the baseline survey and consisted of a comprehensive set of instruments to evaluate various dimensions of care quality: ITERS-R, CIS, and KIDI (MacPhee, 1981). All instruments have adequate internal consistency and concurrent validity (Lopez Boo et al., 2019; Colwell et al., 2013; Al-Maadadi and Ikhlef, 2015), and measure structural and process dimensions of quality.[24] Table 2.1 summarizes these three instruments.

Table 2.1. **Main characteristics of quality instruments**

	Infant/Toddler Environment Rating Scale-Revised Edition (ITERS-R)	**Arnett Caregiver Interaction Scale (CIS)**	**Knowledge of Infant Development Inventory (KIDI)**
Objective	Measures structural and process aspects of childcare quality	Measures child-caregiver interaction styles	Measures caregiver knowledge about child development milestones
Age range	0-30 months	36-60 months	Up to 3 years

[22] On average, the survey was implemented three years after beginning of the personnel support intervention, nine months after infrastructure investments, five months after the provision of equipment, and one month after the delivery of learning materials.
[23] These were the municipalities Potosí and Ravelo. In Potosí, none of the control centers were operating, in Ravelo, none of the treatment centers received the intervention.
[24] In addition to high internal consistency and concurrent validity, the selection of the quality instruments also considered the following criteria: (i) suitability for measurement of center-based quality in a target population of children under 4 years of age; (ii) availability of evidence of successful implementation in Bolivia or similar contexts; (iii) possibility to be implemented under time and training limitations.

Dimensions/ scales	1. Space and Furnishings 2. Personal Care Routines 3. Listening and Talking 4. Activities 5. Interaction 6. Program Structure 7. Parents and Staff	1. Sensitivity 2. Harshness 3. Detachment 4. Permissiveness	1. Single dimension
Type of variables	Structural and process	Structural and process	Structural
Number of items	39	26	58
Assessment method	Observation checklist and interview with center coordinator	Observation checklist	Interview with caregiver and center coordinator
Scoring scale	1 (inadequate quality) to 7 (excellent quality)	1 to 4	0-1
Application time	3.5 hours	1.5 hours	15 minutes
Days of specialized training	3	1	1

Source: Lopez Boo et al. (2016) and Lopez Boo et al. (2019).

The ITERS-R has seven dimensions, one of which focuses on structural aspects of quality ("Space and Furnishings"), while the other six are related to process aspects of quality. Particularly, "Interaction" dimension measures interactions between children and caregiver (Lopez Boo et al., 2016). The ITERS-R data was collected through direct observations of the center's characteristics and activities, as well as through structured interviews with caregivers and center coordinators. ITERS-R contains 39 items that are scored from one (inadequate quality) to seven (excellent quality). Each dimension score is computed as the average score across items within dimension. The overall ITERS-R score is the average score across seven dimensions.

The CIS instrument is mainly used to measure the quality of caregiver-child interactions. It is a checklist of 26 items evaluating four interaction styles of caregivers: sensitivity, harshness, detachment, and permissiveness. The evaluation is based on the frequency of certain caregiver behaviors. Each item of the instrument corresponds to a specific behavior and is scored on a scale between one (never observed) and four (very much observed). Each interaction style score is computed as the average of the corresponding behavior item scores.

The KIDI instrument was implemented to identify the level of knowledge of caregivers of the developmental processes and standards of children's behavior. The KIDI inventory contains 58 statements about parental practices, developmental norms and milestones. The caregivers are asked to indicate whether they agree, disagree or are unsure about the statements. The incorrect responses are scored zero and correct responses are scored one. The final score is the proportion of the correctly answered statements.

In addition to the instruments described above, the follow-up survey included a tailored questionnaire administered to center coordinators. The questionnaire included modules that provided complementary information on the center organization (number of children, group composition, child-caregiver ratios); administrative functions (planning, monitoring and communication with parents); access to basic services; physical infrastructure; and personnel (education, experience, type of contract, level of satisfaction).

Both, the baseline and the follow-up surveys were carried out by a specialized data collection firm with monitoring by the external independent data quality assurance consultant. The survey protocol obtained the Institutional Review Board authorization from the National Committee of Bioethics.

2.4.1 Method

This study uses the randomization design of the program as the key feature that allows to identify the impacts. Given the experimental design in which childcare centers were assigned to the treatment and the control groups based on the random lottery order, the identification strategy is based on comparison of the results between centers in the treatment group and control group. The effects are estimated in the reduced from specification, where the outcome indicator is regressed on the assignment to treatment indicator and

municipality (treatment assignment strata) fixed effects.[25] This estimation allows to identify the Intent-to-treat (ITT) effect parameter.[26]

Specifically, we estimate the following equation:

$$Y_{ij} = \alpha + \beta T_{ij} + \gamma_j + \varepsilon_{ij} \quad (1)$$

Where Y_{ij} is the outcome variable for a childcare center *i* in municipality *j*, T_{ij} is the indicator of assignment to treatment equal to one if a childcare center was assigned to receive the program and zero if a childcare center was assigned to the control group, α is the intercept, γ_j is the municipality-specific fixed effect, ε_{ij} is the error term. The parameter of interest is β, which measures the ITT effect on the outcome variable between treatment and control group.

In addition, to explore the differential association of the program's infrastructure component, we estimate the specification where the infrastructure component indicator is interacted with the assignment to treatment indicator. For this purpose, we estimate the following equation:

$$Y_{ij} = \alpha + \beta T_{ij} + \beta_I T_{ij} * I_{ij} + \gamma_j + \varepsilon_{ij} \quad (2)$$

Where I_{ij} is the indicator that childcare center *i* in municipality *j* received the infrastructure component of the program and $T_{ij} * I_{ij}$ is the interaction between the indicator of the assignment to treatment and the indicator of the infrastructure component. The parameters of interest are β and β_I, which measures the association between the outcome variable and the soft program component, and the marginal contribution of the infrastructure component,

[25] The municipality-specific fixed effect is included because, in the case of this study, the probability of the center to be in the treatment group and control group across municipalities is not the same (Raudenbush et al., 2012).

[26] In the absence of selective attrition, the ITT effect parameter is an unbiased estimate of the average treatment effect (ATE). In the context of the present evaluation, there are indications of non-random attrition with better centers remaining in the control group at follow-up, as well as non-perfect compliance (see Johannsen et al. [2019a] for details). In this context, ATE cannot be recovered from the reduced form, and the estimated ITT parameter is the lower bound of the program effect.

respectively. In all regressions estimated for this study, the standard errors are robust and not clustered at the municipality level, as the randomization was done at the center, not at the municipality level (Abadie et al., 2017).

Finally, the analyses are complemented with the cost-effectiveness analysis of the program soft component and marginal contribution of the infrastructure component. To this end, first, the average per center cost of the soft program component and the average per center cost of the infrastructure program component are computed. Then, the estimate of β from equation (2) is used to compute the cost-effectiveness ratio for the soft intervention component, and the estimate of β_I from equation (2) is used to compute the cost-effectiveness ratio for the marginal contribution of the infrastructure component. Then, the quotient between the former and the latter ratios are computed to compare the cost-effectiveness of the soft program component and the marginal contribution of the infrastructure component. In the results' tables, the cost-effectiveness analysis results are shown only when the estimates of β and β_I are statistically significant at 1%-10% level of statistical significance. All program costs are expressed in 2018 US dollars.

2.5 Results

2.5.1 Baseline balance tests

As discussed in Section 2.4, the baseline survey collected the information only for a sample of 100 centers. During the intervention, some centers were excluded from the study and some centers closed. For that reason, the number of centers for which the baseline and the follow-up survey information is available is 65 (42 centers in the treatment group and 23 centers in the control group). To verify the experimental design validity, we perform the balance tests in pre-treatment characteristics of the centers for this sample. The results are reported in Tables 2.7.A1-2.7.A5 of Appendix 2.7.A.

The balance test results show that centers in treatment and control groups are similar in baseline care quality. The means are not statistically different in any of quality measure (Tables 2.7.A1-2.7.A3) and no statistically significant differences are observed in child development (Table 2.7.A4). There is also a balance in most indicators from the questionnaire administered to center

coordinator, except for some administration indicators and the indicators related to personnel characteristics.

2.5.2 Balance between centers with and without infrastructure component

As mentioned in Section 2.4, not all centers were legally eligible to receive the infrastructure upgrades. The main reason why some centers did not receive infrastructure was that they lacked the tenancy requirement, which is an administrative issue and, a priory, should not be related to center characteristics. If this is the case, then the infrastructure could be considered as quasi-randomly assigned. To sustain this assumption, we perform the tests of balance on observable characteristics of the centers that obtained both infrastructure and soft program component and the centers that received only the soft program component. As shown in Appendix 2.7.B, the centers that received infrastructure and soft component, and the centers that received only soft component are balanced on observable characteristics at baseline, except for one ITERS-R dimension, one CIS interaction style and some caregiver indicators. While we cannot assert that there were no factors that could have influenced the receipt of the infrastructure component, the general balance on the observed characteristics between centers with and without infrastructure component corroborates the assumption that the receipt to the infrastructure upgrades was quasi-random.

2.5.2 Program effects on quality

The estimations of the program impact on quality indicators are reported in Tables 2.2 - 2.5. We begin by presenting the effects on quality measured by ITERS-R and CIS instruments. Then, we present the effects of the program on the personnel indicators from the questionnaire and caregiver knowledge measured by KIDI inventory. Finally, we present the results for the effects of the program on center infrastructure, service and administration indicators.

All results' tables have the same structure. The columns indicate the outcome for which the effect is estimated. In Panel A we show the estimates of the β from equation (1), in Panel B we show the estimates of β and β_I from equation (2), in Panel C we show the results of the cost-effectiveness analysis,

and in Panel D we report the control group mean, standard deviation, and the sample size.

2.5.2.1 ITERS-R

The ITERS-R measures the structural and process quality in childcare centers and is the main outcome indicator in this study. As shown in Table 2.2, the program effect on the average ITERS-R score is 0.88 points. This effect is large (two standard deviations or 62% increase with respect to the control mean) and statistically significant at 1% level of statistical significance. The analysis of the differential association between the infrastructure component and the average ITERS-R score shows that the effect of the program is largely driven by the soft program component, with the estimate of 0.68 ITERS-R points. The estimate on the interaction between the infrastructure component and the assignment to treatment indicator is about twice as small and is only marginally statistically significant (10% level). The cost-effectiveness analysis shows that an increase in the care quality by one ITERS-R point produced by marginal infrastructure upgrades requires more than six times (1/0.15) as much resources as the same increase produced by personnel support and equipment upgrades.

The first dimension of the ITERS-R instrument "Space and Furnishing" is the only ITERS-R indicator measuring structural aspects of care quality. All other dimensions measure process aspects. For this dimension we find that the program had a very large and highly statistically significant effect of 1.158 points. In this dimension, the contribution of the infrastructure component to the overall effect is almost twice as large as the contribution of the soft component, with coefficient estimates in panel B of 1.04 points and 0.625 points, respectively. The analysis of cost-effectiveness shows that it is about two times (1/0.44) as costly to increase the ITERS-R score by one point in this dimension implementing marginal infrastructure upgrades then by implementing personnel support and equipment upgrades.

The remaining six dimensions of ITERS-R are the indicators measuring process aspects of quality. In all six indicators we find large and statistically significant effects ranging from 0.448 for "Parents and Staff" dimension to 1.51 for "Interaction" dimension. In all dimensions except for "Personal Care

Routines" we find that the effects are mostly driven by the soft intervention component. For "Personal Care and Routines" dimension we find that the estimate on the interaction between the infrastructure component and assignment to treatment is statistically significant at 1% level of statistical significance and is almost four times as large as an estimate for the soft intervention component. This result can be explained by the fact that the items in this dimension evaluate the aspects of the process quality which are highly dependent on the quality of infrastructure and materials.[27] The cost-effectiveness analysis for this dimension shows that the cost of increasing ITERS-R score by one point is roughly the same for the soft program component and for additional infrastructure upgrades (1/1.01).

[27] The items include adequate space and materials for delivery of meals, availability of suitable space for sleep and diaper change, sufficient conditions for adequate hygiene practices and essential medicines supply, deficiencies in the infrastructure that may result into health and life risk of children, availability of necessary equipment to provide adequate response in case of emergency.

Table 2.2: Results for ITERS-R, average and by dimension

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Average	Space and Furnishings	Personal Care Routines	Listening and Talking	Activities	Interaction	Program Structure	Parents and Staff
Panel A: Program Effect								
ITT=1	0.881***	1.158***	0.494***	1.136***	0.762***	1.510***	1.391***	0.448***
	[0.127]	[0.135]	[0.114]	[0.236]	[0.126]	[0.308]	[0.219]	[0.114]
Panel B: Differential								
Soft (ITT=1)	0.678***	0.625***	0.167*	1.069***	0.559***	1.390***	1.309***	0.424**
	[0.174]	[0.146]	[0.086]	[0.311]	[0.166]	[0.435]	[0.309]	[0.185]
Infrastructure (ITT*Infrastructure)	0.396*	1.040***	0.638***	0.131	0.398*	0.234	0.160	0.048
	[0.222]	[0.230]	[0.180]	[0.369]	[0.217]	[0.567]	[0.405]	[0.214]
Panel C: Cost Effectiveness								
Soft C/E	56,925	61,756	231,051	36,107	69,138	27,784	29,498.95	9,1046.64
Infrastructure C/E	368,900	140,387	228,608	-	366,998	-	-	-
Soft/Infrastructure	0.15	0.44	1.01	-	0.19	-	-	-
Panel D: Summary Statistics								
Control mean	1.423	1.608	1.133	1.687	1.181	2.021	1.326	1.423
Control SD	(0.425)	(0.493)	(0.340)	(0.954)	(0.331)	(1.210)	(0.682)	(0.405)
N	111	111	111	111	111	111	111	111

Notes: This table presents estimates (Panels A-C) and statistics (Panel D) for ITERS-R average score and dimensions. Each column corresponds to a separate regression. In Panel A outcomes are regressed on the assignment to treatment indicator. In Panel B outcomes are regressed on the assignment to treatment and the interaction of the assignment to treatment and the indicator of infrastructure component interacted with the assignment to treatment. In Panel C Soft C/E shows the average per center cost of the soft program component divided by the coefficient estimate on ITT from Panel B, Infrastructure C/E shows the average per center cost of the infrastructure program component divided by the coefficient

estimate on ITT*Infrastructure. The ratio Soft/Infrastructure is the quotient of the ratios computed above. Maximum and minimum values of the ITERS-R score (average and in each dimension) are 1 and 7, respectively. The coefficients and standard errors are from OLS regressions that include municipality fixed effects. Robust standard errors are in brackets, standard deviations are in parenthesis. Statistical significance: *10%, **5%, ***1%.

2.5.2.2 CIS

The CIS instrument measures structural and process aspects of quality and is complementary to ITERS-R. The results for this instrument are reported in Table 2.3. We find that the program increased "Sensitivity" interaction style of caregivers in treated centers by 0.48 points (22% of the control mean or 70% of the control group standard deviation) and decreased "Detachment" by 0.51 points (26% of the control mean and 76% of the control group standard deviation). We do not find program effects on "Harshness" and "Permissiveness" styles. The coefficient estimates on the interaction between the infrastructure component and the assignment to treatment indicator in Panel B are not statistically significant, which means that there is no differential association between CIS dimensions and infrastructure upgrades.

Table 2.3: Results for CIS interaction styles

	(1) Sensitivity	(2) Harshness	(3) Detachment	(4) Permissiveness
Panel A: Program Effect				
ITT=1	0.481***	-0.091	-0.510***	0.028
	[0.142]	[0.069]	[0.121]	[0.086]
Panel B: Differential				
ITT=1	0.380*	-0.101	-0.406**	0.023
	[0.204]	[0.079]	[0.181]	[0.110]
Infrastructure (ITT*Infrastructure)	0.198	0.019	-0.203	0.010
	[0.237]	[0.099]	[0.195]	[0.122]
Panel C: Cost Effectiveness				
Soft C/E	101,738	-	95,079	-
Infrastructure C/E	-	-	-	-
Soft/ Infrastructure	-	-	-	-
Panel D: Summary Statistics				
Control mean	2.154	1.558	1.896	2.215
Control SD	(0.687)	(0.371)	(0.666)	(0.448)
N	111	111	111	111

Notes: This table presents estimates (Panels A-C) and statistics (Panel D) for CIS dimensions. Each column corresponds to a separate regression. In Panel A outcomes are regressed on the assignment to treatment indicator. In Panel B outcomes are regressed on the assignment to treatment and the interaction of the assignment to treatment and the indicator of infrastructure component of the program. In Panel C, Soft C/E shows the average per center cost of the soft program component divided by the coefficient estimate

on ITT from Panel B, Infrastructure C/E shows the average per center cost of the infrastructure program component divided by the coefficient estimate on ITT* Infrastructure. The ratio Soft/Infrastructure is the quotient of the ratios computed above. Maximum and minimum values of CIS score are 1 and 4 in each dimension. The coefficients and standard errors are from OLS regressions that include municipality fixed effects. Robust standard errors are in brackets, standard deviations are in parenthesis. Statistical significance: *10%, **5%, ***1%.

2.5.2.3 KIDI and Caregiver indicators

In this section we present the results for caregiver knowledge measured by the KIDI inventory and the caregiver indicators from the personnel questionnaire. The results are presented in Table 2.4. For KIDI inventory score we find that the program had positive and statistically significant effect at 1% level of statistical significance increasing knowledge of caregivers by 0.069 points (14% of the control mean or 75% of the control group standard deviation). We do not find differential association between KIDI score and the infrastructure component of the program.

For caregiver indicators retrieved from the personnel questionnaire we find that the program had marginally statistically significant effects (10% level of statistical significance) on satisfaction of caregivers and on receiving training. The effect estimates on receiving feedback and on the aggregated Caregiver index[28] are 0.163 and 0.121, respectively. Both estimates are statistically significant at 1% level of statistical significance. We find large and statistically significant differential association between the infrastructure component and receiving training, and marginally (10% statistical significance) differential association for the Caregiver index.

Table 2.4: Results for KIDI score and Caregiver indicators

	(1) KIDI	(2) Is satisfied=1	(3) Received training=1	(4) Received feedback=1	(5) Caregiver Index
Panel A: Program Effect					
ITT=1	0.069***	0.113*	0.130*	0.163***	0.121***
	[0.024]	[0.058]	[0.076]	[0.062]	[0.037]
Panel B: Differential					

[28] Caregiver Index is the average of the proportions of caregivers in the center who (i) report being satisfied with their work; (ii) work with the contract; (iii) receive feedback on their work; (iv) had training in child development in the last 3 years.

Soft (ITT=1)	0.064*	0.161**	-0.102	0.155*	0.071
	[0.033]	[0.066]	[0.097]	[0.088]	[0.052]
Infrastructure (ITT* Infrastructure)	0.008	-0.086	0.415***	0.014	0.097*
	[0.040]	[0.081]	[0.112]	[0.114]	[0.054]
Panel C: Cost Effectiveness					
Soft C/E	599,392	239,747	-	249,407	-
Infrastructure C/E	-	-	351,872	-	1,499,357
Soft/ Infrastructure	-	-	-	-	-
Panel D: Summary Statistics					
Control mean	0.486	0.823	0.602	0.750	0.669
Control SD	(0.092)	(0.384)	(0.492)	(0.435)	(0.223)
N	88	176	172	176	111

Notes: This table presents estimates (Panels A-C) and statistics (Panel D) for KIDI score and Caregiver indicators. Each column corresponds to a separate regression. In Panel A outcomes are regressed on the assignment to treatment indicator. In Panel B outcomes are regressed on the assignment to treatment and the interaction of the assignment to treatment and the indicator of infrastructure component of the program. In Panel C Infrastructure C/E shows the average per center cost of the soft program component divided by the coefficient estimate on ITT from Panel B, Infrastructure C/E shows the average per center cost of the infrastructure program component divided by the coefficient estimate on ITT*Infrastructure. The ratio Soft/Infrastructure is the quotient of the ratios computed above. The unit of observation in regressions in columns (1) and (5) is a center. The unit of observation in regressions in columns (2), (3) and (4) is a caregiver. Maximum and minimum values for KIDI score are 0 and 1. The outcomes in columns (2), (3) and (4) are binary indicators. Caregiver Index is the average of outcomes in columns (2), (3) and (4) at the center level (see Appendix 2.7.D for details). The coefficients and standard errors are from OLS regressions that include municipality fixed effects. Robust standard errors are in brackets, standard deviations are in parenthesis. Statistical significance: *10%, **5%, ***1%.

2.5.2.4 Infrastructure, service and administration indicators

The results for the infrastructure, service and administration indicators are reported in Table 2.5. The program had a desirable and statistically significant effect on several indicators of quality of physical environment of the centers. The program increased the probability that the center building is in good condition by 0.381. This effect is very large, 366% of the control mean or about one control mean standard deviation, and statistically significant at 1% level of statistical significance. The program also reduced the probability of the center to be in need of repairs by almost 30 percentage points, which is a

very important result given that 100% of control group centers need repairs. For these two indicators we observe large and statistically significant estimates on the interaction between the infrastructure component and the assignment to treatment indicator, suggesting that infrastructure upgrades were relevant for improvements observed in these indicators.

The coefficient estimate on the Construction material index[29] is also positive and statistically significant. It shows that the program increased the proportion of centers built with adequate construction material by 0.156. For this indicator we do not observe differential association for the infrastructure component. For the indicators of the availability of electricity, garden and good illumination, we find that all three indicators are significantly larger in the group of centers that benefited from the infrastructure component (Panel B results). However, the overall program effect reported in Panel A is statistically significant at the conventional 5% level of statistical significance only for the indicator of garden availability.

The endline survey inquired about availability of the equipment and materials for learning, such as books, toys, music equipment, among others. By aggregating these indicators in a composite Learning material index[30] we find that the program had a significant impact increasing the presence of these inputs in the treated centers: the proportion of center coordinators responding affirmatively to the questions about the presence of each input included in the aggregate index increased by 0.392, which represents an improvement of more than 100% with respect to the control mean. We also find statistically significant differential association between Learning material index and the infrastructure component of the program. The analysis of cost-effectiveness shows that the unitary increase in this indicator by adding infrastructure upgrades is almost eight times (1/0.13) as costly as by implementing the soft

[29] Construction material index takes value one if construction materials of roof, floor and walls are not precarious: roof is made of resistant material which is not wood, straw, mud; floor is not bare earth or loose bricks, walls are plastered. Construction of this index is based on the Precarious Toilet Index in Bancalari et al. (2016).

[30] Learning Materials Index is the proportion of affirmative answers to the following questions: the center has at least: 10 books for young children? 3 or more puzzles? Toys for learning different colors, sizes and shapes? A stereo (radio with CD or recorder) to listen to music? Material to cut, color, draw, etc.? Toys for children to play and imitate? Toys to build, such as blocks, cubes, Lego sets, etc.? Balls? Tricycles, wooden horses, other mountable toys? Musical instruments?

program component. For Childcare development and monitoring index[31] and Curriculum, training and monitoring indices[32] we find positive and statistically significant effects of 0.239 and 0.182, respectively. It is about five times (1/0.2) less costly to achieve the unitary increase in Childcare development and monitoring index by implementing the soft component in comparison to implementing additional infrastructure improvements. We do not find differential effects of infrastructure upgrades for Curriculum, training and monitoring index.

[31] Childcare Development Monitoring Index is the proportion of affirmative answers to the following questions: Does the Center have a record of daily attendance of the children? Does this center periodically record the size and weight of each child? Does this Center periodically record vaccines received by each child? Does this center periodically record general health of each child? Does this center periodically record child development? Does this Center provide information to parents/caregivers about their child development? Does this Center daily inform parents/caregivers about how was the child's day?

[32] Curriculum, Training and Monitoring Index is the proportion of affirmative answers to the following questions: Does this Center have an annual staff training plan? Does this Center have a plan of activities for each room or group? Does this Center plan activities for each child according to child's needs? Does this Center have a pedagogical curriculum? Does this Center have regular evaluations of staff performance? Does this Center have any rules or regulations on what to do in case a child has an accident or a medical emergency?

Table 2.5: Results for center infrastructure, service and administration indicators

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Building in good conditions =1	Building needs repair=1	Construction Materials Quality Index	Electricity=1	Garden=1	Good illumination=1	Learning Materials Index	Childcare Development Monitoring Index	Curriculum, Training and Monitoring Index
Panel A: Program Effect									
ITT=1	0.381***	-0.297***	0.156**	0.001	0.228***	0.169*	0.392***	0.239***	0.182***
	[0.085]	[0.063]	[0.071]	[0.087]	[0.081]	[0.092]	[0.043]	[0.044]	[0.048]
Panel B: Differential									
Soft (ITT=1)	-0.010	-0.043	0.089	-0.153	-0.079	-0.086	0.312***	0.172***	0.131**
	[0.069]	[0.045]	[0.090]	[0.117]	[0.066]	[0.110]	[0.063]	[0.059]	[0.063]
Infrastructure (ITT* Infrastructure)	0.761***	-0.495***	0.130	0.299**	0.599***	0.497***	0.155**	0.131*	0.100
	[0.090]	[0.107]	[0.088]	[0.134]	[0.104]	[0.120]	[0.071]	[0.078]	[0.073]
Panel C: Cost Effectiveness									
Soft C/E	-	-	-	-	-	-	123,637	224,918	295,558
Infrastructure C/E	191,669	294,871	-	488,416	243,486	293,561	939,779	1,117,690	-
Soft/Infrastructure	-	-	-	-	-	-	0.13	0.20	-
Panel D: Summary Statistics									
Control mean	0.104	1.000	0.792	0.792	0.125	0.583	0.358	0.443	0.427
Control SD	(0.309)	(0.000)	(0.410)	(0.410)	(0.334)	(0.498)	(0.299)	(0.259)	(0.279)
N	111	111	111	111	111	111	111	111	111

Notes: This table presents estimates (Panels A-C) and statistics (Panel D) for center indicators. Each column corresponds to a separate regression. In Panel A outcomes are regressed on the assignment to treatment indicator. In Panel B outcomes are regressed on the assignment to treatment and the interaction of the assignment to treatment and the indicator of infrastructure component of the program. In Panel C Soft C/E shows the average per center cost of the soft program component divided by the coefficient estimate on ITT from Panel B, Infrastructure C/E shows the average per center cost of the infrastructure program component divided by the coefficient estimate on ITT* Infrastructure, Ratio Soft/Infrastructure is the quotient of the ratios computed above. See Appendix 2.7.D for description of indices. The coefficients and standard errors are from OLS regressions that include municipality fixed effects. Robust standard errors are in brackets, standard deviations are in parenthesis. Statistical significance: *10%, **5%, ***1%.

2.5.3 Robustness checks and falsification tests

2.5.3.1 Robustness checks

As robustness checks of the main results, we estimate the program effects including the vector of baseline characteristic. Specifically, we estimate equations (1) and (2) including the vector of baseline characteristics on which the baseline sample is unbalanced.[33] Because not all centers were interviewed at the baseline, these estimations are done using the sample of 65 centers for which we have baseline and follow-up survey data. The results are reported in Appendix 2.7.C. As seen in Tables 2.7.C1 – 2.7.C4, including baseline characteristics increased the coefficient estimates for the main measures of care quality, which corroborates that the main ITT estimates (Tables 2.2 – 2.4) are lower bounds of the program effects. [34] The effects estimated for the center infrastructure, service and administration indicators in the sample of 65 centers are qualitatively the same and very similar in values to the estimates in the sample of 111 centers (Table 2.7.C.4).

2.5.3.2 Falsification tests

The program did not include actions aimed at increasing coverage of care, or the number of children who enroll and attend childcare centers. As falsification tests, we estimate the effect of the program on coverage indicators. The results are reported in Table 2.6. The ITT estimates in the sample of 111 centers show that the program had negative effect on enrollment and attendance of children. However, the estimates for these

[33] The vector of baseline controls includes the number of classrooms in the center, the indicator that the center has electricity, the indicator that the center is managed by the Departmental Government, the indicator that the center receives funding from parents; the indicator that the center receives funding from NGOs, center operation hours; proportion of women staff; proportion of staff who receive feedback; proportion of staff who report being satisfied with their work; caregiver index .

[34] Additional robustness checks were performed by estimating the program effects using the difference-in-difference (DID) estimator and by correcting for non-random attrition using the inverse probability weights (IPW). The results of these estimations are qualitatively the same and corroborate that the main ITT results are the lower bounds of the program effect estimates. The DID and IPW estimation results can be made available upon request.

indicators become statistically not significant if pre-treatment characteristics of centers are included in the estimations (Appendix 2.7.C, Table 2.7.C5).

Table 2.6: Results for coverage indicators

	(1) Children enrolled in the classroom	(2) Children present in the classroom	(3) Number of classrooms	(4) Number of caregivers
Panel A: Program Effect				
ITT=1	-6.271**	-4.117**	-0.179	-0.519
	[2.740]	[1.943]	[0.147]	[0.316]
Panel B: Differential				
Soft (ITT=1)	-6.549**	-4.755**	-0.224	-0.527*
	[2.538]	[2.098]	[0.165]	[0.304]
Infrastructure (ITT* Infrastructure)	0.543	1.245	0.088	0.014
	[2.512]	[2.206]	[0.179]	[0.270]
Panel C: Summary Statistics				
Control mean	24.229	16.375	1.479	1.938
Control SD	(20.355)	(13.503)	(1.148)	(2.453)
N	111	111	111	111

Notes: This table presents estimates (Panels A-B) and statistics (Panel C) for coverage indicators. Each column corresponds to a separate regression. In Panel A outcomes are regressed on the assignment to treatment indicator. In Panel B outcomes are regressed on the assignment to treatment and the interaction of the assignment to treatment and the indicator of infrastructure component of the program. The coefficients and standard errors are from OLS regressions that include municipality fixed effects. Robust standard errors are in brackets, standard deviations are in parenthesis. Statistical significance: *10%, **5%, ***1%.

2.6 Conclusions

The program "Grow Well to Live Well" was implemented with the objective to contribute to the improvement of development of children under 4 years through better access and quality of childcare services. The program was conceived as a pilot intervention in two prioritized departments in Bolivia (Chuquisaca and Potosí) in order to evaluate the innovative interventions potentially scalable at the national level. In this study we present the results of the evaluation of the program components which financed interventions in childcare centers. The implemented interventions included provision of materials, development of guidelines and protocols of care, training and support to personnel through specialized facilitators and nutritionists. Some

childcare centers also benefited from improvements in physical infrastructure, which included refurbishment of spaces and, in some cases, complete reconstruction of the centers.

Our results show that the program increased quality of care in the beneficiary childcare centers. For the average ITERS-R score we find that the program improved quality of care by 0.88 ITERS-R points, which is a very large effect equivalent to an increase by two standard deviations. The disaggregation of the score in seven ITERS-R subscales shows that the program had important positive impacts in all dimensions, ranging from 0.448 points in "Parents and Staff" dimension to 1.158 in "Space and Furnishings" dimension. These results are remarkable, but they also show that the *status quo* situation in the absence of the program demonstrated by the control group mean is grim. For ITERS-R we also find differential effect of the infrastructure upgrades in "Space and Furnishings" and "Personal Care Routines" dimensions.

The program had desirable effect on the interaction styles between caregivers and children measured by CIS instrument, increasing "Sensitivity" of caregiver by 0.48 points and reducing "Detachment" by 0.51 points. We do not observe program effects on "Harshness" and "Permissiveness". The caregiver knowledge in treated centers also improved by 0.069 (75% of the control group standard deviation) measured by KIDI inventory. The program also had statistically significant positive effect on proportion of caregivers who received feedback and the aggregated Caregiver index. It is worth noting that these results were achieved for community caregivers (community mothers or *madres comunitarias*) who lack formal training in early childhood education and who were trained and supervised by the professionals hired by the program. These results suggest that the "cascade" training and supervision adopted by the program could be a scalable model in the current operational context in the country, where the municipalities are responsible for hiring personnel in the centers.

The evaluation results also show that the program had significant effect on center infrastructure, service and administration indicators. The probability that the center building is in good conditions increased by 0.381, which is more than 100% of the control mean. The program also decreased the proportion of centers in need of repair (by 0.297) and increased the proportion

of centers built with adequate construction materials (by 0.156). We also find that the program was effective in increasing the proportion of the centers with the garden area by 0.228. For all infrastructure indicators except for "Construction material index", we find that the infrastructure upgrades were relevant for achieving observed results. In addition, the program also improved indicators of service and administration, including availability of learning materials, childcare and child development monitoring, curriculum and staff training and monitoring. After controlling for baseline characteristics, we do not find any increase in coverage indicators, which is in accordance with the program logic, focused on improvements of quality of existing care services rather than on increasing care coverage.

In regard to the program components' cost-efficiency, our analyses show that moving quality indicators in desired direction by implementing only soft program component is more cost-effective than by implementing infrastructure upgrades in addition to the soft component investments. This result indicates that investments in improvement of curricular components, training of caregivers and provision of materials, could be a cost-effective alternative of improving childcare centers quality, which is a relevant insight for policy decision in resource-constrained settings, such as those faced in developing countries.

Taken together, these results confirm that a central government program, designed to support locally run childcare centers, can be highly effective in improving childcare quality. The findings are also informative for the policy debate in developing countries on whether to prioritize center coverage or quality improvements. Our results show that a strong support program implemented in poorly funded childcare centers can have positive results, notwithstanding social returns of increasing coverage. An important point to highlight is a critically low level of centers' quality in the absence of the program, which might helped derive the program effects.

Although this evaluation did not measure child development outcomes, the observed improvements in the indicators of process and structural aspects of quality give positive indications for the potential achievement of the final results in children development in the future. The achievement of these long-term results will critically depend on maintaining or increasing quality

standards in the centers. The experience of the GWLW program is a valuable contribution to inform child development policy. Future evaluations of the centers and their impact on children will show whether the program model is effective to promote child development in vulnerable populations.

2.7 Appendix

2.7.A Balance tests on the assignment to treatment

Table 2.7.A1: Balance in ITERS-R score in centers assigned to treatment and control groups

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dimensions ITERS-R	Mean ITT=1	Mean ITT=0	Difference	Standard Error	P-value	N ITT=1	N ITT=0
Space and Furnishings	1.133	1.330	-0.197	[0.132]	0.14	42	23
Personal Care Routines	1.028	1.167	-0.139	[0.123]	0.26	42	23
Listening and Talking	1.254	1.261	-0.007	[0.157]	0.96	42	23
Activities	1.092	1.180	-0.088	[0.076]	0.25	42	23
Interaction	1.310	1.348	-0.038	[0.195]	0.85	42	23
Program Structure	1.060	1.145	-0.085	[0.080]	0.29	42	23
Parents and Staff	1.160	1.346	-0.185	[0.124]	0.14	42	23
Average ITERS-R score	1.131	1.251	-0.120	[0.100]	0.24	42	23

Notes: The table shows the results of the balance tests comparing centers observed at the baseline assigned to receive treatment (ITT=1) and centers assigned to the control group (ITT=0). The estimated difference between these two groups and the associated standard error is in columns (3) and (4), respectively. Column (5) shows the P-value from the formal test of the difference in column (3) being equal to zero. Columns (6) and (7) show sample size of the subsample of centers assigned to receive treatment and the subsample of centers assigned to the control group, respectively. Maximum and minimum values of the ITERS-R score (average and in each dimension) are 1 and 7. Statistical significance: *10%, **5%, ***1%.

Table 2.7.A2: Balance in CIS score in centers assigned to treatment and control groups

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Interaction styles CIS	Mean ITT=1	Mean ITT=0	Difference	Standard Error	P-value	N ITT=1	N ITT=0
Sensitivity	1.690	1.874	-0.184	[0.185]	0.32	41	23
Harshness	1.593	1.710	-0.117	[0.160]	0.47	41	23
Detachment	2.811	2.576	0.235	[0.235]	0.32	41	23

Permissiveness	2.285	2.391	-0.107	[0.102]	0.30	41	23

Notes: The table shows the results of the balance tests comparing centers observed at the baseline assigned to receive treatment (ITT=1) and centers assigned to the control group (ITT=0). The estimated difference between these two groups and the associated standard error is in columns (3) and (4), respectively. Column (5) shows the P-value from the formal test of the difference in column (3) being equal to zero. Columns (6) and (7) show sample size of the subsample of centers assigned to receive treatment and the subsample of centers assigned to the control group, respectively. Maximum and minimum values of the CIS score are 1 and 4 in each dimension. Statistical significance: *10%, **5%, ***1%.

Table 2.7.A3: Balance in COFAS score in centers assigned to treatment and control groups

Interaction quality and Total score	(1) Mean ITT=1	(2) Mean ITT=0	(3) Difference	(4) Standard Error	(5) P-value	(6) N ITT=1	(7) N ITT=0
Good	0.10	0.17	-0.079	[0.093]	0.40	42	23
Regular	0.07	0.04	0.028	[0.059]	0.64	42	23
Bad	0.10	0.17	-0.079	[0.093]	0.40	42	23
Very Bad	0.74	0.61	0.129	[0.124]	0.30	42	23
COFAS score	-231.02	-273.17	42.150	[64.178]	0.51	42	23

Notes: The table shows the results of the balance tests comparing centers observed at the baseline assigned to receive treatment (ITT=1) and centers assigned to the control group (ITT=0). The estimated difference between these two groups and the associated standard error is in columns (3) and (4), respectively. Column (5) shows the P-value from the formal test of the difference in column (3) being equal to zero. Columns (6) and (7) show sample size of the subsample of centers assigned to receive treatment and the subsample of centers assigned to the control group, respectively. The values in columns (1) and (2) are proportions of centers with the interaction quality according to COFAS in all rows but the last one, which shows COFAS score. The Interaction quality according to COFAS is “Good” if COFAS score is from +30 to +130, “Regular” if scores is from -10 to +29, “Bad” if score is from -99 to -11 and “Very Bad” if scores is from -1560 to -100. Statistical significance: *10%, **5%, ***1%.

Table 2.7.A4: Balance in ASQ score in centers assigned to treatment and control groups

Dimension	(1) Mean ITT=1	(2) Mean ITT=0	(3) Difference	(4) Standard Error	(5) P-value	(6) N ITT=1	(7) N ITT=0
Communication (z-score)	-0.061	0.071	-0.132	[0.099]	0.19	233	166
Gross Motor (z-score)	0.035	0.054	-0.019	[0.097]	0.84	233	166
Fine Motor (z-score)	0.074	-0.052	0.126	[0.099]	0.20	233	166
Problem Solving (z-score)	0.007	0.134	-0.127	[0.101]	0.21	233	166
Socio-individual (z-score)	-0.023	0.063	-0.086	[0.101]	0.39	233	166
Overall (z-score)	0.010	0.077	-0.067	[0.099]	0.50	233	166

Notes: The table shows the results of the balance tests comparing centers observed at the baseline assigned to receive treatment (ITT=1) and centers assigned to the control group (ITT=0). The estimated difference between these two groups and the associated standard error is in columns (3) and (4), respectively. Column (5) shows the P-value from the formal test of the difference in column (3) being equal to zero. Columns (6) and (7) show sample size of the subsample of centers assigned to receive treatment and the subsample of centers assigned to the control group, respectively. Z-scores for all dimensions have been standardized in-sample using all sample as reference. Scores for overall, communication, gross motor, fine motor, problem solving, and socio-individual dimensions were computed using questionnaires adapted from the Ages and Stages Questionnaires, 3rd edition. (ASQ-III). Statistical significance: *10%, **5%, ***1%.

Table 2.7.A5: Balance in the observed characteristics of centers assigned to treatment and control group

Dimension	(1) Mean ITT=1	(2) Mean ITT=0	(3) Difference	(4) Standard Error	(5) P-value	(6) N ITT=1	(7) N ITT=0
A. Administration							
Number of classrooms	1.24	1.78	-0.545*	[0.277]	0.05	42	23
Enrollment	18.48	27.17	-8.698	[5.848]	0.14	42	23
Municipal Government is responsible for center administration (yes=1)	0.64	0.74	-0.096	[0.120]	0.42	42	23
NGO is responsible for center administration (yes=1)	0.31	0.13	0.179*	[0.102]	0.08	42	23
Center receives funding from Department (yes=1)	0.21	0.48	-0.264**	[0.124]	0.04	42	23
Center receives funding from children's parents (yes=1)	0.93	0.74	0.189*	[0.101]	0.07	42	23
B. Childcare and curriculum							
Center has at least: 10 books for young children (yes=1)	0.29	0.39	-0.106	[0.125]	0.40	42	23
Center has toys for learning of different colors, sizes and shapes (yes=1)	0.43	0.52	-0.093	[0.131]	0.48	42	23
Center has an annual staff training plan (yes=1)	0.38	0.55	-0.161	[0.134]	0.23	39	22
Center has a pedagogical curriculum (yes=1)	0.14	0.20	-0.065	[0.108]	0.55	37	20

Center periodically evaluates child development (no=1)	0.74	0.88	-0.139	[0.110]	0.21	35	17
C. Infrastructure							
Construction Material Index	0.71	0.83	-0.112	[0.107]	0.30	42	23
Building needs repair (yes=1)	1.00	0.96	0.043	[0.043]	0.32	42	23
Center has electricity (yes = 1)	0.55	0.83	-0.278**	[0.112]	0.02	42	23
Center has a separated kitchen (yes=1)	0.71	0.74	-0.025	[0.117]	0.83	42	23
Center has garden or green area for play (yes=1)	0.17	0.22	-0.051	[0.105]	0.63	42	23
Center has good illumination (yes=1)	0.57	0.48	0.093	[0.131]	0.48	42	23
D. Caregivers							
The highest level of education is incomplete secondary (yes=1)	0.33	0.45	-0.121	[0.077]	0.12	105	66
Received training in early childhood education (yes+1)	0.43	0.44	-0.014	[0.079]	0.86	101	66
Number of years worked in this center	1.55	1.95	-0.401	[0.449]	0.37	105	65
Works with contract (yes=1)	0.54	0.56	-0.018	[0.079]	0.82	105	66
E. Other indicators							
Opening hour	8.31	8.02	0.288**	[0.135]	0.04	42	23
Closing hour	15.71	16.48	-0.764**	[0.309]	0.02	42	23
Number of hours center operates per day	7.40	8.46	-1.052**	[0.404]	0.01	42	23
Proportion of staff who are women	0.91	0.98	-0.078**	[0.033]	0.02	42	22

Proportion of staff who receive feedback	0.18	0.38	-0.202**	[0.100]	0.05	42	22
Proportion of staff who report being satisfied with their job	0.43	0.72	-0.291***	[0.098]	0.00	42	22
Caregiver Index	0.37	0.55	-0.174**	[0.070]	0.02	42	22

Notes: The table shows the results of the balance tests comparing centers observed at the baseline assigned to receive treatment (ITT=1) and centers assigned to the control group (ITT=0). The estimated difference between these two groups and the associated standard error is in columns (3) and (4), respectively. Column (5) shows the P-value from the formal test of the difference in column (3) being equal to zero. Columns (6) and (7) show sample size of the subsample of centers assigned to receive treatment and the subsample of centers assigned to the control group, respectively. See Appendix 2.7.D for details on construction of indices. Statistical significance: *10%, **5%, ***1%.

2.7.B Balance tests on the infrastructure component in the sample of centers that received treatment

Table 2.7.B1: Balance in ITERS-R score in treated centers with and without infrastructure

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dimensions ITERS-R	Mean Infrastructure=1	Mean Infrastructure=0	Difference	Standard Error	P-value	N Infrastructure=1	N Infrastructure=0
Space and Furnishings	1.155	1.110	0.045	[0.091]	0.63	22	20
Personal Care Routines	1.045	1.008	0.037	[0.046]	0.43	22	20
Listening and Talking	1.364	1.133	0.230	[0.184]	0.22	22	20
Activities	1.161	1.016	0.145***	[0.053]	0.01	22	20
Interaction	1.409	1.200	0.209	[0.249]	0.41	22	20
Program Structure	1.098	1.017	0.082	[0.080]	0.31	22	20
Parents and Staff	1.221	1.094	0.127	[0.094]	0.18	22	20
Average ITERS-R score	1.188	1.069	0.120	[0.078]	0.13	22	20

Notes: The table shows the results of the balance tests in the sample of treated centers observed at the baseline. Columns (1) and (2) show means in the sample of centers that received infrastructure component (Infrastructure=1) and centers that did not receive the infrastructure component (Infrastructure=0). The estimated difference between these two groups and the associated standard error is in columns (3) and (4), respectively. Column (5) shows the P-value from the formal test of the difference in column (3) being equal to zero. Columns (6) and (7) show sample size of the group that received the infrastructure component and the group that did not receive it, respectively. Maximum and minimum values of the ITERS-R score (average and in each dimension) are 1 and 7. Statistical significance: *10%, **5%, ***1%.

Table 2.7.B2: Balance in CIS score in treated centers with and without infrastructure

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Interaction styles CIS	Mean Infrastructure=1	Mean Infrastructure=0	Difference	Standard Error	P-value	N Infrastructure=1	N Infrastructure=0
Sensitivity	1.833	1.540	0.293	[0.208]	0.17	21	20
Harshness	1.593	1.594	-0.002	[0.179]	0.99	21	20
Detachment	2.524	3.112	-0.589**	[0.257]	0.03	21	20
Permissiveness	2.286	2.283	0.002	[0.149]	0.99	21	20

Notes: The table shows the results of the balance tests in the sample of treated centers observed at the baseline. Columns (1) and (2) show means in the sample of centers that received infrastructure component (Infrastructure =1) and centers that did not receive the infrastructure component (Infrastructure =0). The estimated difference between these two groups and the associated standard error is in columns (3) and (4), respectively. Column (5) shows the P-value from the formal test of the difference in column (3) being equal to zero. Columns (6) and (7) show sample size of the group that received the infrastructure component and the group that did not receive it, respectively. Maximum and minimum values of the CIS score are 1 and 4 in each dimension. Statistical significance: *10%, **5%, ***1%.

Table 2.7.B3: Balance in COFAS score in treated centers with and without infrastructure

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Interaction quality and Total score	Mean Infrastructure=1	Mean Infrastructure=0	Difference	Standard Error	P-value	N Infrastructure=1	N Infrastructure=0
Good	0.09	0.10	-0.009	[0.093]	0.92	22	20
Regular	0.09	0.05	0.041	[0.080]	0.61	22	20
Bad	0.14	0.05	0.086	[0.090]	0.34	22	20
Very Bad	0.68	0.80	-0.118	[0.137]	0.39	22	20
COFAS score	-208.45	-255.85	47.395	[64.000]	0.46	22	20

Notes: The table shows the results of the balance tests in the sample of treated centers observed at the baseline. Columns (1) and (2) show means in the sample of centers that received infrastructure component (Infrastructure =1) and centers that did not receive the infrastructure component (Infrastructure =0). The estimated difference between these two groups and the associated standard error is in columns (3) and (4), respectively. Column (5) shows the P-value from the formal test of the difference in column (3) being equal to zero. Columns (6) and (7) show sample size of the group that received the infrastructure component and the group that did not receive it, respectively. The values in columns (1) and (2) are proportions of centers with the interaction quality according to COFAS in all rows but the last one, which shows COFAS score. The Interaction quality according to COFAS is "Good" if COFAS score is from +30 to +130, "Regular" if scores is from -10 to +29, "Bad" if score is from -99 to -11 and "Very Bad" if scores is from -1560 to -100. Statistical significance: *10%, **5%, ***1%.

Table 2.7.B4: Balance in ASQ score in treated centers with and without infrastructure

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dimension	Mean Infrastructure=1	Mean Infrastructure=0	Difference	Standard Error	P-value	N Infrastructure=1	N Infrastructure=0
Communication (z-score)	-0.009	-0.129	0.120	[0.138]	0.39	131	102
Gross Motor (z-score)	0.019	0.055	-0.036	[0.129]	0.78	131	102
Fine Motor (z-score)	0.059	0.094	-0.035	[0.136]	0.79	131	102
Problem Solving (z-score)	-0.084	0.124	-0.207	[0.126]	0.10	131	102
Socio-individual (z-score)	-0.032	-0.012	-0.020	[0.138]	0.88	131	102
Overall (z-score)	0.002	0.019	-0.017	[0.137]	0.90	131	102

Notes: The table shows the results of the balance tests in the sample of treated centers observed at the baseline. Columns (1) and (2) show means in the sample of centers that received infrastructure component (Infrastructure =1) and centers that did not receive the infrastructure component (Infrastructure =0). The estimated difference between these two groups and the associated standard error is in columns (3) and (4), respectively. Column (5) shows the P-value from the formal test of the difference in column (3) being equal to zero. Columns (6) and (7) show sample size of the group that received the infrastructure component and the group that did not receive it, respectively. Z-scores for all dimensions have been standardized in-sample using control group sample as reference. Scores for overall, communication, gross motor, fine motor, problem solving, and socio-individual dimensions were computed using questionnaires adapted from the Ages and Stages Questionnaires, 3rd edition. (ASQ-III). Statistical significance: *10%, **5%, ***1%.

Table 2.7.B5: Balance in the observed characteristics of centers with and without infrastructure

Dimension	(1) Mean Infrastructure=1	(2) Mean Infrastructure=0	(3) Difference	(4) Standard Error	(5) P-value	(6) N Infrastructure=1	(7) N Infrastructure=0
A. Administration							
Number of classrooms	1.25	-0.023	[0.167]	0.89	22	20	1.25
Enrollment	18.45	0.050	[2.889]	0.99	22	20	18.45
Municipal Government is responsible for center administration (yes=1)	0.64	0.65	-0.014	[0.152]	0.93	22	20
NGO is responsible for center administration (yes=1)	0.32	0.30	0.018	[0.146]	0.90	22	20
Center receives funding from Department (yes=1)	0.27	0.15	0.123	[0.127]	0.34	22	20
Center receives funding from children's parents (yes=1)	0.95	0.90	0.055	[0.082]	0.51	22	20
B. Childcare and curriculum							
Center has at least: 10 books for young children (yes=1)	0.32	0.25	0.068	[0.142]	0.63	22	20
Center has toys for learning of different colors, sizes and shapes (yes=1)	0.55	0.30	0.245	[0.151]	0.11	22	20
Center has an annual staff training plan (yes=1)	0.47	0.30	0.174	[0.158]	0.28	19	20
Center has a pedagogical curriculum (yes=1)	0.15	0.12	0.032	[0.115]	0.78	20	17
Center periodically evaluates child development (no=1)	0.76	0.72	0.042	[0.152]	0.78	17	18
C. Infrastructure							

Construction Material Index	0.77	0.74	0.123	[0.143]	0.39	22	43
Building needs repair (yes=1)	1.00	1.00	0.000	[0.000]	.	22	20
Center has electricity (yes = 1)	0.64	0.45	0.186	[0.155]	0.24	22	20
Center has a separated kitchen (yes=1)	0.73	0.70	0.027	[0.143]	0.85	22	20
Center has garden or green area for play (yes=1)	0.18	0.15	0.032	[0.117]	0.79	22	20
Center has good illumination (yes=1)	0.68	0.45	0.232	[0.153]	0.14	22	20
D. Caregivers							
The highest level of education is incomplete secondary (yes=1)	0.35	0.31	0.038	[0.093]	0.68	57	48
Received training in early childhood education (yes+1)	0.40	0.46	-0.057	[0.100]	0.57	55	46
Number of years worked in this center	2.14	0.85	1.286**	[0.490]	0.01	57	48
Works with contract (yes=1)	0.65	0.42	0.232**	[0.096]	0.02	57	48
E. Other indicators							
Opening hour	8.33	8.29	0.042	[0.145]	0.77	22	20
Closing hour	15.77	15.65	0.123	[0.298]	0.68	22	20
Number of hours center operates per day	7.44	7.36	0.081	[0.360]	0.82	22	20
Proportion of staff who are women	0.93	0.88	0.045	[0.058]	0.45	22	20
Proportion of staff who receive feedback	0.27	0.08	0.183*	[0.096]	0.06	22	20
Proportion of staff who report being satisfied with their job	0.45	0.39	0.061	[0.126]	0.63	22	20
Caregiver Index	0.45	0.28	0.170**	[0.083]	0.05	22	20

Notes: The table shows the results of the balance tests in the sample of treated centers observed at the baseline. Columns (1) and (2) show means in the sample of centers that received infrastructure component (Infrastructure=1) and centers that did not receive the infrastructure component (Infrastructure =0). The estimated difference between these two groups and the associated standard error is in columns (3) and (4), respectively. Column (5) shows the P-value from the formal test of the difference in column (3) being equal to zero. Columns (6) and (7) show sample size of the group that received the infrastructure component and the group that did not receive it, respectively. See Appendix 2.7.D for details on construction of indices. Statistical significance: *10%, **5%, ***1%.

2.7.C Estimations controlling for baseline characteristics

Table 2.7.C1: Reduced form results for ITERS-R, average and by dimension controlling for baseline characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Average	Space and Furnishings	Personal Care Routines	Listening and Talking	Activities	Interaction	Program Structure	Parents and Staff
Panel A: Program Effect								
ITT=1	1.153***	1.524***	0.737***	1.376**	0.990***	2.086***	1.525***	0.645***
	[0.240]	[0.235]	[0.213]	[0.555]	[0.193]	[0.689]	[0.367]	[0.224]
Panel B: Differential								
Soft (ITT=1)	0.821**	0.716**	0.375	1.140	0.572**	2.010*	1.320**	0.598*
	[0.371]	[0.314]	[0.239]	[0.757]	[0.266]	[1.012]	[0.631]	[0.326]
Infrastructure (ITT*Infrastructure)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Panel C: Summary Statistics								
Control mean	1.406	1.609	1.110	1.667	1.197	1.924	1.275	1.416
Control SD	(0.310)	(0.333)	(0.213)	(0.865)	(0.310)	(1.193)	(0.398)	(0.313)

N	65	65	65	65	65	65	65	65

Notes: This table presents estimates (Panels A-B) and statistics (Panel C) for ITERS-R average score and dimensions. Each column corresponds to a separate regression. In Panel A outcomes are regressed on the assignment to treatment indicator controlling for a vector of controls. In Panel B outcomes are regressed on the assignment to treatment, the vector of controls and the interaction of the assignment to treatment and the indicator of infrastructure component of the program. The vector of controls includes variables on which the baseline sample is unbalanced: the number of classrooms in the center, the indicator that the center has electricity, the indicator that the center is managed by the Departmental Government, the indicator that the center receives funding from parents; the indicator that the center receives funding from NGOs, opening time, closing time, total hours of operation; proportion of women staff; proportion of staff who receive feedback; proportion of staff who report being satisfied with their work; Caregiver index. All regressions include municipality fixed effects. Robust standard errors are in brackets, standard deviations are in parenthesis. Statistical significance: *10%, **5%, ***1%.

Table 2.7.C2: Reduced form results for CIS interaction styles controlling for baseline characteristics

	(1) Sensitivity	(2) Harshness	(3) Detachment	(4) Permissiveness
Panel A: Program Effect				
ITT=1	0.621***	-0.097	-0.587***	-0.024
	[0.197]	[0.103]	[0.171]	[0.143]
Panel B: Differential				
ITT=1	0.462*	-0.088	-0.489**	-0.081
	[0.269]	[0.116]	[0.222]	[0.173]
Infrastructure (ITT*Infrastructure)	0.290	-0.017	-0.179	0.104
	[0.280]	[0.114]	[0.203]	[0.151]
Panel C: Summary Statistics				
Control mean	2.104	1.536	1.913	2.246
Control SD	(0.610)	(0.400)	(0.725)	(0.515)
N	65	65	65	65

Notes: This table presents estimates (Panels A-B) and statistics (Panel C) for CIS dimensions. Each column corresponds to a separate regression. In Panel A outcomes are regressed on the assignment to treatment indicator controlling for a vector of controls. In Panel B outcomes are regressed

on the assignment to treatment, the vector of controls and the interaction of the assignment to treatment and the indicator of infrastructure component of the program. The vector of controls includes variables on which the baseline sample is unbalanced: the number of classrooms in the center, the indicator that the center has electricity, the indicator that the center is managed by the Departmental Government, the indicator that the center receives funding from parents; the indicator that the center receives funding from NGOs, opening time, closing time, total hours of operation; proportion of women staff; proportion of staff who receive feedback; proportion of staff who report being satisfied with their work; Caregiver index. All regressions include municipality fixed effects. Robust standard errors are in brackets, standard deviations are in parenthesis. Statistical significance: *10%, **5%, ***1%.

Table 2.7.C3: Reduced form results for KIDI score and Caregiver indicators

	(1) KIDI	(2) Is satisfied=1	(3) Received training=1	(4) Received feedback=1	(5) Caregiver Index
Panel A: Program Effect					
ITT=1	0.076***	0.145*	0.128	0.210**	0.138**
	[0.027]	[0.084]	[0.105]	[0.081]	[0.057]
Panel B: Differential					
Soft (ITT=1)	0.095**	0.148	-0.132	0.227**	0.075
	[0.038]	[0.097]	[0.134]	[0.109]	[0.073]
Infrastructure (ITT* Infrastructure)	-0.036	-0.006	0.433***	-0.029	0.115
	[0.043]	[0.108]	[0.145]	[0.139]	[0.070]
Panel C: Summary Statistics					
Control mean	0.483	0.816	0.652	0.776	0.700
Control SD	(0.075)	(0.391)	(0.482)	(0.422)	(0.242)
N	50	105	101	105	65

Notes: This table presents estimates (Panels A-B) and statistics (Panel C) for KIDI score and Caregiver indicators. Each column corresponds to a separate regression. In Panel A outcomes are regressed on the assignment to treatment indicator controlling for a vector of controls. In Panel B outcomes are regressed on the assignment to treatment, the vector of controls and the interaction of the assignment to treatment and the indicator of infrastructure component of the program. The vector of controls includes variables on which the baseline sample is unbalanced: the number of classrooms in the center, the indicator that the center has electricity, the indicator that the center is managed by the Departmental Government, the indicator that the center receives funding from parents; the indicator that the center receives funding from NGOs, opening time, closing time, total hours of operation; proportion of women staff; proportion of staff who receive feedback; proportion of staff who report being satisfied with their work; Caregiver index. All regressions include municipality fixed effects. Robust standard errors are in brackets, standard deviations are in parenthesis. Statistical significance: *10%, **5%, ***1%.

Table 2.7.C4: Reduced form results for center infrastructure, service and administration indicators controlling for baseline characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Building in good conditions=1	Building needs repair=1	Construction Materials Quality Index	Electricity=1	Garden=1	Good illumination=1	Learning Materials Index	Childcare Development Monitoring Index	Curriculum, Training and Monitoring Index
Panel A: Program Effect									
ITT=1	0.281**	-0.287***	0.116	-0.139	0.258**	0.090	0.310***	0.241***	0.159**
	[0.128]	[0.085]	[0.095]	[0.135]	[0.113]	[0.142]	[0.067]	[0.063]	[0.072]
Panel B: Differential									
Soft (ITT=1)	-0.106	-0.056	0.048	-0.319*	-0.105	-0.253	0.204*	0.152*	0.088
	[0.116]	[0.080]	[0.117]	[0.174]	[0.088]	[0.157]	[0.101]	[0.084]	[0.092]
Infrastructure (ITT* Infrastructure)	0.709***	-0.423***	0.124	0.330*	0.664***	0.627***	0.194**	0.164	0.130
	[0.129]	[0.143]	[0.093]	[0.167]	[0.122]	[0.145]	[0.095]	[0.108]	[0.097]

Panel C: Summary Statistics									
Control mean	0.174	1.000	0.783	0.870	0.130	0.652	0.443	0.466	0.435
Control SD	(0.388)	(0.000)	(0.422)	(0.344)	(0.344)	(0.487)	(0.276)	(0.244)	(0.300)
N	65	65	65	65	65	65	65	65	65

Notes: This table presents estimates (Panels A-B) and statistics (Panel C) for center indicators. Each column corresponds to a separate regression. In Panel A outcomes are regressed on the assignment to treatment indicator controlling for a vector of controls. In Panel B outcomes are regressed on the assignment to treatment, the vector of controls and the interaction of the assignment to treatment and the indicator of infrastructure component of the program. The vector of controls includes variables on which the baseline sample is unbalanced: the number of classrooms in the center, the indicator that the center has electricity, the indicator that the center is managed by the Departmental Government, the indicator that the center receives funding from parents; the indicator that the center receives funding from NGOs, opening time, closing time, total hours of operation; proportion of women staff; proportion of staff who receive feedback; proportion of staff who report being satisfied with their work; Caregiver index. All regressions include municipality fixed effects. Robust standard errors are in brackets, standard deviations are in parenthesis. Statistical significance: *10%, **5%, ***1%.

Table 2.7.C5: Reduced form results for coverage indicators controlling for baseline characteristics

	(1) Children enrolled in the classroom	(2) Children present in the classroom	(3) Number of classrooms	(4) Number of caregivers
Panel A: Program Effect				
ITT=1	-7.245	-2.148	-0.257	-0.609
	[4.885]	[2.569]	[0.266]	[0.554]
Panel B: Differential				
Soft (ITT=1)	-7.270	-3.083	-0.295	-0.660
	[4.535]	[2.742]	[0.287]	[0.529]
Infrastructure (ITT* Infrastructure)	0.045	1.710	0.070	0.093
	[2.905]	[2.420]	[0.230]	[0.306]
Panel C: Summary Statistics				
Control mean	25.478	16.043	1.696	2.130
Control SD	(23.712)	(10.615)	(1.396)	(2.735)
N	65	65	65	65

Notes: This table presents estimates (Panels A and B) and statistics (Panel C) for coverage indicators in the sample of centers observed at baseline and follow-up. Each column corresponds to a separate regression. In Panel A outcomes are regressed on the assignment to treatment indicator controlling for a vector of controls. In Panel B outcomes are regressed on the assignment to treatment, the vector of controls and the interaction of the assignment to treatment and the indicator of infrastructure component of the program. The vector of controls includes variables on which the baseline sample is unbalanced: the number of classrooms in the center, the indicator that the center has electricity, the indicator that the center is managed by the Departmental Government, the indicator that the center receives funding from parents; the indicator that the center receives funding from NGOs, opening time, closing time, total hours of operation; proportion of women staff; proportion of staff who receive feedback; proportion of staff who report being satisfied with their work; Caregiver index. All regressions include municipality fixed effects. Robust standard errors are in brackets, standard deviations are in parenthesis. Statistical significance: *10%, **5%, ***1%.

2.7.D Description of indices

1: Learning Materials Index is the proportion of affirmative answers to the following questions: The Center has at least: ten books for young children? tree or more puzzles? Toys for learning of different colors, sizes and shapes? A stereo (radio with CD or recorder) to listen to music? Material to cut, color, draw, etc.? Toys for children to play and imitate? Toys to build (blocks, cubes, Lego sets, etc.? Balls? Tricycles, wooden horses, other mountable toys? Musical instruments?

2: Childcare and Development Monitoring Index is the proportion of affirmative answers to the following questions: Does the Center have a record of daily attendance of the children? Does this center periodically record the size and weight of each child? Does this Center periodically record vaccines received by each child? Does this center periodically record general health of each child? Does this center periodically record child development? Does this Center provide information to parents/caregivers about their child development? Does this Center daily inform parents/caregivers about how was the child's day?

3: Curriculum, Staff Training and Monitoring Index is the proportion of affirmative answers to the following questions: Does this Center have an annual staff training plan? Does this Center have a plan of activities for each room or group? Does this Center plan activities for each child according to child's needs? Does this Center have a pedagogical curriculum? Does this Center have regular evaluations of staff performance? Does this Center have any rules or regulations on what to do in case a child has an accident or a medical emergency?

4: Caregiver Index is the average of the proportions of caregivers in the center who (i) report being satisfied with their work; (ii) work with the contract; (iii) receive feedback on their work; (iv) had child development training in the last 3 years.

5: Construction material index takes value one if construction materials of roof, floor and walls are not precarious: roof is made of resistant material which is not wood, straw, mud; floor is not bare earth or loose bricks, walls are plastered. Construction of this index is based on the precarious toilet index in Bancalari et al. (2016).

Chapter 3: The Effects of a Home-visiting Program on Early Childhood Development in Bolivia

3.1 Introduction[35]

Inadequate parental care and home environment during the early years can result in development deficits, perpetuating and increasing with age (Baker et al., 2019; Fort et al., 2020). Children born in vulnerable families with parents who lack knowledge of adequate parenting practices are particularly exposed to these development risks. Literature shows that home-visiting programs could be a viable solution to improve parental practices for better child development and outcomes in adulthood (Attanasio et al., 2021; Gertler et al., 2014; Gertler et al., 2021). However, there is still a question on how to maintain the effectiveness of these programs when rolled out at scale in developing countries, particularly in the contexts of high socio-economic vulnerability.

This study presents the results of a large scale randomized controlled trial of a home-visiting intervention targeting low-income disadvantaged families in Bolivia through the "Grow Well to Live Well" (GWLW) program, implemented by the country's Ministry of Health. Prior to the intervention, the development scores of target children on different early childhood development dimensions (including gross and fine motricity, problem resolution, socio-individual development and communication) were very low.[36] Given the consensus in the literature that early child development inequalities are in large part determined by inequality in family and home environment,[37] the GWLW program provided a promising solution for mitigating these inequalities and preventing their perpetuation in adulthood.

[35] This chapter is based on the impact evaluation of the "Grow Well to Live Well" early childhood intervention in Bolivia funded by the Inter-American Development Bank (IDB). The impact evaluation report is published as an IDB Technical Note 1790. This research was conducted at the IDB under the direction of Julia Johannsen and Sebastian Martinez, Cecilia Vidal supported the tasks of data collection and acquisition.

[36] Gertner et al. (2016).

[37] In Bolivian context, Celhay et al. (2018), using the information from the Health and Nutrition Survey 2012, show that children aged 6 - 36 months from the poorest 20% families have a motor development score 0.21 standard deviations lower than the score of children from the richest 20%.

The operational arm of the GWLW program evaluated in this study consisted of home visits by trained community workers of families with children aged 6 to 36 months. The community workers offered children's parents guidance and counseling on parenting and early simulation, including manufacturing of home-made toys, teaching of how to talk, play and interact with children.

The program was implemented in 36 municipalities of two prioritized departments in Bolivia: Chuquisaca and Potosí. The program was implemented at the intervention unit (IU) level: communities in the rural areas and neighborhoods in the urban areas. Of the pool of eligible IUs, 315 were assigned to treatment group and 364 to control group. Within IUs assigned to the treatment group, households with at least one child aged 6 to 30 months were identified as eligible to receive the intervention. The program targeted both rural and urban areas.

This study contributes to the literature by providing evidence from an experimental evaluation of a home-visiting program implemented at large scale in a vulnerable setting of a developing Latin American country. The variation stemming from the randomized assignment of IUs in treatment and control groups allows rigorous estimation of the program treatment effects and identification of the impacts of the program on child development. This study contributes to the body of knowledge on the determinants of early childhood development, specifically, to the strand of the literature focused on the quality of parent-child interaction and home environment (see, e.g., Francesconi and Heckman, 2016; Neidell, 2000; Del Bono et al., 2016; Fiorini and Keane, 2014 for studies in high-income countries). In addition, this study also contributes to the body of literature evaluating the adaptation of the Reach Up curriculum in developing countries. In Latin America, the evidence includes the study by Attanasio et al. (2014) in Colombia and Araujo et al. (2016) in Perú. Both studies evaluated the implementation of the Reach Up curriculum through large nationwide programs, while the present study is an efficacy trial implemented in targeted disadvantaged communities in Bolivia. The evaluation also provides the measures of child development and home environment by multiple instruments, including the Ages and Stages Questionnaires® (ASQ-3) adapted to the local context of Bolivia, the socioemotional scale of the Regional Program of Child Development Indicators (PRIDI, for its acronym in Spanish), home environment indicators

from the Home Observation Measurement of the Environment (HOME), and selected items from the UNICEF Family Care Indicators (FCI). Finally, the study also contributes to the discussion on the implementation challenges of development programs in low-income countries (e.g., Bouguen et al., 2018) by showing that particularly low take-up rates can hinder the realization of the expected program impacts

The main findings show that the program had large impact on child development in rural areas. The reduced form estimation shows that the program improved the main child development indicator by 0.14 SD, representing an increase of about 80% with respect to the control mean. Across development dimensions, the program improved children's communication, fine motor, and problem-solving skills. The results for the intermediate outcomes, in line with the program's theory of change, corroborate that the impacts on child development were achieved through improvements in parent-child interactions and better cognitive stimulation. The program results were limited by imperfect compliance, which was particularly low in the urban areas. Arguably, this was the main factor contributing to no observed program effects on children in urban households. Overall, the results show that, after a year of exposure to the program and a total of 40 home visits on average, beneficiary children in rural areas had significant improvements in different dimensions of child development, which gives a solid ground for a possible scalability of the program in other regions of the country.

The reminder of this chapter is structured as follows. Section 3.2 discusses the relevant literature. Next, Section 3.3 presents the intervention and Section 3.4 describes the data and statistical analysis. Finally, Section 3.5 reports the results and Section 3.6 concludes. Some additional results and information are presented in Section 3.7 that acts as an Appendix.

3.2 Literature review

3.2.1 The determinants of early childhood development

A growing body of evidence indicates that foundations for healthy and productive lives are formed at a very early age (Grantham-McGregor and

Smith, 2016). Inadequate health and nutrition, parenting practices with limited interactions between parents and children, home environments with few books, toys and lack of other learning opportunities, can negatively affect cognitive and socioemotional development of children. Early developmental deficits can have lifelong consequences, including lower levels of school attendance and performance, lower income in adulthood, greater dependence on the health care system and higher crime rates (Naudeau et al., 2011; Walker et al., 2011). The environment and socioeconomic status of the household affect the opportunities for physical and mental development from birth, putting at a disadvantage children born in vulnerable households without access to adequate child development services (Lozoff et al., 2006; Rubio-Codina et al., 2015; Schady et al., 2015).

3.2.2 Parental and caregiver time and childhood outcomes

The existing literature highlights the importance of parenting as one of the major determinants of child development in the early years (Francesconi and Heckman, 2016). Several studies find that parental time spent on activities relevant for development matters for better child outcomes. Using National Longitudinal Survey of Youth Child-Mother file, Neidell (2000) showed that uninterrupted parental time investments in the first year of life had lasting effects on socio-emotional development of children. Price (2010) found that additional mother-child reading increased children's reading performance. Del Bono et al. (2016) used the UK Millennium Cohort data to show that maternal time is an important quantitative determinant of early childhood skill formation and has long-term effects on cognitive skill development. Fiorini and Keane (2014) analyzed the diaries data from the Longitudinal Study of Australian Children and found that the time children spend with parents on educational activities is the most productive input for cognitive skills development. For disadvantaged families, where parents often lack knowledge about children's capacities and parenting practices (Cunha et al., 2013), home-visiting interventions focused on improving parenting practices and fostering farther-child interactions can be a viable solution to achieve better child outcomes.

3.2.3 The Reach Up Early Childhood Parenting program

Home-visiting interventions aim to improve development of children by strengthening parental capacities and guiding families on parent-child interaction, stimulation activities or nutrition. Examples of such interventions in Latin America and the Caribbean include the Kallpa Wawa program in Bolivia, the Cuna Mas program in Peru, the Roving Caregivers program in the Caribbean and the Reach Up[38] Early Childhood Parenting program in Jamaica. The Reach Up is an early childhood stimulation program based on the successful Jamaican Home Visit intervention (Grantham-McGregor and Walker, 2015). The Reach Up model has been broadly and rigorously evaluated, showing benefits on child development and parenting practices (Grantham-MacGregor and Smith, 2016). Specifically, the program generated positive effects on children's cognitive development, school performance, grade progression and general knowledge (Grantham-McGregor et al., 1991; Walker et al., 2000). There is also evidence that the positive effects of the program remain in the long term, with improvements in adult education, mental health, income, and reductions in violent behavior (Gertler et al., 2021; Gertler et al., 2014; Walker et al., 2011; Walker et al., 2010).

3.2.4 Trials and adaptations of the Reach Up curriculum

Structured curricula similar to Reach Up were adapted, implemented and evaluated in other countries. In Bangladesh, several studies on the effectiveness of Reach Up program found positive effects on child development and behavior (Hamadani et al., 2006; Nahar et al., 2012). In Colombia, the Reach Up model was implemented in the framework of the Colombia Conditional Cash Transfer program known as "Families in Action" (Familias en Acción), finding positive effects on cognitive development and language (Attanasio et al., 2014). In Peru, the Reach Up model was adapted and implemented by the government at the national level through the nationwide program of home visits "Cuna Mas". The evaluation of "Cuna Mas" found positive effects on cognitive development and language (Araujo et al., 2016).

[38] More information on the Reach Up Early Childhood Parenting program can be found on www.reachupandlearn.com (accessed on December 23, 2021).

3.3 Intervention

As presented in Chapter two Section 2.3, the GWLW program included three operational arms. The second operational arm, which consisted in home visits, is the focus of this study.[39] The home visits were delivered by trained community workers organized in teams of one to three people covering beneficiary families in three to five communities. Each community worker had a workload of 15-20 families. The teams of community workers were monitored by coaches-leaders who had at least a college-level degree in early childhood development. According to the program guidelines, each beneficiary family would receive four 45-minutes home visits per month. The program targeted only the youngest child in the beneficiary household.

During home visits, trained community workers worked with the caregivers[40] offering guidance and counseling on parenting and early stimulation. The curricular content of the visits included manufacturing and use of home-made toys, teaching caregivers in a playful way how to talk, play, teach and interact with their child. It was expected that these activities would improve caregivers' knowledge about childcare and stimulation practices, increase the use of development-relevant practices in caregiver-child interactions and, put into practice, would lead to better child development outcomes.

Because the resources of the program were limited, the intervention prioritized municipalities with higher level of poverty and health needs, and with eligible health networks for other GWLW program operational arms. The assignment to intervention was done at the level of IU: community in rural areas and neighborhood in urban areas. The eligible IUs were selected according to two criteria: (i) a minimum of 200 inhabitants, (ii) absence of childcare centers.

The program was implemented as a Randomized Controlled Trial. The urban/rural areas of the eligible municipalities delimited the randomization

[39] The GWLW program consisted of three operational arms: (i) child stimulation centers, (ii) a home-visiting program, (iii) a program to improve the quality of center-based childcare. See Chapter two Section 2.3 for details.
[40] In this study we use the term "caregivers" to refer to parents of a child or other main caregivers within the household.

stata. Within each randomization stratum, the IUs were assigned to receive or not to receive the intervention. In total, 820 IUs (441 rural communities and 379 urban neighborhoods) were eligible to participate. Within each stratum, eligible IUs were randomly assigned the order of entrance to the program and listed accordingly. Eligible IUs from the top of the list were assigned to receive the intervention (treatment group) while the rest of the IUs were assigned to the control group. Before program implementation, the list of eligible IUs was refined and validated.[41] The final list consisted of 315 IUs in the treatment group and 364 IUs in the control group.

Within each IU, the program identified eligible households with at least one child aged 6 to 30 months.[42] To ensure maximum exposure to the program, the priority was given to households with smaller children. The rest of the households were assigned to the replacement group. The replacement households were assigned to receive the program if the prioritized beneficiary household refused the intervention, emigrated, or the beneficiary child exited the program because of death or reaching the age limit of 36 months.[43] The endline survey revealed that not all households assigned to the treatment group received the intervention. The participation rate was particularly low in the urban areas. The program take-up and its implications are discussed in Section 3.5.2.

The randomization was done in 2012, but the intervention was implemented between June 1, 2017, and May 30, 2018. In the first years of the program, the activities were focused on planning the program implementation, gathering baseline information, training the personnel, developing guidelines and manuals. The group of national and international professionals developed the program Guidelines "Early Childhood Development Guide with a Community Approach for Girls and Boys from 6 to 36 Months of Age". The Guidelines were based on the Reach Up curriculum and adapted to the

[41] For example, initially eligible IUs in municipalities Las Carreras (Chuquisaca) and San Antonio de Esmoruco (Potosí) were excluded because their population was smaller than 200 inhabitants.

[42] The restriction of the maximum age 30 months was introduced to ensure that the beneficiary children were exposed to the program for at least 6 months before they reached the maximum program eligibility age of 36 months.

[43] The detailed description of the selection of program beneficiaries is available in Johannsen et al. (2019b).

sociocultural context of Bolivia. The Guidelines contained the detailed intervention protocol, specifying activities and parenting practices that community workers taught parents and caregivers during each home visit. Prior to the intervention, the Guidelines were tested and validated in the groups of non-beneficiary children in the departments of Chuquisaca and Potosi.

3.4 Data and method

3.4.1 The endline survey

The data for the program evaluation was retrieved from the endline survey (ES).[44] The ES was carried out between June and July 2018, one year after the start of the intervention and immediately after the intervention concluded. The ES included modules on childcare practices, direct observations of child-caregiver interactions, and parental time inputs. To evaluate development outcomes, the survey included measures of communication, gross motor, fine motor, cognitive and socioemotional development (described in Sections 3.4.2 and 3.4.3). In addition, the ES collected background information of socioeconomic characteristics of all household members, retrospective information on child endowments (such as birth weight and antenatal care) and participation in early childhood programs.

The survey had a probabilistic sample design with IUs being the primary sampling units. The sample frame was the list of all treatment and control IUs. In rural areas, the selection of treatment IUs was probabilistic, with the distribution across municipalities being proportional to the distribution of all IUs in the sample frame. In each municipality, the equivalent number of control IUs was selected from the bottom of the treatment assignment list. In urban areas, a random sample of IUs in treatment and control groups were selected. The final survey sample consisted of 100 urban (50 treated and 50

[44] The program also implemented a baseline survey in 2014 on a sample of children in eligible communities. The objective of the baseline survey was to provide the information on the target population and to inform the program design (Gertner et al., 2016). The information from the baseline survey was not used in this study because the children interviewed at baseline in 2014 were no longer eligible to receive the program when it started in 2017.

control) and 140 rural (70 treated and 70 control) IUs. In each rural IU, a random sample of 12 households was selected, while in urban IUs the ES was implemented in all eligible households. The household was eligible for the interview if it had at least one child aged 12 to 44 months.[45] In households with more than one child aged 12 to 44 months, the survey collected information from the youngest child. The final analysis sample included 1,052 and 1,461 children in urban and rural areas, respectively.

The survey was carried out by a specialized data collection firm with monitoring by the external independent data quality assurance consultant. The survey protocol obtained the Institutional Review Board authorization from the National Committee of Bioethics.

3.4.2 Child development outcomes

Child development was measured using five dimensions of the Ages and Stages Questionnaires (ASQ) and seven questions of the Regional Project of Child Development Indicators ("Proyecto Regional de Indicadores de Desarrollo Infantil", or PRIDI in Spanish) questionnaire.

ASQ questionnaires (Squires et al., 2009) constitute a developmental screening tool used to identify potential developmental delays in the first five years of life. It is largely used in household surveys due to easy parental comprehension, reliability and implementational cost-effectiveness (Kerstjens et. al., 2009, and studies cited there). In this study, child development was assessed using the communication, fine motor, gross motor, problem solving and socio-individual scales of the third edition of the ASQ (ASQ-3). The questionnaires were applied through 13 age-group questionnaires.[46] To increase variability in the sample, additional items of

[45] The age range of eligible children was defined to maximize the probability of interviewing children that were exposed to the program. Since the program started in June 2017 targeting children aged 6 - 30 months, by the time of the endline survey, the age of these children would be between 18 and 42 months (i.e., enrollment age plus 12 months). The age range established for the endline survey was 12 – 44 months to include children incorporated in the program as replacements.

[46] Age groups: 1) 11m 0d – 12m 30d; 2) 13m 0d – 14m 30d; 3) 15m 0d – 16m 30d; 4) 17m 0 d – 18m 30d; 5) 19m 0d – 20m 30d; 6) 21m 0d – 22m 30d; 7) 23m 0d – 25m 30d; 8) 26m 0d – 27m 30d; 9) 28m 0d – 30m 30d; 10) 31m 0d – 33m 30d; 11) 34m 0d – 36m 30d; 12) 37m 0d – 42m 30d; 13) 43m 0d – 48m 30d.

decreasing and increasing difficulty were added to the six core items in each developmental domain. Similar adaptations were implemented in other studies (Rubio-Codina et al., 2016; Fernald et al., 2012). The language was adapted to the local context of Bolivia. Questions about tasks that the child is (or is not) able to perform were asked to a caregiver by an interviewer, while some specific items were directly administered to a child. Each question/item was scored 10, 5 or 0 depending on whether the caregiver reported that the child could perform the task always, sometimes, or never, respectively. Raw scores were constructed for each domain as the sum of the scores across items. The raw scores were transformed into within age-group standardized z-scores, with mean zero and standard deviation (SD) one.

In addition to the five dimensions covered in the ASQ-3, the ES included seven questions from the socioemotional scale of PRIDI questionnaire.[47] PRIDI questionnaire aims to identify young children at risk of social or emotional difficulties and measures the child's skills to recognize emotions and to handle and adapt to new situations. Based on the response category for each question ("yes", "sometimes" and "no"), a raw score was constructed by adding all responses. An index of socio-emotional child development was computed as a simple average of the seven raw scores (Kling et al., 2007).[48]

3.4.3 Home environment outcomes

The home environment quality was measured using the abbreviated version of "Responsivity" and "Acceptance" dimensions of the HOME inventory (Caldwell and Bradley, 2001) and selected items from the UNICEF's FCI.

HOME is a well-known observational measure of the quality of cognitive stimulation and emotional support provided in a child's family. Several versions of HOME inventory are available: Infant/Toddler (IT) HOME for children birth to 3, Early Childhood (EC) HOME for children ages 3 to 6, and Middle Childhood (MC) HOME for children aged 6 to 10. For scoring, the assessor enters a plus sign for each item if the behavior is observed or reported

[47] For more information about the program and questionnaires visit https://www.iadb.org/es/sectores/educacion/pridi/inicio (accessed on December 23, 2021).
[48] A detailed information on construction of this and other indices is presented in Appendix 3.7.A.

and a minus sign if it is not. Each subscale score and the total inventory score are computed by counting the number of plus signs. In this study, the abbreviated versions of the IT HOME "Responsivity" and "Acceptances" subscales were employed.[49] "Responsivity" subscale measures the degree of responsiveness of caregivers to a child. The implemented shortened version of this subscale comprises six items. The plus signs were scored as one and minus signs were scored as zero. "Acceptance" subscale measures parental acceptance of misbehavior and avoidance of restriction and punishment. The implemented shortened version of this subscale has five items. The scoring of this subscale is reversed. Therefore, the plus signs were scored zero and minus signs were scored one. Standardized z-scores were computed for each subscale and the overall HOME score using the sample mean and SD.

The FCI is a short and easy to administer test developed by UNICEF (Frongillo et al., 2003). The test was validated against HOME in vulnerable socioeconomic contexts by Hamadani et al. (2010). Selected items of the FCI inventory were used to collect the information on parental support and stimulation of child development. Caregivers were asked about six activities in which an adult in the household was engaged with the child over the past three days: singing, reading, telling stories, counting or naming things, playing and going out. The information from these six questions was used to construct the Child Stimulation Index, which shows a proportion of positive answers to the questions about adult engagement in each of six play activities (Kling et al., 2007). In addition, the FCI collected the information on practices of discipline and punishment. This information was used to construct the severe discipline practices and the rules and routines indices.[50]

3.4.4 Method

[49] IT-HOME full version has 45 items and six subscales: (1) Responsivity: the extent of responsiveness of the parent to the child; (2) Acceptance: parental acceptance of suboptimal behavior and avoidance of restriction and punishment; (3) Organization: regularity and predictability of the environment; (4) Learning Materials: provision of appropriate play and learning materials; (5) Involvement: extent of parental involvement; and (6) Variety in daily stimulation. Eighteen items are based on observation, 15 on interview, and 12 on either observation or interview.

[50] See Appendix 3.7.A for the details of indices construction.

This study uses the randomization design of the program as the key feature that allows to identify the impacts. Given the experimental design of the intervention in which the IUs were assigned to treatment based on the random lottery order, the identification strategy is based on comparison of the results between households in the treatment IUs and control IUs. In the absence of selective attrition and with perfect compliance, the difference between treatment and control groups is an unbiased estimate of the average treatment effect (ATE). In case of this intervention, not all households complied with the assigned intervention group, which implies that the ATE cannot be directly recovered. Instead, the empirical strategy focuses on estimation of the associated program effect parameters: the intent-to-treatment (ITT) effect and the Local Average Treatment Effect (LATE). The ITT effect is an estimate of the effect on those assigned to treatment, regardless of their take-up (Angrist and Pischke, 2008). The LATE provides an estimate of the treatment effect for compliers, i.e., those who are induced by their assignment to comply (Imbens and Angrist, 1994; Angrist et al., 1996). Under imperfect compliance, the ITT usually provides a lower bound of the ATE, while the LATE typically provides an upper bound.[51]

The results presented in this study are estimated in the following econometric models:

Model 1: The ITT effect, or reduced form effect, for households in urban and rural areas. The ITT parameter is retrieved by regressing the outcome indicator on the assignment to treatment indicator and the randomization strata fixed effect:

$$Y_{ij} = \alpha + \beta T_{ij} + \gamma_j + \varepsilon_{ij} \quad (1)$$

Where Y_{ij} is the outcome indicator for child i in strata j; T_{ij} is a treatment assignment indicator equal to one if the IU was assigned to treatment and zero otherwise; γ_j is strata fixed effect, and ε_{ijt} is the error term. The parameter of interest is β, which measures the ITT effect on the outcome indicator between treatment and control group.

[51] J-PAL Abdul Latif Jameel Poverty Action Lab Research Resources. Available online: https://www.povertyactionlab.org/resource/data-analysis, (accessed on December 21, 2021).

Model 2: The ITT effect, estimated by regressing the outcome indicator on the assignment to treatment indicator, the randomization strata fixed effect, and the set of control variables included to improve efficiency and account for some imbalances in fixed or pre-determined characteristics:[52]

$$Y_{ij} = \alpha + \beta T_{ij} + X'_{ij}\delta + \gamma_j + \varepsilon_{ij} \quad (2)$$

In this model, X_{ij} is a set of child, caregiver and household controls. All other variables and parameteres are defined as in model (1).

Model 3: The ITT effect estimated by regressing the outcome indicator on the assignment to treatment indicator, the randomization strata fixed effect and the set of control variables, differentiating between rural and urban households:

$$Y_{ij} = \alpha + \beta_R T_{ij} * Rural + \beta_U T_{ij} * Urban + X'_{ij}\delta + \gamma_j + \varepsilon_{ij} \quad (3)$$

In this model, the parameters of interest are β_R and β_U. They measure the ITT effect on the outcome indicator between treatment and control group in rural and urban households, respectively. These parameters are estimated on the interaction of the assignment to treatment indicator T_{ij} and the indicator of rural/urban IU. All other variables and parameters included in the model are defined as in model (2).

Model 4: The LATE effect, estimated differentiating between rural and urban households and including the randomization strata fixed effect and the control variables. The LATE parameters are estimated in two stages. In the first stage, the interactions between the actual treatment and rural/urban indicator are

[52] The control variables include: the child's age in months, the indicator that the child is female, the indicator that the child has a health card, the indicator that the child's mother had more than four prenatal controls, the indicator that the child's caregiver can read, write, caregiver's years of education, marital status, the indicator that the caregiver is indigenous, the household size, number of rooms, home ownership, and the wealth index. Six control variables – child has a health card, child's caregiver can read, write, marital status, caregiver is indigenous, and the number of rooms – are included to control for the imbalance in pre-determined and time-invariant characteristics. See Section 3.5.1 for details. Other control variables are included to improve precision of the estimated parameters (Martinez et al., 2018).

regressed on the interactions of the assignment to treatment and the indicator of rural/urban IU, strata fixed effects, and the control variables. In the second stage, the outcome indicators are regressed on the predicted values from the first stage regressions, strata fixed effects, and control variables. Formally, the following equations are estimated:

Stage 1:

$$D_{ij} * Rural = \alpha_1 + \beta_{R1}\, T_{ij} * Rural + \beta_{U1}\, T_{ij} * Urban + X'_{ij}\delta_1 + \gamma_j + \varepsilon_{ij} \quad (4.1)$$

$$D_{ij} * Urban = \alpha_2 + \beta_{R2}\, T_{ij} * Rural + \beta_{U2}\, T_{ij} * Urban + X'_{ij}\delta_2 + \gamma_j + \varepsilon_{ij} \quad (4.2)$$

Stage 2:

$$Y_{ij} = \alpha + \beta_{R\_LATE}\, D_{ij} * \widehat{Rural} + \beta_{U\_LATE}\, D_{ij} * \widehat{Urban} + X'_{ij}\delta + \gamma_j + \varepsilon_{ij} \quad (4.3)$$

The parameters of interest are β_{R_LATE} and β_{U_LATE}. They measure the LATE effect on the outcome indicator for the households that complied with the treatment in rural and urban areas, respectively. These parameters are estimated on the predicted values of $D_{ij} * \widehat{Rural}$ and $D_{ij} * \widehat{Urban}$, retrieved from the first-stage regressions (equations 4.1 and 4.2), where D_{ij} is an indicator of whether the household received or not received the intervention. All other variables and parameters are defined as in models (1) – (3).

In all regressions the standard errors are robust and clustered at the IU level to account for correlation within randomization units. Given that the parameters are estimated in a sample with probabilistic design, all regressions include sampling weights.[53]

3.5 Results

3.5.1 Balance tests

[53] In addition to models (1) – (4), the estimation of LATE effects was done for the pooled sample, and the estimation of the ITT without covariates differentiating between rural and urban households. Results of these analyses can be made available upon request.

To verify the experimental design validity, we perform balance tests using the information from the endline survey on pre-determined and time-invariant characteristics of children, their households, and caregivers in the sample of rural and urban households. These results are reported in Table 3.1.

As seen in Table 3.1, we find balance in most time-invariant and pre-determined characteristics of children and their households. Notably, we observe balance in key child characteristics, such as gender, age, weight at birth. We also observe balance in main household characteristics which might be potentially correlated with the outcomes, such as household income, wealth index, land ownership. Most of imbalance is observed in the caregiver characteristics. Specifically, we find statistically significant differences in means at five percent level of statistical significance for the indicators that the caregiver is indigenous and that she can read, and statistically significant differences at ten percent level for the indicators that the caregiver can write and is married. We do not observe any statistically significant differences in means for the household characteristics, except for the number of bedrooms, which is statistically significant at ten percent level of statistical significance. As it was mentioned in the previous section, in models (2) – (4) we control for the imbalance in pre-determined and time-invariant characteristics by including as regressors the variables on which we find statistically significant differences between control and treatment group means, along with some additional regressors which we included to improve precision of the estimates.

Table 3.1: Comparison of treatment and control group characteristics

	Rural					Urban				
	Treatment mean	Control mean	Difference	P-value	Sample	Treatment mean	Control mean	Difference	P-value	Sample
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A: Child characteristics and endowments										
Child is female	0.45	0.49	-0.039	0.132	1461	0.52	0.48	0.031	0.313	1052
Child's age in months	28.80	28.28	0.568	0.271	1461	28.20	28.49	-0.230	0.741	1052
Child has a health card	0.10	0.10	-0.007	0.728	1333	0.16	0.21	-0.055*	0.059	1008
Child's mother had four or more prenatal controls	0.70	0.75	-0.023	0.367	1461	0.85	0.90	-0.055	0.105	1045
Week of pregnancy at 1st antenatal care check-up	8.98	8.82	-0.137	0.697	1433	9.10	9.15	-0.114	0.802	1045
Birth attended in health facility	0.60	0.64	-0.026	0.439	1461	0.95	0.94	0.012	0.510	1052
Child's weight at birth (grams)	3321.60	3300.50	20.078	0.511	1097	3252.15	3292.05	-39.347	0.200	963
B: Caregiver characteristics										
Caregiver is female	0.98	0.98	0.006	0.461	1461	0.97	0.97	-0.006	0.595	1052
Caregiver's age	31.78	31.44	0.348	0.534	1461	30.61	30.07	0.403	0.472	1052
Caregiver is married/in civil union	0.51	0.56	-0.045*	0.099	1461	0.42	0.37	0.048	0.122	1052
Caregiver is head of household	0.13	0.13	0.003	0.878	1461	0.17	0.13	0.037	0.154	1052
Mother tongue is Spanish	0.25	0.32	0.001	0.975	1461	0.63	0.63	-0.009	0.871	1052
Caregiver is indigenous	0.77	0.68	0.080**	0.011	1461	0.55	0.53	0.020	0.683	1052
Caregiver reads	0.78	0.82	-0.052**	0.027	1461	0.93	0.97	-0.038**	0.018	1052
Caregiver writes	0.78	0.81	-0.042*	0.052	1461	0.93	0.97	-0.042**	0.014	1052
Years of education	5.66	5.56	0.054	0.841	1461	9.86	10.50	-0.766	0.191	1052

Number of hours worked weekly	43.28	44.48	-0.138	0.921	592	43.43	40.87	2.603	0.251	345
Self-employed or independent worker	0.39	0.34	0.044	0.239	592	0.50	0.52	-0.032	0.644	346
C: Household characteristics										
Household size	5.45	5.44	-0.008	0.947	1461	4.74	4.59	0.163	0.152	1052
The dwelling is a house	0.84	0.83	-0.007	0.760	1461	0.74	0.70	0.024	0.620	1052
Household owns the dwelling	0.82	0.81	0.005	0.842	1461	0.45	0.41	0.023	0.537	1052
Construction material index[1]	0.45	0.46	0.001	0.986	1461	0.88	0.90	-0.018	0.397	1052
Bathroom connected to the sewerage	0.14	0.12	-0.013	0.737	1461	0.86	0.82	0.038	0.346	1052
Has electricity	0.79	0.82	-0.003	0.926	1461	0.99	0.99	-0.005	0.448	1052
Number of bedrooms	2.59	2.46	0.150*	0.051	1461	2.41	2.46	-0.062	0.733	1052
Dwelling is a room	0.15	0.16	0.006	0.785	1461	0.25	0.28	-0.021	0.676	1052
Wealth index[1]	-0.63	-0.55	-0.032	0.675	1461	0.79	0.83	-0.033	0.686	1052
Household owns agricultural land	0.81	0.72	0.052	0.141	1461	0.20	0.19	0.013	0.671	1052
Logarithm of monthly household income	6.54	6.79	-0.171	0.132	1418	7.88	7.92	-0.042	0.544	1040

Notes: 1: See Appendix 3.7.A for the index construction methodology. Columns (1) and (2) show means in the treatment group, column (3) shows the difference between control and treatment group means, column (4) shows the P-value from the formal test of the equality of the treatment and control group means, column (5) shows the number of observations in the overall sample (treatment and control) for rural households. Columns (6)-(10) show the same information for urban households. The differences between means are estimated in the regressions where each child, caregiver or household characteristic variable is regressed on the assignment to treatment indicator controlling for the randomization strata fixed effects. The sample of caregivers reporting the hours worked and the employment type is restricted to those caregivers who reported that they worked at least one hour last week. All statistics calculations and estimations use sampling weights. Statistical significance: * 10%, ** 5%, *** 1%.

3.5.2 Program take-up and participation

This section presents the results of the program take-up. Table 3.2 show that program take-up was no complete. In particular, 38% of the eligible households received at least one program visit in rural areas and 14% in urban areas. A possible explanation of this low take-up rate is that almost half (43%) of households in rural IUs assigned to treatment and 68% of households in urban IUs assigned to treatment reported not knowing about the program. Plausible reasons include staff shortage (i.e., field teams were understaffed at certain moments of the program and could not cover all eligible households); migration, especially for children of working-age parents; difficulty to identify IUs' limits by program staff, who may have considered different geographical community limits than those used by the survey teams.

Although the compliance with the treatment assignment was not perfect, the duration and frequency of the program visits in the households that received the intervention was in accordance with the program protocol. As shown in Table 3.2, on average, the frequency of visits was around three to four per month and the duration of the program was about 12 months. Overall, treated children received about 40 program visits.

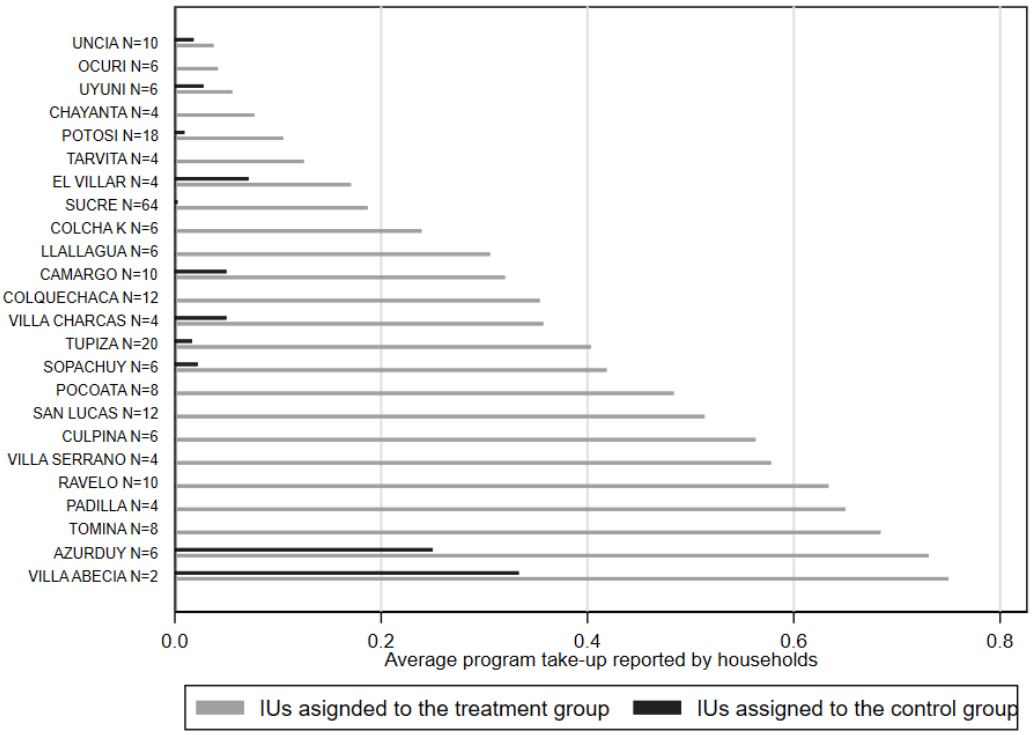
In addition to the low take-up among treated households, we also observe a small contamination of the control group: one percent of the control households reported having received the program in rural IUs and two percent in urban IUs. The analysis of the program take-up and control contamination by IUs reveals that the extent of the problem varies across municipalities. On Figure 1 we illustrate these differences: while all municipalities in which the endline survey was implemented present partial compliance, only 11 municipalities present control contamination problem, with particularly high percentage of treated households in IUs assigned to the control group in Azurduy and Villa Abecia.

Table 3.2: Program take-up and participation

	Rural					Urban				
	Treatment mean	Control mean	Difference	P-value	Sample	Treatment mean	Control mean	Difference	P-value	Sample
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Receives or has received program visits	0.38	0.01	0.370***	0.000	1461	0.14	0.02	0.114***	0.000	1052
Caregiver knows about the program	0.57	0.21	0.386***	0.000	1461	0.32	0.22	0.093**	0.026	1052
Visits per month[1]	3.46	4.01	-0.384	0.138	296	3.71	3.04	0.823	0.113	100
Months household received program visits[1]	11.93	6.72	3.221	0.251	296	11.45	14.96	-0.327	0.920	100
Total number of visits[1]	40.85	26.32	10.329	0.303	296	43.96	45.40	9.136	0.522	100

Notes: 1: Conditional on having received the program. Columns (1) and (2) show means in the treatment group, column (3) shows the difference between control and treatment group means, column (4) shows the P-value from the formal test of the equality of the treatment and control group means, column (5) shows the number of observations in the overall sample (treatment and control) for rural households. Columns (6)-(10) show the same information for urban households. The differences between means are estimated in the regressions where the program take-up or participation indicator is regressed on the assignment to treatment indicator controlling for the randomization strata fixed effects. All statistics calculations and estimations use sampling weights. Statistical significance: * 10%, ** 5%, *** 1%.

Figure 3.1: Program participation by municipalities



Notes: This figure shows the average participation rate reported by households in treated and control Intervention Units (IUs) across municipalities. "N" indicates the number of IUs in each municipality.

Given the partial compliance with the treatment assignment and the contamination of the control group, the magnitude of the minimum detectable effect size (MDES) at the design stage of the intervention is smaller than the MDES accounting for partial compliance (Duflo et al., 2007). In case of this study, the reduction in statistical power caused by imperfect compliance is specifically severe in the subsample of urban IUs, where participation rate in the treatment group is only 14% and the adjusted difference between treatment and control group compliance with assignment to treatment is 11%. The extent of the problem is large and is reflected in the estimates of the program results presented in the following section of this study.

3.5.3 Program effects on child development

Table 3.3 presents program effects on children's main development outcomes measured using standardized z-scores based the answers of ASQ-3[54] and PRIDI questionnaires. The table is structured in four columns, each column showing the results for the models presented in section 3.4.4. Specifically, the results for models one, two and three are presented in columns one through three, respectively. The second stage results for the LATE model (equation 4.3) are presented in the column four. The overall ITT effect β is reported in line "ITT=1", the ITT or LATE effects for rural and urban households, β_R, β_U, β_{R_LATE}, β_{U_LATE}, are reported in lines "ITT(LATE)*Rural" and "ITT(LATE)*Urban." The table also reports the control means in the overall sample, rural and urban areas, and the number of observations. The regressions estimated for models two through four include control variables. The table is divided in seven panels, each one showing the results for the overall ASQ-3 score, five ASQ-3 dimensions, and socio-emotional dimension based on PRIDI questionnaire index.

Table 3.3: Program effects on child development (ASQ-3 z-scores and PRIDI index)

	ITT	ITT + controls	ITT Rural/Urban + controls	LATE Rural/Urban + controls
Overall ASQ-3 (Z-score)	(1)	(2)	(3)	(4)
ITT = 1	0.048	0.067		
	[0.062]	[0.057]		
ITT(LATE)*Rural			0.140**	0.351**
			[0.063]	[0.147]
ITT(LATE)*Urban			0.013	0.122
			[0.087]	[0.782]
Control mean	0.015	0.015		
Control mean (rural)			-0.179	-0.179
Control mean (urban)			0.165	0.165
Sample	2499	2499	2499	2499
Communication ASQ-3 (Z-score)				
ITT = 1	0.038	0.043		
	[0.050]	[0.049]		
ITT(LATE)*Rural			0.170***	0.424***

[54] In Appendix 3.7.B1 we present the results for the raw ASQ-3 scores.

			[0.053]	[0.131]
ITT(LATE)*Urban			-0.050	-0.443
			[0.072]	[0.655]
Control mean	-0.008	-0.008		
Control mean (rural)			-0.134	-0.134
Control mean (urban)			0.090	0.090
Sample	2499	2499	2499	2499
Gross Motor ASQ-3 (Z-score)				
ITT = 1	-0.004	0.007		
	[0.067]	[0.065]		
ITT(LATE)*Rural			-0.028	-0.069
			[0.063]	[0.156]
ITT(LATE)*Urban			0.032	0.286
			[0.102]	[0.925]
Control mean	0.030	0.030		
Control mean (rural)			-0.068	-0.068
Control mean (urban)			0.106	0.106
Sample	2499	2499	2499	2499
Fine Motor ASQ-3 (Z-score)				
ITT = 1	0.051	0.067		
	[0.047]	[0.045]		
ITT(LATE)*Rural			0.116**	0.290**
			[0.057]	[0.142]
ITT(LATE)*Urban			0.031	0.275
			[0.066]	[0.592]
Control mean	0.022	0.022		
Control mean (rural)			-0.139	-0.139
Control mean (urban)			0.145	0.145
Sample	2499	2499	2499	2499
Problem Solving ASQ-3 (Z-score)				
ITT = 1	0.018	0.035		
	[0.055]	[0.054]		
ITT(LATE)*Rural			0.166***	0.413***
			[0.063]	[0.150]
ITT(LATE)*Urban			-0.061	-0.538
			[0.079]	[0.710]
Control mean	0.022	0.022		
Control mean (rural)			-0.188	-0.188
Control mean (urban)			0.184	0.184
Sample	2499	2499	2499	2499
Socio-individual ASQ-3 (Z-score)				
ITT = 1	0.050	0.065		
	[0.051]	[0.048]		
ITT(LATE)*Rural			0.026	0.067
			[0.057]	[0.137]
ITT(LATE)*Urban			0.094	0.838
			[0.071]	[0.690]
Control mean	-0.015	-0.015		

Control mean (rural)			-0.059	-0.059
Control mean (urban)			0.019	0.019
Sample	2499	2499	2499	2499
Socio-emotional PRIDI				
ITT = 1	0.022	0.029**		
	[0.016]	[0.014]		
ITT(LATE)*Rural			0.049***	0.122***
			[0.014]	[0.035]
ITT(LATE)*Urban			0.014	0.125
			[0.022]	[0.207]
Control mean	0.504	0.504		
Control mean (rural)			0.448	0.448
Control mean (urban)			0.548	0.548
Sample	2513	2513	2513	2513

Notes: Scores for overall, communication, gross motor, fine motor, problem solving, and socio-individual dimensions were computed using questionnaires adapted from the Ages and Stages Questionnaires, 3rd edition. (ASQ-3). Z-scores were computed with respect to in-age-groups means and standard deviations for the overall ASQ-3 and each dimension. The score for socio-emotional dimension was computed using 7 selected items of the PRIDI questionnaire. Results in columns (3) – (4) are from regressions that include controls for child's age in months, sex, has health card, >=4 antenatal care controls, caregiver can read, write, education in years, marital status, indigenous, household size, number of rooms, home ownership, wealth index, strata fixed effect, and are estimated using sampling weights. Standard errors in brackets are adjusted for clustering at the IU level. Statistical significance: * 10%, ** 5%, *** 1%.

As seen in Table 3.3, while we do not observe statistically significant program effects in the models that pool together rural and urban households (except for PRIDI's socio-emotional dimension), we do find large and statistically significant program effects at the conventional level of statistical significance (five or one percent) for rural households in the models which differentiate between rural and urban households. Specifically, our results show that the program was successful in increasing the overall ASQ-3 score in rural households by 0.14 SD in the reduced form estimation and by 0.35 SD in the LATE model. This implies a 78% increase (ITT model) with respect to the control mean. We also find a large and statistically significant program effect in three ASQ-3 dimensions: communication, fine motor, and problem solving, ranging from 0.116 SD to 0.424 SD. On the contrary, we do not observe effects in gross motor and socio-individual dimensions. A plausible explanation to this is that the program did not specifically focus on gross motor development (e.g., keeping equilibrium, running) and did not specifically target the child's satisfaction of self-help needs (e.g., getting dressed) measured in the socio-individual dimension of the ASQ-3. Finally, we observe statistically significant program effects in socio-emotional

dimension in the overall sample (0.029 points of PRIDI index) and in the sample of rural households (0.049 to 0.122 points of PRIDI index).[55]

3.5.4 Program effects on child's environment

In this section we present the program effects on the child environment, specifically, the quality of the child-caregiver interaction measured using standardized HOME z-scores,[56] and child stimulation, sever discipline practices, rules and routines indices based on the FCI questionnaire. These results are intermediate outcomes and represent the program mediating pathway towards observed changes in child development indicators. The results are presented in Table 3.4. The rows and columns of this table have the same structure as in Table 3.3 described in Section 3.5.3. The overall HOME score and the results for two child-caregiver interaction scales are presented in the first three panels. The results for the FCI child stimulation, severe discipline practices, and rules and routines indices are presented in the last three panels.

Table 3.4: Program effects on parent-child interaction and stimulation (HOME z-scores and FCI indices)

	ITT	ITT + controls	ITT Rural/Urban + controls	LATE Rural/Urban + controls
Overall HOME (Z-score)	(1)	(2)	(3)	(4)
ITT = 1	-0.047	-0.021		
	[0.075]	[0.071]		
ITT(LATE)*Rural			0.146***	0.362***
			[0.056]	[0.134]
ITT(LATE)*Urban			-0.144	-1.283
			[0.116]	[1.028]
Control mean	0.041	0.041		

[55] It is worth noting that, while ASQ-3 includes the social-individual domain, this scale mostly assesses child's self-help needs. In contrast, the socio-emotional domain of development focuses on such behavioral areas as self-regulation, compliance, social-communication, adaptive functioning, affect and interaction with people. The experts recommend complementing the ASQ-3 assessment with socio-emotional evaluations for a complete screening of development factors: https://agesandstages.com/free-resources/articles/using-asq-3-and-asqse-2-together/ (accessed on December 21, 2021).
[56] In Appendix 3.7.B2 we present the results for the raw HOME scores.

Control mean (rural)			-0.187	-0.187
Control mean (urban)			0.218	0.218
Sample	2513	2513	2513	2513
Responsiveness HOME (Z-score)				
ITT = 1	-0.034	-0.005		
	[0.079]	[0.077]		
ITT(LATE)*Rural			0.198***	0.490***
			[0.058]	[0.140]
ITT(LATE)*Urban			-0.155	-1.381
			[0.124]	[1.120]
Control mean	0.044	0.044		
Control mean (rural)			-0.201	-0.201
Control mean (urban)			0.234	0.234
Sample	2513	2513	2513	2513
Acceptance HOME (Z-score)				
ITT = 1	-0.047	-0.044		
	[0.053]	[0.051]		
ITT(LATE)*Rural			-0.084	-0.210
			[0.064]	[0.158]
ITT(LATE)*Urban			-0.014	-0.131
			[0.075]	[0.655]
Control mean	0.006	0.006		
Control mean (rural)			-0.020	-0.020
Control mean (urban)			0.026	0.026
Sample	2513	2513	2513	2513
Child stimulation index FCI				
ITT = 1	0.023	0.037**		
	[0.022]	[0.017]		
ITT(LATE)*Rural			0.059***	0.149***
			[0.017]	[0.035]
ITT(LATE)*Urban			0.020	0.178
			[0.028]	[0.259]
Control mean	0.487	0.487		
Control mean (rural)			0.393	0.393
Control mean (urban)			0.559	0.559
Sample	2513	2513	2513	2513
Severe discipline practices index FCI				
ITT = 1	0.017	0.018		
	[0.012]	[0.012]		
ITT(LATE)*Rural			0.014	0.036
			[0.009]	[0.023]
ITT(LATE)*Urban			0.021	0.184
			[0.020]	[0.181]
Control mean	0.179	0.179		
Control mean (rural)			0.185	0.185
Control mean (urban)			0.175	0.175
Sample	2513	2513	2513	2513
Rules and routines index FCI				

ITT = 1	-0.007	-0.001		
	[0.022]	[0.021]		
ITT(LATE)*Rural			-0.008	-0.019
			[0.027]	[0.064]
ITT(LATE)*Urban			0.004	0.036
			[0.030]	[0.270]
Control mean	0.441	0.441		
Control mean (rural)			0.388	0.388
Control mean (urban)			0.482	0.482
Sample	2513	2513	2513	2513

Notes: The overall, responsiveness and acceptance HOME scores were computed using a reduced set of items from the HOME inventory. The activity participation index, severe discipline practice, and rules and routines indices show the proportion of affirmative answers on the relevant questions of the FCI questionnaire. See Appendix 3.7.A for the index construction methodology. For HOME inventory, the raw scores were transformed in standardized z-scores using sample mean and standard deviation. The results in columns (2) – (4) are from regressions that include controls for child's age in months, sex, has health card, >=4 antenatal care controls, caregiver can read, write, education in years, marital status, indigenous, household size, number of rooms, home ownership, wealth index, strata fixed effect, and are estimated using sampling weights. Standard errors in brackets are adjusted for clustering at IU the level. Statistical significance: * 10%, ** 5%, *** 1%.

As seen in Table 3.4, the program had large and statistically significant effects on child-caregiver interaction quality in rural households. Specifically, the program increased the overall HOME score in rural households in a range from 0.146 SD in the ITT model to 0.362 in the LATE model. This impact is mostly driven by a large increase in the responsiveness dimension (98% with respect to the control mean in the ITT model). We do not observe program effects in the acceptance dimension in rural households, and in either dimension in urban households. The results for the stimulation index show that the program had a large and statistically significant effect (0.059 points in the ITT model and 0.149 points in the LATE mode) improving cognitive stimulation of children in rural households. We do not observe changes in the severe discipline and rules and routines indices.

Overall, the findings in this section are consistent with the program having a positive impact on child development outcomes through increase in cognitive stimulation and better quality of child-caregiver interactions. This is expected and in line with the program logic, considering that the intervention specifically focused on improving caregiver skills to interact and stimulate children. The absence of the program effects in the two of FCI indices focused on rules and punishment is also consistent with the program logic, since the

program curriculum did not address these areas.[57] Regarding the HOME's acceptance dimension, arguably, the items included in this dimension are difficult to measure in a single visit by a survey enumerator.[58]

3.5.5 Unanticipated effects on health outcomes

The program design pre-identified various dimensions of early child development as primary outcomes of interest (communication, cognitive and psychosocial skills). However, given that the program was delivered by the Bolivian Ministry of Health, home visitors were identified by families as health professionals. Motivated by these aspects of the program implementation, we also look at the effects on child health outcomes registered in the endline survey. These results are reported in Table 3.5. The table has the same structure as Tables 3.3 and 3.4.

Table 3.5: Program effects on health indicators

	ITT	ITT + controls	ITT Rural/Urban + controls	LATE Rural/Urban + controls
	(1)	(2)	(3)	(4)
Child was sick in the last 4 weeks				
ITT = 1	0.001	0.009		
	[0.029]	[0.028]		
ITT(LATE)*Rural			-0.060**	-0.148**
			[0.027]	[0.066]
ITT(LATE)*Urban			0.059	0.526
			[0.043]	[0.416]
Control mean	0.464	0.464		
Control mean (rural)			0.525	0.525
Control mean (urban)			0.417	0.417
Sample	2513	2513	2513	2513
Child had a diarrhea in the last 2 weeks				
ITT = 1	-0.016	-0.014		
	[0.022]	[0.021]		
ITT(LATE)*Rural			-0.049**	-0.122**
			[0.022]	[0.054]

[57] In fact, the estimation of the effects for these two indices can be considered falsification tests, since no changes should be observed in these indicators as a result of the program.
[58] The acceptance dimension includes items such as mother shouting, scolding or hitting a child, prohibiting or depriving them, expressing annoyance or hostility towards a child during the survey enumerator visit. Arguably, caregivers can refrain from this kind of behavior during the enumerator's visit even if they do practice it regularly.

ITT(LATE)*Urban			0.011	0.098
			[0.033]	[0.299]
Control mean	0.232	0.232		
Control mean (rural)			0.269	0.269
Control mean (urban)			0.203	0.203
Sample	2513	2513	2513	2513
Child received health controls on time				
ITT = 1	0.011	0.021		
	[0.024]	[0.025]		
ITT(LATE)*Rural			0.051**	0.126**
			[0.025]	[0.059]
ITT(LATE)*Urban			-0.000	-0.002
			[0.039]	[0.352]
Control mean	0.573	0.573		
Control mean (rural)			0.535	0.535
Control mean (urban)			0.603	0.603
Sample	2513	2513	2513	2513

Notes: The outcomes presented in this table are binary variables taking value of one if the answer on the relevant question is affirmative and zero otherwise. The results in columns (2) – (4) are from regressions that include controls for child's age in months, sex, has health card, >=4 antenatal care controls, caregiver can read, write, education in years, marital status, indigenous, household size, number of rooms, home ownership, wealth index, strata fixed effect, and are estimated using sampling weights. Standard errors in brackets are adjusted for clustering at IU the level. Statistical significance: * 10%, ** 5%, *** 1%.

As seen in Table 3.5, in rural households, the program reduced the probability of a child being sick in the last four week (by 0.06 - 0.148), reduced the probability of diarrhea (by 0.049-0.122), and increased the probability of getting the health check-ups on time (by 0.051-0.126). These findings are not completely surprising. As mentioned earlier, the program home visitors were identified by families as health professionals because the program was implemented by the Ministry of Health. In addition, while home visitors were not health workers, regional and national program coordinators were mostly trained health professionals. Even though the intervention itself did not include references to health care practices, anecdotal evidence exists that the use of primary health services was encouraged by the program staff during home visits.

3.6 Conclusions

This study presents the results of the impact evaluation of the home-visiting intervention in Bolivia implemented as a second operational arm of the "Grow Well to Live Well" program. The objective of the intervention was to

improve child development over a comprehensive set of dimensions, including communication, cognitive and socioemotional development. The program targeted children between 6 and 36 months of age and consisted of weekly 45-minutes home visits by trained community workers. During the visits trained community workers worked with caregivers offering guidance and counseling on parenting and early stimulation. These activities were expected to improve caregiver's knowledge about childcare and stimulation, resulting in a better child-caregiver interaction and stimulation practices, and ultimately translating into better child development outcomes.

The results of the evaluation show that the program had large and significant impact on development of children in rural households. Our main results are for the overall ASQ-3 score, and they show that the program improved child development in a range from 0.14 (ITT model) to 0.35 (LATE model) SD. Across ASQ-3 dimensions, the program improved children's communication, fine motor, and problem solving skills. We also observe program impacts on socio-emotional development measured by the index based on PRIDI questionnaire (ranging from 0.049 to 0.122 points, depending on the model). It is remarkable that these results were achieved with a relatively low compliance (37%). Although the compliance was not perfect, the households that did receive the visits did so according to the program protocol.

We also observe large and statistically significant program effects on the intermediate outcomes. Specifically, we observe that the program improved parent-child interaction measured by the overall HOME score (in a range from 0.146 to 0.362 SD), and induced more cognitive stimulation of children, measured by the FCI child stimulation index (0.059-0.149 points, depending on the model). Consistent with the program logic, we do not observe program results on punishment practices and compliance with rules and routines. Overall, these intermediate results corroborate that the plausible mediating pathway for the observed impacts on child development is better quality of child-caregiver interactions and stimulation. While not anticipated at the program design, we also observe some improvements in child health indicators. These results most likely occurred because the intervention was implemented by the Ministry of Health and that the use of primary health services was encouraged by home visitors.

Regarding the urban areas, the evaluation did not find any effects on children in urban households, except for the socio-emotional index. These results can be explained by a particularly low program compliance (11%) in urban areas. The project team investigated plausible causes of the low compliance rate and identified several contributing factors such as staff shortage, household migration, possible inconsistencies in the IUs' boundaries used by survey teams and program staff.

From the public policy perspective, the important aspect when choosing how to spend scarce recourses is the intervention cost. According to the program administrative records, the cost of home visits in rural areas was US$2,627 per child or $64.3 per one home visit. While these figures seem high, they are in line with the cost of similar interventions.[59] It is also worth noting that if the program is scaled up, the cost per child will most likely go down, and even more so if it is possible to deploy the intervention upon already existing public services infrastructure.[60] An alternative approach to reducing program costs could be implementing the program curriculum through a less expensive delivery model, for example, parenting interventions.[61] Busso et al. (2017) show that parenting interventions, in which groups of parents receive training on childcare practices and stimulation, can be highly effective and are relatively inexpensive.

In summary, the results of this impact evaluation show that, after a year of exposure to the program and a total of 40 home visits on average, beneficiary children in rural areas had significant improvements in their fine motor, communicational and cognitive development. These developmental improvements were most likely achieved through improvements in parent-child interactions and an increase in the frequency and quality of early stimulation activities, which are the main mechanisms by which the program sought to influence children's development. Given that the program was implemented only in the prioritized municipalities, these results are encouraging for a possible scalability, since they provide solid and rigorous

[59] Program Project Completion Report: https://www.iadb.org/en/project/bo-l1064 (accessed on December 23, 2021).
[60] Bos et al. (2021).
[61] See, for example, Walker et al. (2015).

evidence of the effectiveness of the intervention. The cost of the program is high, but it is in line with costs of similar programs and could be reduced if the program is implemented at scale. When scaling up, specific care should be taken in supervision of the program implementation with the focus on achieving higher compliance rate. Subsequent evaluations will be relevant to show whether the program results in rural areas are sustained over time and generate beneficial effects on children in the medium and long run.

3.7 Appendix

3.7.A Description of indices

1: Construction material index measures the quality of the household dwelling's construction materials. The index takes value of one if the construction materials of roof, floor and walls are not precarious: roof is made of resistant material which is not wood, straw, mud; floor is not bare earth or loose bricks, walls are plastered. Higher index value indicates better quality of construction materials. This index is based on the precarious bathroom index in Bancalari et al. (2016).

2: Wealth Index combines information on the household's ownership of assets, dwelling characteristics and access to basic services. The wealth index used in this study includes the information of the ownership of the following assets: refrigerator, radio, television, living room set, air conditioning, washing machine, microwave, stove, motorcycle, automobile, bicycle, cellphone, landline phone, computer, tablet; the information on the sources of water, fuel, electricity; availability of bathrooms, kitchen, quality of construction material, roof and floor of the home; number of household members per bedroom. The selection of the variables and the computation of the index is based on the methodology used in the Demographic and Health Surveys (DHS) Program (Rutstein and Johnson, 2004). The index ranges from -2.1 to 2.66, with larger number corresponding to wealthier households.

3: Socio-emotional development index is the proportion of affirmative answers to the following questions about the child's socio-emotional development from the PRIDI study: does the child like to draw or paint ?; can the child play 15 minutes or more without adult attention ?; does the child

like to meet adults ?; does the child have a preference to play with favorite friends ?; does the child worry about a known sick or injured person ?; does the child has preferences for some games or activities ?; does the child cry when the caregiver has to leave? The index ranges from zero to one, with larger values indicating higher socio-emotional development.

4: Child stimulation index, or an index of participation in child activities, is a proportion of positive answers to the six UNICEF Child Care Inventory Indicators (FCI UNICEF) questions about mother's involvement in the following child stimulation activities: reading books, telling stories, singing songs, walking outside, playing games, and naming objects, teaching counting and drawing. The index ranges from zero to one, with higher value indicating better child stimulation practices.

5: Severe discipline practices index is the proportion of the affirmative answers to the FCI UNICEF inventory on the following disciplining questions: shaking; shouting; hitting on the butt with the hand; hitting with an object; calling foolish or lazy; hitting on the head, face, pulling ears; hitting on the arm, hand or leg; beating up. The index ranges from zero to one, with higher value indicating more severe discipline practices.

6: Rules and routines index is the proportion of affirmative answers to the FCI UNICEF questions about the rules and routines regarding food the child should eat, the child's bedtime, chores, the moments when the family gets together. The index ranges from zero to one, with higher value indicating more rules and routines.

3.7.B Program effects on raw scores

Table 3.7.B1: Program effects on child development (ASQ-3 raw score)

	ITT	ITT + controls	ITT Rural/Urban + controls	LATE Rural/Urban + controls
Overall ASQ-3	(1)	(2)	(3)	(4)
ITT = 1	0.339	0.496		
	[0.503]	[0.472]		
ITT(LATE)*Rural			1.176**	2.942**
			[0.528]	[1.250]
ITT(LATE)*Urban			-0.003	-0.008
			[0.719]	[6.406]
Control mean	37.296	37.296		
Control mean (rural)			35.679	35.679
Control mean (urban)			38.541	38.541
Sample	2499	2499	2499	2499
Communication ASQ-3				
ITT = 1	0.437	0.464		
	[0.643]	[0.638]		
ITT(LATE)*Rural			1.943***	4.844***
			[0.727]	[1.714]
ITT(LATE)*Urban			-0.620	-5.483
			[0.934]	[8.509]
Control mean	35.651	35.651		
Control mean (rural)			33.979	33.979
Control mean (urban)			36.939	36.939
Sample	2499	2499	2499	2499
Gross Motor ASQ-3				
ITT = 1	-0.021	0.110		
	[0.832]	[0.793]		
ITT(LATE)*Rural			-0.385	-0.951
			[0.776]	[1.948]
ITT(LATE)*Urban			0.473	4.200
			[1.246]	[11.377]
Control mean	40.096	40.096		
Control mean (rural)			39.023	39.023
Control mean (urban)			40.924	40.924
Sample	2499	2499	2499	2499
Fine Motor ASQ-3				
ITT = 1	0.637	0.871		
	[0.671]	[0.637]		
ITT(LATE)*Rural			1.769**	4.429**
			[0.784]	[1.936]
ITT(LATE)*Urban			0.213	1.930
			[0.939]	[8.368]

Control mean	36.024	36.024		
Control mean (rural)			33.933	33.933
Control mean (urban)			37.635	37.635
Sample	2499	2499	2499	2499
Problem Solving ASQ-3				
ITT = 1	0.125	0.331		
	[0.673]	[0.679]		
ITT(LATE)*Rural			2.079***	5.174***
			[0.810]	[1.933]
ITT(LATE)*Urban			-0.950	-8.412
			[0.983]	[9.005]
Control mean	35.150	35.150		
Control mean (rural)			32.458	32.458
Control mean (urban)			37.225	37.225
Sample	2499	2499	2499	2499
Socio-individual ASQ-3				
ITT = 1	0.515	0.702		
	[0.580]	[0.542]		
ITT(LATE)*Rural			0.476	1.213
			[0.664]	[1.596]
ITT(LATE)*Urban			0.867	7.726
			[0.806]	[7.693]
Control mean	39.556	39.556		
Control mean (rural)			39.000	39.000
Control mean (urban)			39.985	39.985
Sample	2499	2499	2499	2499

Notes: Scores for the overall, communication, gross motor, fine motor, problem solving, and socio-individual dimensions were computed using questionnaires adapted from the Ages and Stages Questionnaires, 3rd edition. (ASQ-3). The ASQ-3 test score is computed for each of the 13 age groups. The answers are scored 0 (No), 5 (Sometimes), 10 (Always). The raw ASQ-3 score can take values from 0 to 60 in each dimension. The results in columns (2) – (4) are from regressions that include controls for child's age in months, sex, has health card, >=4 antenatal care controls, caregiver can read, write, education in years, marital status, indigenous, household size, number of rooms, home ownership, wealth index, strata fixed effect, and are estimated using sampling weights. Standard errors in brackets are adjusted for clustering at IU the level. Statistical significance: * 10%, ** 5%, *** 1%.

Table 3.7.B2: Program effects on parent-child interaction (HOME raw score)

	ITT	ITT + controls	ITT Rural/Urban + controls	LATE Rural/Urban + controls
Overall HOME	(1)	(2)	(3)	(4)
ITT = 1	-0.097	-0.042		
	[0.154]	[0.146]		
ITT(LATE)*Rural			0.300***	0.741***
			[0.114]	[0.275]
ITT(LATE)*Urban			-0.294	-2.624
			[0.237]	[2.102]
Control mean	8.542	8.542		
Control mean (rural)			8.075	8.075
Control mean (urban)			8.905	8.905
Sample	2513	2513		
Responsiveness HOME				
ITT = 1	-0.062	-0.010		
	[0.144]	[0.141]		
ITT(LATE)*Rural			0.362***	0.896***
			[0.106]	[0.255]
ITT(LATE)*Urban			-0.284	-2.527
			[0.226]	[2.050]
Control mean	3.942	3.942		
Control mean (rural)			3.493	3.493
Control mean (urban)			4.290	4.290
Sample	2513	2513		
Acceptance HOME				
ITT = 1	-0.035	-0.033		
	[0.039]	[0.038]		
ITT(LATE)*Rural			-0.062	-0.156
			[0.047]	[0.117]
ITT(LATE)*Urban			-0.011	-0.097
			[0.055]	[0.485]
Control mean	4.601	4.601		
Control mean (rural)			4.582	4.582
Control mean (urban)			4.615	4.615
Sample	2513	2513		

Notes: The HOME inventory short version has 2 dimensions: responsiveness and acceptance. The responsiveness dimension has 6 items, and acceptance dimension has 5 items. Items in the responsiveness dimension are scored yes = 1, no = 0, items in the acceptance dimension are scored yes = 0, no = 1. The total score in each dimension is the sum of the item scores. The maximum score in the responsiveness dimension is 6, 5 in the acceptance dimension, and 13 in the overall score. The

results in columns (2) – (4) are from regressions that include controls for child's age in months, sex, has health card, >=4 antenatal care controls, caregiver can read, write, education in years, marital status, indigenous, household size, number of rooms, home ownership, wealth index, strata fixed effect, and are estimated using sampling weights. Standard errors in brackets are adjusted for clustering at IU the level. Statistical significance: * 10%, ** 5%, *** 1%.

Chapter 4: More Money More Learning? Evidence from Exogenous Spending Variation in Brazil

4.1 Introduction

Does increase in spending on education matter for student achievement? Policy makers and practitioners are under constant pressure of stakeholders demanding better quality of education. Thus, it is essential to understand whether achieving these results implies higher costs and additional resources that likely avert funds from other demands on public expenditure or increase tax burden. Given policy objectives of improving education quality and achieving better student outcomes, how much do we know about the effectiveness of providing additional resources for achieving these goals? The effectiveness of resource-based policies in education was under scrutiny since Coleman Report (Coleman et al., 1966). Hanushek (2003, 2006) concludes that there is little evidence of the relationship between increase in resources for education and changes in student outcomes. On the other hand, recent rigorous research finds that increasing funding for education can improve student outcomes (Jackson et al., 2016; Lafortune et al., 2018). It is not clear how these two results can be reconciled (Hanushek, 2003). Among suggested explanations, it was pointed out that the way how money in education is spent can be as relevant (or even more so) as the level of provided resources (Vegas and Coffin, 2015). All in all, as of today, there is no consensus on whether additional educational spending translates into better student outcomes.

In this study I use the data from a sample of Brazilian municipalities to analyze the effect of educational expenditure on student test scores. To address the endogeneity problem stemming from spending decisions, I exploit the allocation mechanism of the federal transfer *Fundo de Participação dos Municípios* (FPM) across Brazilian municipalities. In this allocation mechanism, the amount of transfer received by municipalities is determined by their population size, which generates exogenous and discontinuous jumps in the funds received by municipalities at given population thresholds. These discontinuities provide a source of exogenous variation in resources available to local governments which I exploit in the identification strategy. Specifically, I implement an instrumental variables (IV) approach in which endogenous discretional spending on education is

instrumented with federal transfers, which I compute using the transfer allocation formula.

This study contributes to several strands of literature. First, this evaluation contributes to a growing body of literature evaluating the effect of unconditional or unrestricted resources available for spending on education. Specifically, the study provides evidence of the unconditional spending on education in a developing country, while most existing evidence on this kind of spending comes from developed countries. The existing evaluations identify the effects by exploiting plausible exogenous changes in spending induced by school finance reforms (Card and Payne, 2002; Guryan, 2001; Jackson et al., 2016; Lafortune et al., 2018), Great Recession (Jackson et al., 2018), state financing formulas of school districts (Gigliotti and Sorensen, 2018; Papke, 2005), changes in property values (Miller, 2017), geographical discontinuities (Gibbons et al., 2017). Several recent rigorous studies present evidence of causal effects of educational spending on long-run outcomes, such as college enrollment and attainment (Hyman, 2017), labor market outcomes (Jackson et al., 2016) and poverty and crime (Johnson and Jackson., 2017).

This study also adds to the literature that evaluates the impact of providing more educational resources in developing countries. The main contribution is that this study focuses on unconditional education spending, while the evidence from developing countries comes mostly from experimental evaluations of programs providing resources earmarked for specific spending categories, such as materials (Das et al., 2013), textbooks and school supplies (Mbiti et al., 2018), textbooks and classroom constriction (Glewwe et al., 2009). In the Brazilian context, existing literature focuses on estimating the effects of the equalization Fund for Maintenance and Development of Basic Education and Teacher Appreciation (FUNDEF), which redistributes resources earmarked for basic education within states between municipalities (Haddad et al., 2017; Gordon and Vegas, 2004). In this study, I estimate the causal effect of educational spending on student achievement in the short run. To this end, I exploit a plausibly exogenous variation in the discretional expenditure on education in Brazilian municipalities stemming from sharp discontinuities in federal transfers at population thresholds.

Finally, this study contributes to the literature by exploiting an exogenous change in resources available to local governments. For the specific case of the FPM transfer, several studies have utilized its allocation rule to identify the impact of actual FPM using Regression Discontinuity design (e.g., Brollo et al., 2013; Bastos and Straume, 2016). To the best of my knowledge, this is the first study that uses the FPM allocation rule to study the short-run causal effects of unconditional educational spending on student outcomes. Specifically, I implement an instrumental variables (IV) approach, which provides a credible source of exogenous variation in spending on education. The validity of the empirical strategy used in this study hinges on the assumption that there are no other mechanisms through which changes in federal transfers at population thresholds affect student achievement other than discretional spending on education. I provide much evidence suggesting that federal transfer is a strong and plausible exogenous instrument. Specifically, I show that there is a statistically significant relationship between federal transfers and endogenous discretional spending on education. I also provide evidence that there is no selective sorting of municipalities at population thresholds which generate discontinuities in the transfers and spending on education. Finally, I show that there is no relationship between federal transfers and student outcomes that are not affected by changes in discretional spending on education but are potentially affected by changes in other spending chapters such as health or social protection.

The results indicate that a 1% increase in federal transfer leads to an increase in total spending on education by 0.17% and an increase in discretional spending on education by 0.38%. This change in resources translates into an increase in standardized test scores. Specifically, I find that each 10% increase in spending on education boosts students test scores by about 0.13 of a standard deviation (SD). These estimates are very similar to the results reported in other studies: Jackson et al. (2018) find that a decrease in funding to schools by 10% lead to a decline in test scores by about 0.078 of a SD, and Lafortune et al. (2018) find that the similar increase in spending improved student achievement between 0.12 and 0.24 of a SD. I also explore the mechanisms through which additional resources available for spending on education affect student achievement. To this end I look at the relationship between spending on education and levels of school inputs. The results show

that traditional school inputs, such as class size, teacher level of education and quality of school infrastructure, do not mediate the relationship between increase in spending and changes in student achievement.

The reminder of this chapter is structure as follows. Section 4.2 reviews recent empirical evidence on the relationship between educational spending and student achievement. Section 4.3 presents the FPM transfer and explains its allocation mechanism. Section 4.4 describes the data. Section 4.5 presents the econometric strategy. The results are reported in Section 4.6. Section 4.7 concludes. Some additional results and information are presented in the Appendix section 4.8.

4.2 Related literature

The literature on the effects of spending on education can be grouped into two large strands. The first strand evaluates the effects of unconditional and/or unrestricted additional resources. The second strand evaluates the effect of additional funds earmarked for specific spending items or school inputs. This study contributes to the first strand of the literature evaluating the effect of unconditional or unrestricted resources available for spending on education. To the best of my knowledge, this is the first study that evaluates short run causal effect of an increase in educational spending on test scores in the context of a developing middle-income country using an IV strategy. While there are studies that estimated effectiveness of educational spending on student achievement using instrumental variables (Jackson et al., 2018; Miller, 2017), this is the first one that uses as an instrument federal transfers computed using the transfer allocation formula.

Several recent studies have contributed to the body of the research on unconditional spending on education.[62] Most of these studies use the event-based approach and exploit large and permanent increases in school spending stemming from the passage of the School Finance Reforms (SFR) in the United States (US). For example, in a recent study Lafortune et al. (2018) use

[62] For studies that evaluate the effects of earmarked spending in developed countries see, e.g., Dragoset et al. (2017), Gamse et al. (2008), Martorell et al. (2016), Leuven et al. (2007). For studies in developing countries, see, e.g., Carneiro et al. (2015), Das et al. (2013), Pradhan et al. (2011), Mbiti et al. (2018), Glewwe et al. (2009).

the US National Assessment of Educational Progress (NAEP) data to evaluate the impact of post-1990 SFR on spending and achievement of low-income school districts. The study exploits the plausible randomness in the reforms timing in an event-study framework. The authors conclude that reforms lead to increases in spending in low-income school districts and increased the district-level student achievement. They find that the implied impact of the increase in the annual spending by 10%[63] is between 0.12 and 0.24 SD of student test scores. Card and Payne (2002) analyze the consequences of the SFR in 1980s. They find that states where school finance system was declared unconstitutional increased funding of low-income districts, which helped closing the gap in spending between richer and poorer districts. In this study the effect of closing spending gaps on student achievement is estimated using the self-selected sample of SAT-takers. The findings show that equalization of spending leads to narrowing of test score outcomes across family background groups. Another related study is the work by Jackson et al. (2018) evaluating the effect of educational spending cuts induced by the Great Recession of 2008 on student outcomes. Using the panel data covering 2000-2015, the authors employ an IV approach instrumenting educational spending with the interaction between the share of state educational K12 spending and the number of years post-recession. Using the NAEP test scores data, they find that a 10% decline in per student spending led to decline in test scores by 0.078 of a SD. A study by Miller (2017) uses different IV approach to estimate the effects of spending on student achievement. Miller (2017) instruments the endogenous spending with a simulated school revenue calculated by interacting changes in property values with fixed school finance formulas. Using 2009-2013 Stanford Education Data Archive test scores, he finds that a 10% increase in spending increases fourth grade test scores by about 0.09 SD. Among studies that evaluate the effects of SFR on other dimensions including some educational outcomes, Jackson et al. (2016) find that a 10% increase in annual per student spending induced by SFR lead to 0.31 more completed years of education, 7% higher wages, 3.2 percentage point reduction in adult poverty. Johnson and Jackson (2017) find that an increase in K12 spending raised educational attainment and earnings, and reduced likelihood of poverty and incarceration in adulthood.

[63] Lafortune et al. (2018) report the effects for $1,000 spending increase, which is roughly 10% of the average spending $9,540.35 (footnote 2 in Jackson et al., 2018).

In regard to evaluation of the effects of spending on education in the short run, Gigliotti and Sorensen (2018) evaluate the effect of a plausible exogenous variation in a per-student expenditure stemming from the New York State aid formula which allows school districts to keep the levels of state aid when their student enrollment declines. They find that an additional per-student spending of 10% can increase educational performance from about 0.03 SD to 0.09 SD.[64] Papke (2005) estimates the effect of a change in per-pupil expenditure induced by the state educational aid resulting from Michigan's school finance reform. She finds that a 10% increase in per-student spending is associated with a two percentage point increase in the pass rate on the end of the year examination.

This study also relates to the literature that examines the effects of educational funding in Brazil. While I evaluate the effect of changes in discretional unrestricted spending on education induced by changes in the federal transfer, other studies evaluate changes in resources available to Brazilian municipalities through the state transfer FUNDEF. For example, Haddad et al. (2017) examine the effect of changes in FUNDEF on student test scores using the panel data for a period 2003-2009. The authors use an IV approach exploiting the information prior to the reform that introduced FUNDEF to construct a simulated FUNDEF transfer and use it as an instrument for the actual FUNDEF transfer. They find that increasing FUNDEF resources translate into a small increase in test scores. Gordon and Vegas (2005) investigate the effect of FUNDEF on enrollment, school spending, teacher credentials, class size and student achievement. Using the same methodology as Haddad et al. (2017), Gordon and Vegas (2004) find that FUNDEF-induced spending increases enrollment, reduces class size and helps meet federal mandate stipulating that teachers must have at least a secondary education degree. They do not find that FUNDEF-induced spending improves student outcomes, except for some evidence for low-achieving and non-white students.[65]

[64] Gigliotti and Sorensen (2018) estimate an increase in test scores from 0.015 to 0.045 SD stemming from an increase of $1,000 in per student spending, which is less than 5% of average education spending in New York.

[65] For a review of other studies that analyze the effects of FUNDEF see Haddad et al. (2017).

4.3 The FPM transfers

Brazil is one of the most decentralized federal countries in the world. Its government is structured into three tiers: Federal government (Union), State and Municipal governments. Administratively, Brazil is composed of one Federal district (where the capital city Brasilia is located), 26 States and over 5,500 municipalities. Both states and municipalities have autonomous administrations, they collect their own taxes and receive transfers from the higher tier governments. State governments are responsible for maintaining state highway systems, low-cost housing programs, public infrastructure, telephone companies, and transit police. Both state and municipal governments are responsible for public primary and secondary schools and public hospitals. Municipal governments are also responsible for water, sewerage, and garbage services.

The revenue of municipalities consists of resources received from federal and state governments and the revenue from local taxes. A large part of the resources available for spending on public goods provided by local governments comes from federal and state transfers. An important source of funding is the federal transfer FPM. In the sample of municipalities used in this study, this transfer represents about 60% of all transfers received by local governments and 30% of their total revenue. FPM is an automatic federal transfer to local governments. The law mandates that 25% of its resources must be allocated to spending on education.[66] 15% of the FPM funds is automatically "discounted" to the state fund FUNDEF,[67] which further redistributes resource across municipalities to cover mandatory primary education spending. After the FUNDEF discount, at least 10% of the FPM must be spent on education.

This study focuses on the FPM transfer because of its unique feature, which is its assignment rule that establishes that the amounts received by local governments are defined by population of municipalities. According to this assignment rule, the FPM transfer received by municipalities jumps

[66] Mandated by the Federal Constitution. Other than this restriction, distribution of the FPM transfer across expenditure chapters is unregulated.
[67] For details on the FUNDEF fund and other institutions involved in financing Brazilian system of fundamental education see Appendix 4.8.A.

discontinuously at given population thresholds. These discontinuities provide exogenous variation in levels of resources received by local government which I exploit in the IV approach.[68]

The FPM allocation mechanism groups municipalities in population brackets. These population brackets determine the coefficient which is applied to the share of the FPM transfer received by the state. Municipalities with higher population brackets have larger coefficient and hence receive more FPM funds, while municipalities with lower population brackets have smaller coefficients and receive less FPM funds. Each of 26 states receives its own share of the FPM. This share is than redistributed within state according to coefficients that change depending on population.[69]

Specifically, let FPM_{ikt} be the transfer received by municipality i in state k in year t. The FPM transfer allocation mechanism works as follows:

$$FPM_{ikt} = \frac{FPM_{kt}\omega_i}{\sum_{i\epsilon k}\omega_i}$$

Where FPM_{kt} is the amount assigned to the state k in year t, ω_i is the FPM coefficient of municipality i determined by the i's municipality population size.

In defining the sample of municipalities in this analysis, I follow Brollo et al. (2013) and focus on municipalities with the population above 6,792 inhabitants and below 50,940.[70] These municipalities represent about 90% of all Brazilian municipalities and 34% of the total population. The sample

[68] Several studies have utilized the FPM allocation rule to identify the impact of actual FPM transfers using the Regression Discontinuity (RD) design. For example, Brollo et al. (2013) evaluate the effect of the FPM on corruption using fuzzy RD design, Litschig and Morrison (2013) look at the effect of the transfers on development outcomes using sharp RD design, Bastos and Straume (2016) use fuzzy RD design to measure the effect of the FPM on enrollment in public and private preschools.

[69] The allocation of the FPM transfer to state capitals and the federal district Brasilia does not follow the general rule. Therefore, the state capitals and the federal district are excluded from the analysis.

[70] As noted in Brollo et al. (2013), imposing these limits on the sample restricts the interpretation of the results to municipalities of similar size excluding very large and very small municipalities.

covers seven population thresholds: 10,189; 13,585; 16,981; 23,773; 30,564; 37,356; and 44,148. The intervals between the first three thresholds are 3,396, while the intervals between the following four thresholds are twice as large (6,792). Table 4.1 shows the population brackets and the associated FPM coefficients ω_i.

Table 4.1: FPM coefficients, actual and theoretical transfers

(1) Population Brackets	(2) FPM Coefficient	(3) Actual transfers	(4) Theoretical transfer	(5) Municipalities in year 2006
6,793 – 10,188	0.6	1.58	1.56	482
10,189 – 13,584	0.8	2.11	2.12	381
13,585 – 16,980	1	2.64	2.67	297
16,981 – 23,772	1.2	3.15	3.19	421
23,773 – 30,564	1.4	3.67	3.73	236
30,565 – 37,356	1.6	4.22	4.27	154
37,357 – 44,148	1.8	4.65	4.73	102
44,149 – 50,940	2	5.23	5.32	56

Notes: Actual and theoretical FPM transfers are in constant million US dollars in 2016; Sample comprises municipalities with test scores data in 2007 and spending and school inputs data in year 2006 for which all spending variables are positive and reported spending on education is at least 25% of total revenue.

The coefficient ω_i is assigned to each municipality by the Brazilian Federal Court of Audit (TCU). The assignment of coefficient is done annually and is based on the population estimates for the previous year calculated by Brazilian Institute of Geography and Statistics (IBGE).[71] In this analysis I use the IBGE population estimates to calculate theoretical FPM transfers. As seen in columns three and four of Table 4.1, the IBGE population estimates do not perfectly predict the actual FPM transfers. While the process that IBGE follows in producing its population estimates makes manipulation of these figures unlikely, it was noted in Brollo et al. (2013) that discrepancies may occur because of the imperfect adjustments of coefficients when municipalities split or merge, or other distortions in the FPM allocation procedure.

[71] IBGE employs a top-down method for population estimates. That is, the estimation of population at municipal level must be consistent with the estimates at the state level, which in turn must be consistent with population estimates at the federal level. The Online Appendix for Brollo et al. (2013) provides a detailed description of the process followed by IBGE for computation of population estimates.

4.4 Data and descriptive statistics

4.4.1 FPM transfers and population data

For the actual FPM transfers received by each municipality I use the FPM data for years 2002-2006 from the Brazilian National Treasury.[72] The main analysis focuses on the data for year 2006, but I also present the estimations using 2002-2005 FPM data.

I compute the theoretical FPM transfers using the allocation rule described in the previous section and apply it to the IBGE population estimates reported on the IBGE website.[73] Since the amount of the federal transfer received by each municipality is computed according to the IBGE estimates of population in previous year, I use population estimates in year *t-1* to compute the theoretical FPM transfers in year *t*. Specifically, I use population data in year 2005 to compute theoretical transfers that municipalities would receive in year 2006, population data in 2004 to compute theoretical transfers received in 2005, and so on.

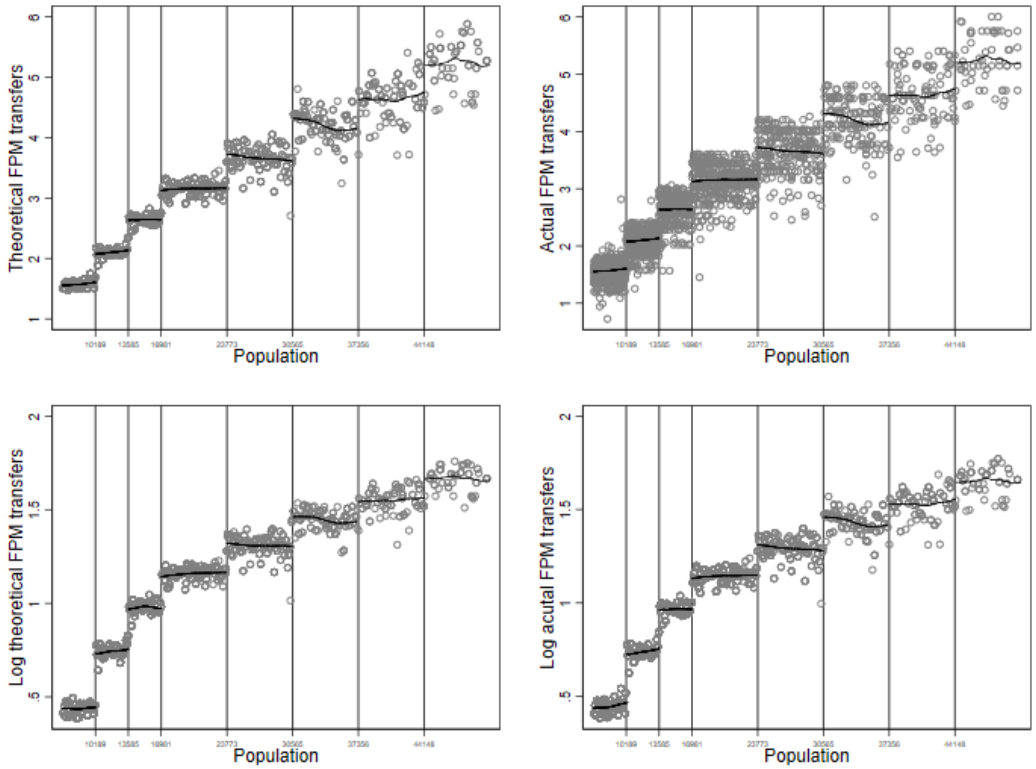
The third and the fourth columns of Table 4.1 show the actual and the theoretical FPM transfers received by each municipality in year 2006 by population brackets. As seen, the average actual transfers are very similar to the theoretical transfers based on the IBGE population estimates. On average, a municipality received US $2,799 million in 2006.[74] The average of the predicted transfers is slightly larger and amounts to US $2,826 million. There are 2,129 municipalities in the analytical sample, and about 74% of them are in the first four population brackets.

Figure 4.1 depicts the FPM transfers against population. Panels in the top row show the relationship for the transfers expressed in levels, while the bottom row show the relationship for the transfers expressed in natural logarithm (logs). Panels on the left are for theoretical transfers, and panels on the right

[72] https://www.tesourotransparente.gov.br/temas/estados-e-municipios/transferencias-a-estados-e-municipios (accessed on October 3, 2021).
[73] https://www.ibge.gov.br/estatisticas-novoportal/sociais/populacao/9103-estimativas-de-populacao.html?=&t=downloads (accessed on October 3, 2021).
[74] All monetary amounts are expressed in constant US 2016 dollars.

are for the actual transfers. The circles of the scatterplot correspond to the local means where the FPM transfers are averaged over cells of 100 inhabitants. The black lines are the smoothed averages computed separately for each population bracket. The labels on the horizontal axis and vertical lines show the population thresholds that delimit population brackets. As seen, there are clear jumps in both theoretical and actual FPM transfers at population thresholds. The jumps are similar in size across all thresholds and amount to about US $0.5 million. Also, note that there is a variation in transfers received by municipalities within the population brackets. This happens because different states receive different shares of the FPM transfers which are afterwards distributed across municipalities according to the population thresholds.

Figure 4.1: Actual and theoretical FPM transfers



Notes: Actual and theoretical FPM transfers are shown in constant million US dollars in 2016. The sample comprises municipalities with population between 6,793 and 50,940 inhabitants, with the test scores data in 2007 and spending and school inputs data in year 2006, for which all spending variables are positive, test scores data is available, and reported spending on education is at least 25% of the total revenue.

4.4.2 Education financing data

Municipal revenue and spending data was retrieved from the Brazilian National Treasury portal.[75] This data is an accounting register which shows revenues and expenditures by spending and revenue chapters. Unlike administrative federal transfer data, municipal spending and revenue is self-reported by municipalities. Thus, a concern in regard to the data quality may arise. To ensure the quality of the data, I rely on the Brazilian Constitution mandate that municipalities must spend at least 25% of their revenue on education. I use this rule and limit the initial sample of municipalities to only those for which the reported spending on education is at least 25% of the reported total revenue. Municipalities that pass this restriction make up the main study sample.[76]

To facilitate the interpretation of the results, I transform all spending and revenue variables in logs and use the transformed variables in all estimations. To keep the number of observations constant across all estimations, I restrict the main sample to the observations for which spending and revenue variables are nonzero and test scores are reported. The municipalities included in the sample after this restriction make up the analytical sample.[77] I check the quality of the data in the analytical sample by looking at the correlation between the FPM transfer self-reported by municipalities and the FPM transfer registered in the administrative data. I find that in year 2006 the correlation is 0.98, while in years 2002-2005 the correlation ranges from 0.89 to 0.98.

The descriptive statistics of spending and revenue variables are presented in Panel A of Table 4.2. The first and the second columns of Table 4.2 show means and standard deviations of the variables in the main sample, which comprises municipalities observed in 2006 with population between 6,793 and 50,940 inhabitants and spending on education at least 25% of the total revenue. The third and the fourth column show the means and standard

[75] The data was obtained from the Brazilian National Treasury portal in 2019 from the electronic link https://www.tesouro.fazenda.gov.br/finbra-financas-municipais (accessed on May 20, 2019).
[76] Application of this rule reduces the initial sample of municipalities by 7%.
[77] Moving from the main sample to the analytical sample reduces the number of observations by 3%.

deviations of the variables in the analytical sample, where I keep the main sample observations for which all spending variables are positive and test scores data is reported.

Table 4.2: Summary statistics

	Main sample		Analytical sample	
Variable	Mean	SD	Mean	SD
Panel A: Revenue and Spending				
Theoretical FPM transfer	2.80	1.06	2.83	1.06
Actual FPM transfer	2.78	1.04	2.79	1.04
Total revenue	9.56	6.22	9.67	6.28
Total spending	9.57	6.13	9.68	6.18
Spending on education	3.23	2.03	3.27	2.04
Mandatory spending on education*	1.79	1.12	1.81	1.12
Discretional spending on education**	1.44	1.24	1.46	1.25
Panel B: Population and enrollment				
Population	18,533.2	10,226.79	18,701.92	10,267.01
Total enrollment	1,877.37	1,197.29	1,896.08	1,196.82
Enrollment in municipal schools	1,519.78	1,080.18	1,547.76	1,072.66
Panel C: Inputs in Municipal Schools				
Class size	23.04	4.16	23.19	4.03
Share teachers with higher education	0.33	0.29	0.33	0.29
Share schools with electricity	0.98	0.06	0.98	0.05
Share schools with water supply	0.73	0.27	0.74	0.26
Share schools with sewerage	0.32	0.38	0.33	0.38
Share schools with playground	0.09	0.21	0.09	0.21
Share schools with computer room	0.15	0.28	0.16	0.28
Share schools with library	0.29	0.33	0.29	0.33
Number of municipalities	2,199	2,199	2,129	2,129

Notes: *Mandatory spending on education amounts to resources received by municipalities from FUNDEF fund. **Discretional spending on education is Spending on education minus FUNDEF. Absolute monetary values are expressed in constant million US dollars in 2016. Main sample comprises municipalities with population between 6,793 and 50,940 inhabitants with test scores data in 2007 and spending and school inputs data in year 2006 for which reported spending on education is at least 25% of total revenue. Analytical sample is restricted to observations with positive spending variables and non-missing test scores data.

As shown in the first column of Table 4.2, the average municipality in the main sample received in 2006 US $9.56 million in total revenue and spent

slightly more – US $9.57 million. The average spending on education was US $3.23 million[78] and the average FPM transfer was US$ 2.8 million. The FUNDEF resources represented about 55% of the total spending on education. On average, municipalities allocated $1.44 million to the discretional spending on education. This discretional spending is the endogenous variable that I use in the analysis to estimate the effect of changes in the education expenditure on student achievement. I use this specific indicator of education expenditure because it represents the funds that are not affected by any specific spending chapter.

4.4.3 Student achievement and school inputs data

The measure of student achievement is the national standardized test *Prova Brasil*. *Prova Brasil* data is available online from the National Institute of Studies and Research (INEP).[79] The test started in 2005 and is administered every 2 years. In the analysis I use the 2007 *Prova Brasil* round. In this year the test was taken in all public schools with at least 30 students.[80] Students take the test at the exit of primary school (grade 4) and secondary school (grade 8). While primary school attendance is mostly universal, dropout rate in secondary education in Brazil is one of the largest among the OECD countries (OECD, 2018). Because additional resources may affect the composition of who takes the exit exam in secondary school, I limit the analysis to fourth grade test scores. Since I analyze the effects of the municipal educational spending, in the main analysis I use fourth grade tests scores in the municipal schools. In the falsification tests I also use fourth grade test scores in the state schools. For the purposes of the analysis, I use math and language test scores. I also compute a combined test score as the average of raw test scores in math and language. I standardize raw test scores

[78] This gives a per student spending of US $1,704 which is larger than per primary student spending computed using the World Bank World Development Indicators WDI (According to WDI, in US dollars of 2015 Brazil spent $1,205 per primary education student in 2006 and $1,334 in 2007). This discrepancy may happen because the spending data is for all levels of education, but the enrollment data is only for primary education students.
[79] http://portal.inep.gov.br/educacao-basica/saeb/resultados (accessed on October 3, 2021).
[80] In 2005 *Prova Brasil* test was administered only in urban schools with at least 20 students enrolled. I do not use 2005 *Prova Brasil* data because the sample consists mostly of rural and small urban municipalities underrepresented in the 2005 round. The test was also administered in 2009 and 2011, but because of the changes in the denomination of primary and secondary education it was administered in grades 5 and 9.

to mean zero and standard deviation one and aggregate them at the municipality level weighting by the number of test takers.

To explore the mechanisms through which the changes in the level of resources affect test scores, I look at the relationship between transfers and school inputs. For school inputs I employ the school census data, which is available on the INEP webpage[81] for years 2002-2006. Brazilian school census is compulsory. It is conducted annually by the Ministry of Education in cooperation with the state-level education departments. School census gathers the data on student enrollment, teachers, infrastructure, among other information. The quality of the census data is verified in yearly inspections in a random sample of schools.

From the school census I use the information on the number of students and the number of classes to compute the class size. I also use the information on the number of teachers by the level of education to compute the share of teachers with tertiary education. Finally, I use the binary indicators of the availability of school infrastructure and equipment (whether school has electricity, water supply, sewerage, playground, computer room and library) and, following Katz et al. (2001), I construct an index of school infrastructure quality.

Summary statistics on the school inputs are reported in Panel C of Table 4.2. In the sample of analyzed municipalities, the average class has 23 students. About one third of teachers have tertiary education. Almost all schools have electricity. The coverage of the water supply is lower – about 73% – and only about one third of schools have sewerage. Computer rooms are available in 15% of schools, almost 30% of schools have libraries and only 9% of schools have playgrounds.

4.4.4 Time invariant characteristics of municipalities

To check sorting of municipalities around population thresholds I perform the balance tests on observable pre-determined characteristics of municipalities. To this end, I use the data on the time invariant attributes of

[81] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados (accessed on October 3, 2021).

municipalities available from the Brazilian Institute of Research in Applied Economics (IPEA). More specifically, I look at the following attributes defined at the municipal level: the area of the municipality (measured in squared kilometers), the geographical location (altitude, latitude, longitude), the distance to the state capital and to the federal union capital. In addition, I also test for balance on proxies of pre-treatment development indicators measured by income per capita and the share of black population reported in the 1991 population census. The results of this analysis are presented in Section 4.6.1.

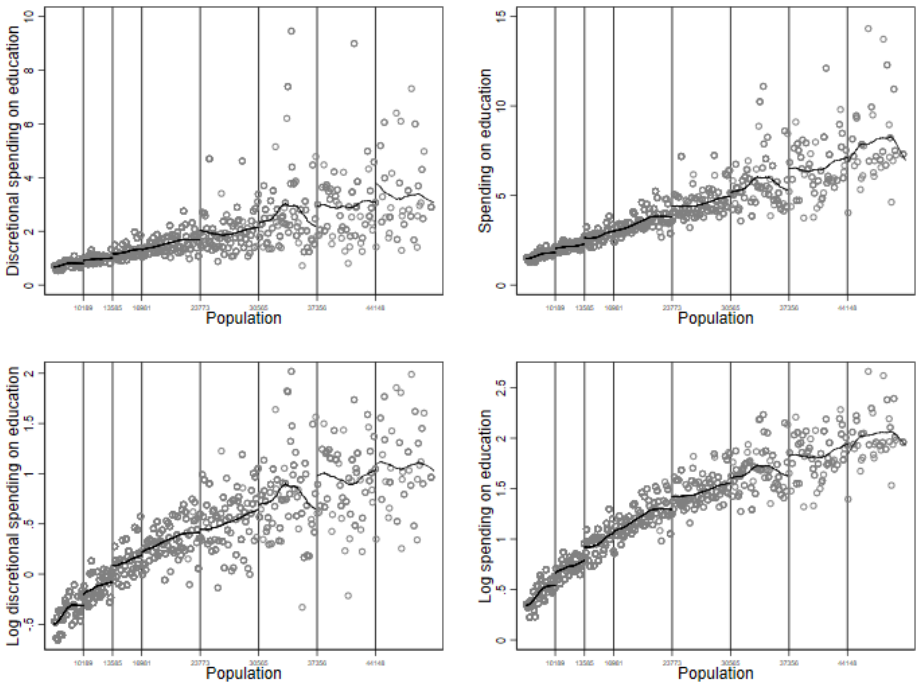
4.5 Econometric strategy

The objective of this study is to estimate the causal effect of educational spending on student achievement. The key challenge of this analysis is the potential bias in the relationship between public spending on education and student outcomes. This bias may occur for different reasons. For example, municipalities might be willing to spend more on education because they are wealthier or value education more, which by itself may be correlated with student achievement. Hence, the estimation of a simple Ordinary Least Squares regression of student outcomes on educational spending would most likely yield biased estimates. If panel data is available, a possible solution could be analyzing the changes within municipalities over time. In this case, all potential sources of bias that are constant in time would be controlled for. However, we would still have to deal with the sources of bias that are not constant over time. Another solution consists in looking at the variation in education spending that occurs because of the reasons unrelated to other policies and changes that might affect student outcomes. This is the strategy that I use in the analysis.

To estimate the causal effect of educational spending on student achievement I exploit plausible exogenous variation in spending on education induced by changes in the theoretical FPM transfer at population thresholds. The causal interpretation of the results hinges on the assumption that student test scores in municipalities with different levels of theoretical FPM transfer are not differentially affected by the changes in the theoretical FPM for reasons other than through educational spending. This requires, first of all, that the theoretical FPM transfer is a good shifter of spending on education. As it was

explained in Section 4.3, a share of the actual FPM transfer must be allocated by municipalities on discretional educational spending. The evidence presented in Table 4.1 and Figure 4.1 shows that the actual and theoretical FPM transfers have similar levels and exhibit similar jumps at population thresholds. On Figure 4.2 I present the results of the same analysis as in Figure 4.1 for spending on education and discretional spending on education. The jumps in spending variables are observed at almost all FPM population thresholds, even though the changes in the levels of spending variables are not as sharp as they are in the theoretical and actual FPM transfers. Given this evidence, I argue that the theoretical FPM is a strong instrument for educational spending. I formally test this assertion in Section 4.6.2 where I estimate the reduced form relationships between the theoretical FPM transfers and spending variables.

Figure 4.2: Spending on education at FPM population thresholds



Notes: The figure shows discretional spending on education and total spending on education at population thresholds. Spending variables are in constant million US dollars in 2016. Sample comprises municipalities with population between 6,793 and 50,940 inhabitants with test scores data in 2007 and spending and school inputs data in year 2006 for which all spending variables are positive, test scores data is available, and reported spending on education is at least 25% of total revenue.

The second requirement for the estimates to have causal interpretation is that the theoretical FPM transfer is, in fact, an exogenous instrument. I focus the analysis on exploiting of changes in the theoretical FPM at given population thresholds. Therefore, so long as the changes in the instrument at these thresholds are not related to student achievement, through the mechanisms other than spending on education, we could claim that the instrument is exogenous. This implies that, first of all, there are no other policies and programs in place other than FPM that exploit these population thresholds. This concern was explored in Brollo et al. (2013), and Litschig and Morrison (2013). To my knowledge, there are no other programs that used these same cutoffs in the period of analysis. Second, there should not be any strategic sorting of municipalities in the vicinity of the population thresholds. I formally test the absence of strategic sorting in Section 4.6.1 by checking that there are no systemic differences between municipalities just below and just

above population thresholds in pre-determined characteristics of municipalities. Also, in Section 4.6.1 I formally test for "bunching" of municipalities at population thresholds. Overall, the results suggest that strategic sorting of municipalities at population thresholds is not a concern in the study.[82] Finally, in Section 4.6.5 I show that changes in educational spending of municipalities induced by the changes in the theoretical FPM are unrelated to the changes in student outcomes in the state school system. Given that state schools do not benefit from the municipal spending on education, this corroborates the assumption that theoretical transfers affect student test scores only through educational spending.

To isolate the plausible exogenous variation in educational spending, I use an IV regression model where I instrument the endogenous spending on education with theoretical FPM transfer. Specifically, in the estimation model, I compare changes in student achievement in municipalities just below and just above population thresholds with lower/higher level of resources induced by theoretical FPM transfer. Formally, I estimate:

$$S_i = f(P_i) + \beta_T Z_i + \gamma_s + e_i \quad (1)$$

$$Y_i = f(P_i) + \beta \hat{S}_i + \gamma_s + \mu_i \quad (2)$$

where S_i is the endogenous treatment expressed as log of discretional spending on education in municipality i; Y_i is the outcome of interest (test scores or school inputs) in municipality i; $f(P_i)$ is a third-order polynomial of population in municipality i; Z_i is the exogenous instrument expressed as a log of theoretical FPM transfer; $\hat{S}_i$ is a prediction of the endogenous treatment estimated in equation (1); γ_s are state fixed-effects; e_i and μ_i are error terms.

In the first stage (equation 1) the endogenous educational spending is regressed on the exogenous instrument, flexible population polynomial and

[82] This is in line with what concluded earlier studies that utilized theoretical FPM as an instrument in the same period of time (Brollo et al., 2013), as well as the studies that used the FPM data for 80ties (Litschig and Morrison, 2013). However, Litschig (2012) detects evidence of manipulation of the 1991 population estimates, which determined the FPM transfers through 90ties.

state fixed effects. In the second stage (equation 2), the impact of educational spending is estimated by regressing the outcome Y_i against $\hat{S}_i$, unbiased fitted values from equation (1), population polynomial and state fixed effects. For some analysis I also estimate the reduced form relationships between theoretical FPM transfer and the outcomes of interest by estimating equation (1) with test scores and municipal spending on the left-hand side. In all estimations I cluster standard errors at the municipality level. In some specifications I also add the vector of predetermined characteristics of municipalities to check for robustness and reduce residual variance.

4.6 Empirical results

4.6.1 Balance and validity tests

One of the requirements for the instrument to be exogenous is that it is unrelated to pre-determined characteristics of municipalities. I test this assumption by checking that observed predetermined characteristics do not differ between municipalities just below and just above population thresholds. I analyze such time-invariant municipal characteristics as the area of the municipality (measured in squared kilometers), its altitude and geographical coordinates, distance to the state capital and distance to the federal union capital. In addition, I also test for the balance in income per capita and share of black population in 1991. For the balance tests I estimate discontinuities in these characteristics using equation (1) with municipalities' characteristics on the left-hand side and log theoretical FPM transfer on the right-hand side, controlling for a third-order population polynomial and state fixed effects. The results of the balance tests in the analytical sample are reported in Table 4.3. They show that there are no significant discontinuities in the pre-determined characteristics of municipalities, except for Area at the 10% level of statistical significance.[83]

Table 4.3: Balance tests of time-invariant and pre-determined characteristics of municipalities

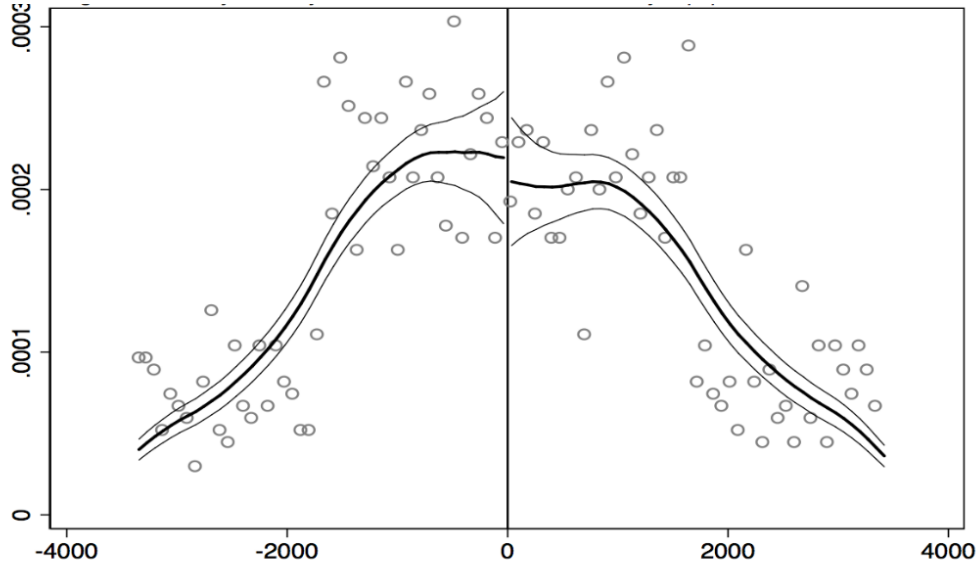
[83] In the Appendix 4.8.B Table 4.8.B.1 I show the results of the balance test in the main sample of municipalities that includes observations without test scores and with zero spending/revenue. In this sample I do not find statistically significant discontinuities in any of the considered pre-determined characteristics of municipalities.

	Area	Elevation	Latitude	Longitude	Distance to federal capital	Distance to state capital	Share black 1991	Income per capita 1991
Log (Theoretical FPM)	1,910.85*	-78.86	0.04	-0.15	60.59	-59.05	0.001	-0.002
	(1,161.12)	(74.47)	(0.51)	(0.56)	(55.69)	(53.96)	(0.02)	(0.01)
Municipalities	2,125	2,125	2,125	2,125	2,125	2,125	1,869	1,873

Notes: Estimates from reduced-form regressions of pre-determined municipal characteristics on the log of theoretical FPM transfer. All regressions include a third-order population polynomial and state fixed effects. Robust standard errors in parentheses are clustered at the municipal level. *, **, *** represent significance at the 10%, 5%, and 1% level, respectively. Sample comprises municipalities with population between 6,793 and 50,940 inhabitants with test scores data in 2007 and spending and school inputs data in year 2006 for which all spending variables are positive, test scores data is available, and reported spending on education is at least 25% of total revenue.

Another requirement for the instrument to be exogenous is the absence of "bunching" of observations just below or just above thresholds. This condition can be tested by looking at the discontinuity in the density of the IBGE population estimates at the population thresholds used in the analysis. To test this, I pool the observations for all seven threshold and center them around one normalized threshold with population equal to zero. I follow McCrary (2008) and formally test for density discontinuity by running a kernel regression of the log of density on each side of the normalized threshold. The results displayed in Figure 4.3 indicate that there is no evidence of municipalities sorting in the vicinity of population thresholds.

Figure 4.3: McCrary's density test for discontinuities in the density of population at the thresholds



Notes: Weighted kernel estimation of the log density (by population size), performed separately on either side of pooled FPM revenue-sharing thresholds. The sample is the same as that employed in the regression analysis. Optimal binwidth and binsize as in McCrary (2008).

4.6.2 First stage estimates

Table 4.4 shows the results for the reduced form relationships between log of theoretical FPM transfers and log of spending variables. The objective of this analysis is to check whether additional recourses transferred to municipalities translate into increase in spending, specifically in discretional spending on education. In column (1) of Table 4.4 I report the estimates for equation (1). In column (2) of Table 4.4 I show the results for equation (1), adding in the model specification pre-determined characteristics of municipalities.

As seen, there is strong positive association between theoretical and actual FPM transfers. The point estimates are positive, slightly smaller than one and statistically significant at the 1% level of statistical significance. These results are in line with the findings in earlier research that leveraged Brazilian FPM

transfer data (Brollo et al., 2013; Bastos and Straume, 2016).[84] These results also anticipate that theoretical FPM should be a good shifter of discretional spending on education, since the FPM transfer received by municipalities is an important source of the funds that municipalities allocate to this end.

Table 4.4: First stage regressions of spending variables on theoretical FPM transfers

	(1)	(2)
Log (actual FPM)	0.94	0.93
	(0.02)***	(0.02)***
Log (Discretional spending on education)	0.38	0.42
	(0.11)***	(0.10)***
Log (Total spending)	0.27	0.30
	(0.07)***	(0.06)***
Log (Spending on education)	0.17	0.17
	(0.08)**	(0.08)**
Log (Mandatory spending on education FUNDEF)	0.03	0.00
	(0.11)	(0.11)
Number of municipalities	2,129	1,869
Covariates	No	Yes

Notes: Estimates from reduced-form regressions of spending variables expressed in logs on log of theoretical FPM transfers. All regressions include a third-order population polynomial and state fixed effects. Robust standard errors in parentheses are clustered at the municipal level. *, **, *** represent significance at the 10%, 5%, and 1% level, respectively. Sample comprises municipalities with population between 6,793 and 50,940 inhabitants with test scores data in 2007 and spending and school inputs data in year 2006 for which all spending variables are positive, test scores data is available, and reported spending on education is at least 25% of total revenue.

This intuition is corroborated by the results of estimating the reduced form relationship between log of theoretical FPM and log of discretional spending

[84] The fact that the coefficient estimate is not exactly equal to one can occur because of measurement error or misclassification by TCU of municipalities that are below threshold to be just above threshold (Brollo et al. 2013).

on education. The coefficient estimate is positive and statistically significant at the 1% level. It indicates that a 1% increase in the theoretical FPM transfer is associated with a 0.38%-0.42% increase in discretional spending on education. Overall, this evidence supports the relevance of the theoretical FPM as an instrumental variable for the discretional spending on education.

I also analyze the response of municipalities on changes in theoretical FPM transfers in other expenditure categories related to education. I observe significant shifts in total spending and spending on education induced by changes in theoretical FPM. The effect of log theoretical FPM transfers on log of total spending and log of spending on education is positive and statistically significant at the 1-5% level. The coefficient estimates show that a 1% increase in theoretical transfers implies a 0.27%-0.30% increase in total spending and 0.17% increase in spending on education. As expected, I do not observe any statistically significant relationship between theoretical FPM transfers and mandatory spending on education funded with FUNDEF resources. This result is expected because the funds that municipalities receive from FUNDEF depend on enrollment and should be unrelated to the changes in the theoretical FPM at population thresholds.

4.6.3 Reduced form and instrumental variable estimates for test scores

4.6.3.a Reduced form estimates

Table 4.5 shows the results of estimating the reduced form relationships between log of theoretical FPM transfers and test scores. In column (1) of Table 4.5 I show the results of estimating equation (1) with test scores on the left-hand side. In column (2) I estimate the same model as in column (1) including pre-determined characteristics of municipalities as controls.

The results show a positive and statistically significant relationship between theoretical FPM transfer and student achievement. In the specification without covariates, I find that a 1% increase in the theoretical FPM transfer is associated with a 0.23% increase of a SD in the combined test score, 0.21% increase of a SD in math and 0.20% increase of a SD in language. These results are robust to including controls for pre-determined characteristics of municipalities. All coefficient estimates are statistically significant at the 5%

level except for math without controls, which is statistically significant at the 10%.

Table 4.5: Reduced form estimates for test scores on theoretical FPM transfers

	(1)	(2)
Combined 4th grade	0.23	0.27
	(0.11)**	(0.12)**
Math 4th grade	0.21	0.25
	(0.11)*	(0.11)**
Language 4th grade	0.20	0.24
	(0.09)**	(0.10)**
Number of municipalities	2,129	1,869
Covariates	No	Yes

Notes: Estimates from reduced-form regressions of standardized *Prova Brasil* 4th grade test scores in municipal schools on log of theoretical FPM transfers. All regressions include a third-order population polynomial and state fixed effects. Robust standard errors in parentheses are clustered at the municipal level. *, **, *** represent significance at the 10%, 5%, and 1% level, respectively. Sample comprises municipalities with population between 6,793 and 50,940 inhabitants with test scores data in 2007 and spending and school inputs data in 2006 for which all spending variables are positive, test scores data is available, and reported spending on education is at least 25% of total revenue.

4.6.3.b Instrumental variables estimates

In this section I quantify the relationship between discretional educational spending and student achievement. The IV model specified in equation (2) provides an estimate of the causal effect of discretional spending on student test scores and allows to perform tests of statistical significance of this estimate. The results of estimating the effect of discretional educational spending on student test scores are reported in Table 4.6.

The coefficient estimates in the parsimonious model without covariates are 0.60 for combined score, 0.55 for math and 0.53 for language, statistically significant at the 10%. These estimates indicate that for every 1% increase in discretional spending combined test scores increase by 0.59% of a SD, test scores in math increase by 0.55% and by 0.53% in language. Adding controls leaves these estimates largely unchanged.

Table 4.6: IV estimates for test scores on discretional spending on education

	(1)	(2)
Combined 4th grade	0.60	0.63
	(0.33)*	(0.31)**
Math 4th grade	0.55	0.58
	(0.32)*	(0.30)*
Language 4th grade	0.53	0.56
	(0.29)*	(0.27)**
Number of municipalities	2,129	1,869
Covariates	No	Yes

Notes: IV Estimates for standardized *Prova Brasil* 4th grade test scores in municipal schools from regressions where log of discretional spending on education is instrumented with log of theoretical FPM transfers. All regressions include a third-order population polynomial and state fixed effects. Robust standard errors in parentheses are clustered at the municipal level. *, **, *** represent significance at the 10%, 5%, and 1% level, respectively. Sample comprises municipalities with population between 6,793 and 50,940 inhabitants with test scores data in 2007 and spending and school inputs data in year 2006 for which all spending variables are positive, test scores data is available, and reported spending on education is at least 25% of total revenue.

Given that discretional spending on education represents about 45% of total spending on education, these results mean that a 10% increase in spending on education implies an increase in test scores of about 13% of a SD.[85] This result is similar to the findings in some previous studies. Specifically, Jackson et al. (2018) find that a 10% decrease in spending induced by Great Recession reduced test scores by about 7.8% of a SD, while Lafortune et al. (2018) find that the implied impact of the equivalent increase in spending is between 0.12 and 0.24 SD.

[85] Because discretional spending represents 45% of total spending on education, a 10% increase in spending on education is equivalent to 22% (10%/45%) increase in discretional spending. In turn, a 22% increase in discretional spending on education is expected to increase test scores by 13% (0.60*22%).

To put the cost of the estimated increase in test scores in monetary values, I use the fact that the average municipality in the sample allocates US $3.2 million to spending on education. The estimates suggest that increasing educational spending by US $322,900 (a 10 percent change) would increase test scores by about 13% of a SD. In per student terms, this means that an increase in educational spending per student by US $212[86] would increase test scores by 13% of a SD, or that a per student spending of about US $16 is needed to raise student test score by 1% of a SD.

4.6.4 Exploring mechanisms

Table 4.7 shows the results of estimating reduced form relationships between log of theoretical FPM transfers and some traditional school inputs: class size, proportion of teachers with higher education, and an index measuring the quality of school infrastructure. The objective of this analysis is to investigate the mechanisms through which additional resources transferred to municipalities translate into increase in student achievement. The results in Table 4.7 suggest that no systematic relationship between federal transfers and levels in traditional school inputs is observed in the analyzed data.

A possible explanation to this result can be that the resources available to municipalities from FPM transfers are not spent on traditional school inputs. That is, municipalities have other sources of funding to cover their needs in spending on traditional inputs. An example of such source is a state fund FUNDEF. Gordon and Vegas (2004) study FUNDEF reform and conclude that a part of new resources available to schools through FUNDEF were used to reduce class size and to comply with the federal legislation mandating that teachers should have at least a secondary education degree. Gordon and Vegas (2004) also conclude that it is not clear that the observed reductions in the average class size induced by FUNDEF reform resulted in improvements in student achievement. Another source of funding of essential school inputs is the federal government, which participates in providing resources to schools through federal programs and funds (World Bank, 2002). However, the amounts spent by federal government on fundamental education in Brazil

[86] In the sample average enrollment in municipal schools is 1,519 students. This means that a 10% increase in discretional spending per student is US $ 322,900 / 1,519 = US $212.45.

are significantly smaller than the amounts spent by sub-national governments (see Appendix 4.8 Section A for details).

Besides investment in traditional school inputs, there are many other ways of improving student achievement. For instance, tutoring, guided technology, specialized instruction in small groups, or provision of small budgets for purchase of materials and school supplies have been proved to be effective in increasing test scores in primary education (Cristia, 2017). Thus, it can be the case that spending on education induced by the influx of resources from the FPM transfer is spent by municipalities, at least in part, on these interventions. Unfortunately, school census does not collect the data on non-traditional school inputs and activities. Consequently, I cannot test the hypothesis that non-traditional inputs and alternative interventions are the mechanism through which educational spending induced by the FPM transfers translates into better student outcomes.

Table 4.7: Reduced form estimates for school inputs on theoretical FPM transfers

	(1)	(2)
Class size	0.32	0.26
	(1.15)	(1.24)
Share of teachers with higher education	0.07	0.08
	(0.06)	(0.06)
Infrastructure index	-0.03	-0.05
	(0.0376)	(0.04)
Number of municipalities	2,129	1,869
Covariates	No	Yes

Notes: Estimates from reduced-form regressions of municipal school inputs on log of theoretical FPM transfers. All regressions include a third-order population polynomial and state fixed effects. Robust standard errors in parentheses are clustered at the municipal level. *, **, *** represent significance at the 10%, 5%, and 1% level, respectively. Sample comprises municipalities with population between 6,793 and 50,940 inhabitants with test scores data in 2007 and spending and school inputs data in year 2006 for which all spending variables are positive, test scores data is available, and reported spending on education is at least 25% of total revenue.

4.6.5 Falsification tests

The interpretation of the estimates reported in Section 4.6.3b as causal effects of educational spending on student achievement relies on the validity of the exclusion restriction. That is, the credibility of these estimates depends on the assumption that there is no other mechanism through which test scores of students in municipal schools are affected by changes in levels of FPM transfers at population thresholds other than discretional spending on education. While it is not possible to test this assumption formally, I provide evidence that supports the argument that the exclusion restriction holds in the analysis.

In Table 4.8 I show the results of estimating reduced form relationships between fourth grade state school student test scores and log of the theoretical FPM transfers. I show the results for the combined test scores, math and language separately. I do not find any statistically significant relationship between theoretical FPM transfers and educational outcomes of students in state schools.[87] This is an important result, because FPM transfers are the source of funding that can be allocated to different chapters of municipal expenditure. Consequently, changes in the levels of the theoretical FPM are potentially related to different dimensions, such as health, social protection, public safety, which can eventually affect achievement of students who do not benefit from municipal spending on education. Small and statistically insignificant results suggest that the main channel through which increases in federal transfers affect test scores is through the increases in discretional educational spending. This corroborates the validity of the exclusion restriction and sustains the credibility of the IV estimates.

[87] A concern may arise because the sample of municipalities with state school test scores is a subset of municipalities with municipal school test scores. To address this concern in Appendix 4.8 Table 4.8.C.1 I report the results of estimating Table 4.5 regressions in the sample of municipalities used to produce Table 4.8 results. The size of the estimates in specification without covariates is roughly the same and sometimes lager in specification with covariates. The estimates in specification without covariates lose precision because of larger standard errors, but significance of the estimates in specification with covariates is mostly unchanged.

Table 4.8: Reduced form estimates for test scores on theoretical FPM transfers in state schools

	(1)	(2)
Combined 4th grade	0.06	0.11
	(0.17)	(0.17)
Math 4th grade	0.02	0.05
	(0.16)	(0.16)
Language 4th grade	0.08	0.15
	(0.15)	(0.15)
Number of municipalities	1,091	1,004
Covariates	No	Yes

Notes: Estimates from reduced-form regressions of standardized *Prova Brasil* 4th grade test scores in state schools on the log of theoretical FPM transfers. All regressions include a third-order population polynomial and state fixed effects. Robust standard errors in parentheses are clustered at the municipal level. *, **, *** represent significance at the 10%, 5%, and 1% level, respectively. Sample comprises municipalities with population between 6,793 and 50,940 inhabitants with test scores data in 2007 and spending and school inputs data in year 2006 for which all spending variables are positive, test scores data is available, and reported spending on education is at least 25% of total revenue.

4.6.6 Estimations in previous years

4.6.6.a Transfers and spending

The main results presented in this study are estimated using the transfer and spending data for year 2006. In this section I analyze the reduced form relationships between spending variables and theoretical transfers in years 2002-2005. These results are reported in Table 4.9. Odd-numbered columns show the results for the parsimonious specification without covariates and even-numbered columns show the results for specification with covariates.

Table 4.9: Spending variables and theoretical FPM transfers in years 2002-2005

	(1) 2002	(2) 2002	(3) 2003	(4) 2003	(5) 2004	(6) 2004	(7) 2005	(8) 2005
Log (actual FPM transfer)	0.59	0.56	0.72	0.68	0.82	0.79	0.86	0.84
	(0.03)***	(0.03)***	(0.02)***	(0.02)***	(0.02)***	(0.02)***	(0.02)***	(0.02)***
Log (Discretional spending on education)	0.24	0.33	0.40	0.38	0.52	0.51	0.42	0.37
	(0.13)*	(0.14)**	(0.12)***	(0.12)***	(0.14)***	(0.15)***	(0.13)***	(0.14)***
Log (Total spending)	0.19	0.17	0.26	0.23	0.23	0.20	0.26	0.25
	(0.07)***	(0.07)**	(0.07)***	(0.07)***	(0.08)***	(0.08)**	(0.07)***	(0.08)***
Log (Spending on education)	0.11	0.09	0.18	0.15	0.13	0.13	0.15	0.15
	(0.08)	(0.09)	(0.08**	(0.08)*	(0.09)	(0.09)	(0.09)*	(0.09)*
Log (FUNDEF)	-0.04	-0.17	-0.06	-0.14	-0.02	-0.02	0.09	0.09
	(0.12)	(0.13)	(0.12)	(0.12)	(0.11)	(0.12)	(0.11)	(0.12)
Municipalities	2,035	1,815	2,192	1,961	1,924	1,717	1,977	1,745
Covariates	No	Yes	No	Yes	No	Yes	No	Yes

Notes: Estimates from reduced-form regressions of spending variables expressed in logs on log of theoretical FPM transfers. All regressions include a third-order population polynomial and state fixed effects. Robust standard errors in parentheses are clustered at the municipal level. *, **, *** represent significance at the 10%, 5%, and 1% level, respectively. Sample comprises municipalities with population between 6,793 and 50,940 inhabitants with test scores data in 2007 and spending and school inputs data in years 2002-2005 for which all spending variables are positive, test scores data is available, and reported spending on education is at least 25% of total revenue.

Overall, I observe the same pattern in the relationship between theoretical FPM transfers and spending variables using the data for years 2002-2005 as in the main analysis using the data for 2006. The coefficient estimates for the actual FPM transfers are positive, statistically significant and smaller than one in all years. The size of the coefficient estimates ranges from 0.56-0.58 in 2002 to 0.84-0.86 in 2005 (the estimates in year 2006 reported in Table 4.4 are 0.93-0.94). The increase in the coefficient estimates from 2002 to 2006 might be explained by improvements in the quality of population estimates and the data on the actual FPM transfers.

The coefficient for the endogenous discretional spending on education ranges in size from 0.24 to 0.52. The average value is 0.39, which is similar to the average estimate of 0.40 obtained using 2006 data. All estimates for the discretional spending on education are statistically significant at the conventional level, excluding the estimate obtained in parsimonious specification using 2002 data. The coefficient estimates for the total spending are statistically significant. The size ranges from 0.17 to 0.26, which is slightly smaller than the estimates using 2006 data (0.27-0.30). Spending on education is the only variable in the set of expenditure variables for which I do not observe the 2006 pattern in years 2002-2005 consistently. For almost all years I find coefficient estimates that are not statistically significant at the 5% level, except for a specification without covariates using 2003 data. However, the average size of the coefficient estimate is 0.13, which is slightly smaller than 0.17 obtained using 2006 data. The coefficient estimate for FUNDEF is, as expected, statistically not different from zero in all years of analysis. All in all, this evidence shows that in years preceding 2006 theoretical FPM transfers were significant shifters of the actual transfers and endogenous discretional spending on education.

4.6.6.b Student test scores

Table 4.10 reports the results of the reduced form and IV estimations for student test scores in year 2007 and transfer and spending variables in years 2002-2005. Odd-numbered columns show the results for the parsimonious specification without covariates and even-numbered columns show the results for the specification with covariates. The results reported in Table 4.10 suggest that there is no clear pattern in the relationship between student test

scores, discretional spending on education and the instrument in previous years. While I observe some statistically significant coefficients in years 2003 and 2004, I do not observe any in years 2004-2005, except for language in the model with covariates in year 2004 (statistically significant at the 10%).

Table 4.10: Reduced form and IV estimates for test scores in 2006 and spending in 2002-2005

	(1) 2002	(2) 2002	(3) 2003	(4) 2003	(5) 2004	(6) 2004	(7) 2005	(8) 2005
Combined 4th grade								
Reduced form								
Log (theoretical FPM)	0.20	0.27	0.24	0.30	0.13	0.19	0.07	0.19
	(0.12)*	(0.12)**	(0.11)**	(0.11)***	(0.12)	(0.12)	(0.11)	(0.12)
IV estimation								
Log (Discretional spending on education)	0.84	0.81	0.59	0.78	0.26	0.37	0.16	0.35
	(0.61)	(0.47)*	(0.31)*	(0.38)**	(0.24)	(0.26)	(0.27)	(0.34)
Math 4th grade								
Reduced form								
Log (theoretical FPM)	0.19	0.25	0.20	0.25	0.07	0.13	0.06	0.11
	(0.11)*	(0.12)**	(0.11)*	(0.11)**	(0.12)	(0.12)	(0.11)	(0.12)
IV estimation								
Log (Discretional spending on education)	0.76	0.75	0.49	0.66	0.14	0.25	0.14	0.30
	(0.58)	(0.45)*	(0.29)*	(0.35)*	(0.23)	(0.24)	(0.27)	(0.33)
Language 4th grade								
Reduced form								
Log (theoretical FPM)	0.19	0.24	0.24	0.29	0.17	0.21	0.06	0.12
	(0.10)*	(0.11)**	(0.10)**	(0.10)***	(0.10)	(0.11)*	(0.10)	(0.11)
IV estimation								
Log (Discretional spending on education)	0.76	0.72	0.59	0.76	0.32	0.41	0.15	0.33
	(0.55)	(0.42)*	(0.28)**	(0.34)**	(0.21)	(0.24)*	(0.24)	(0.30)
N	2,035	1,815	2,192	1,961	1,924	1,717	1,977	1,745
Covariates	No	Yes	No	Yes	No	Yes	No	Yes

Notes: Estimates from reduced-form regressions and IV estimates for *Prova Brasil* 4th grade test scores in municipal schools for regressions where log of discretional spending on education is instrumented with log of theoretical FPM transfers. In each column a test score in year 2006 is regressed on spending o transfers in the year shown in the column heading. In IV specification, spending in the year shown in the column heading is instrumented with the theoretical FPM in the same year. All regressions include a third-order population polynomial and state fixed effects. Robust standard errors in parentheses are clustered at the municipal level. *, **, *** represent significance at the 10%, 5%, and 1% level, respectively. Sample comprises municipalities with population between 6,793 and 50,940 inhabitants with test scores data in 2007 and spending and school inputs data in years 2002-2005 for which all spending variables are positive, test scores data is available, and reported spending on education is at least 25% of total revenue.

A possible explanation to these results is the sample variation. Because I was not able to perfectly match student test scores to spending variables in years 2002-2005, the sample of municipalities in estimations varies from year to year. Another reason of sample variation is the natural population growth and administrative decisions of splitting and merging municipalities. About 7% of municipalities in the analytical sample moved from one population bracket to another in year 2006. On average, between 2002 and 2006, about 12% of municipalities in year *t* were in a different population bracket than in year *t-1*. As shown in Table 4.11, the largest number of movements across population thresholds occurred in years 2002 and 2005. The spikes in the number of switching municipalities in these years were followed by a sharp decline in the number of movements.

I explore to what extent these movements across population thresholds could have affected the results by estimating Table 4.10 regressions in the sample of non-switching municipalities that remained in the same population threshold in all years of analysis. These results are displayed in Table 4.12. As seen, the coefficient estimates in both reduced form models and IV specifications are about twice as large for municipalities that never switch population brackets. I also observe statistically significant coefficient estimates in year 2004 in specifications with covariates and in year 2005 for math (statically significant at the 10%). These results suggest that sample composition is not trivial to the analysis. Thus, the evidence obtained in the main analysis may not necessarily be replicated in other years. This is a drawback that puts limitations on the external validity of the results.

In addition to sample variation, selective sorting of student might be an issue driving the results in Table 4.10. Theoretical transfers in year t can be related to discretional spending on education in year *t-1* or *t-2*, or earlier years. This may result into selective migration of students that is not addressed in the study. I leave for future research the exploration of the issues related to dynamic student sorting across municipalities.

Table 4.11: Movements of municipalities across population thresholds

Year	(1) Number of Municipalities	(2) Up	(3) Down	(4) Up - Down	(5) Up + Down	(6) (5)/(1)*100%
2002	2,035	399	179	220	578	28%
2003	2,192	122	49	73	171	8%
2004	1,924	87	16	71	103	5%
2005	1,977	181	42	139	223	11%
2006	2,129	115	27	88	142	7%
Total		904	313	591	1217	
Average		180.80	62.60	118	243.40	12%

Notes: This tables shows the number of movements of municipalities across population thresholds. Column (1) shows the number of municipalities in the analytical sample in the given year. Column (2) shows the number of municipalities in year t which, in respect to year t-1, moved up at least one population threshold. Column (3) shows the number of municipalities in year t which, in respect to year t-1, moved down at least one population threshold. Column (4) shows the difference of the figures reported in columns (2) and (3). Column (5) shows the sum of the figures reported in columns (2) and (3). Column (6) is the ratio of the figure reported in column (5) and column (1).

Table 12: Reduced form and IV estimates for test scores in 2006 and spending in 2002-2005 for the sample of non-switchers

	(1) 2002	(2) 2002	(3) 2003	(4) 2003	(5) 2004	(6) 2004	(7) 2005	(8) 2005
Combined 4th grade								
Reduced form								
Log (theoretical FPM)	0.40	0.57	0.42	0.56	0.27	0.44	0.19	0.30
	(0.18)**	(0.18)***	(0.18)**	(0.18)***	(0.19)	(0.18)**	(0.18)	(0.19)
IV estimation								
Log (Discretional spending on education)	1.20	1.33	0.75	0.93	0.52	0.81	0.29	0.55
	(0.79)	(0.69)*	(0.37)**	(0.39)**	(0.40)	(0.44)*	(0.29)	(0.39)
Math 4th grade								
Reduced form								
Log (theoretical FPM)	0.40	0.56	0.41	0.54	0.26	0.44	0.20	0.32
	(0.17)**	(0.18)***	(0.18)**	(0.18)***	(0.18)	(0.18)**	(0.18)	(0.18)*
IV estimation								
Log (Discretional spending on education)	1.20	1.31	0.72	0.91	0.50	0.81	0.31	0.59
	(0.78)	(0.67)*	(0.37)**	(0.39)**	(0.40)	(0.43)*	(0.29)	(0.39)
Language 4th grade								
Reduced form								
Log (theoretical FPM)	0.33	0.47	0.36	0.47	0.23	0.36	0.14	0.22
	(0.16)**	(0.16)***	(0.15)**	(0.16)***	(0.16)	(0.16)**	(0.16)	(0.16)
IV estimation								
Log (Discretional spending on education)	0.97	1.10	0.64	0.78	0.44	0.65	0.21	0.41

	(0.67)	(0.59)*	(0.32)**	(0.34)**	(0.35)	(0.37)*	(0.25)	(0.34)
N	1,163	1,051	1,256	1,133	1,107	992	1,127	997
Covariates	No	Yes	No	Yes	No	Yes	No	Yes

Notes: Estimates from reduced-form regressions and IV estimates for *Prova Brasil* 4th grade test scores in municipal schools for regressions where log of discretional spending on education is instrumented with log of theoretical FPM transfers. In each column a test score in year 2006 is regressed on spending o transfers in the year shown in the column heading. In the IV specification, spending in the year shown in the column heading is instrumented with the theoretical FPM in the same year. All regressions include a third-order population polynomial and state fixed effects. Robust standard errors in parentheses are clustered at the municipal level. *, **, *** represent significance at the 10%, 5%, and 1% level, respectively. Sample comprises non-switching municipalities that do not change population brackets in years 2002-2006, with population between 6,793 and 50,940 inhabitants, with test scores data in 2007 and spending and school inputs data in years 2002-2005 for which all spending variables are positive, test scores data is available, and reported spending on education is at least 25% of the total revenue.

4.7 Conclusions

In this study I estimate the causal effect of educational spending on student achievement in the short run. To this end I exploit a plausibly exogenous variation in the discretional spending on education in Brazilian municipalities stemming from sharp discontinuities in federal transfers at population thresholds. The results suggest that more federal transfers resources generate increases in educational spending, which translates into better student achievement. Specifically, the results indicate that a 10% increase in educational spending (about $212 per student in the sample) boosts student test scores by 0.13 of a SD.

From a public policy perspective, however, it is interesting to know whether increasing spending on education is a cost-effective way of improving student achievement. To answer this question, I compare the results with the estimates of costs and effects of some primary education interventions reviewed in Cristia (2017). In particular, Cristia (2017) estimates that in Colombia reducing class-size from 25 to 20 students implies an increase in per student cost by 27.1% and generates an increase in test scores by 0.08 of a SD; extending school day from 4 to 7 hours implies an increase in per student cost by 81.2% and increase in student test scores by 0.04 of a SD; and increasing teachers' years of education rises costs by 24% and has no effect on test scores. On the other hand, Cristia (2017) identifies many low-cost interventions for which the average effect is 0.10 of a SD and requires increasing per student cost by 2%.[88] In my analysis I find that increasing spending on education in municipal schools in Brazil by 10% is associated with an increase by a 0.13 of a SD in standardized test score. Given the results obtained in Cristia (2017), it appears that increasing discretional spending on education is a relatively cost-effective way of improving student outcomes, although not the most efficient.

[88] Cristia (2017) estimates costs of the interventions in Colombia in 2014. Particularly, the cost of reducing class was estimated at $281, extending school day length at $842, more years of teacher education at $248. The average cost of low-cost interventions was estimated at $21. I present these costs as a % increase with respect to the average per student spending in primary education in Colombia, which was $1,037 in 2014.

Another question raised in this study is the importance of understanding the mechanisms through which additional funds translate into improvements in student achievement. I investigated this issue by looking at the effects of changes in spending on changes in traditional school inputs, such as class size, teacher level of education and quality of infrastructure. I find no evidence that traditional school inputs mediate the relationship between increase in spending and changes in student achievement. I hypothesize that investments in non-tradition inputs and alternative interventions such as tutoring, small class instruction, funding for materials, guided technology with extra time, which proved to be effective in other settings and countries, can be the mechanisms through which additional resources affect student test scores. Lack of the data on non-traditional inputs and alternative interventions did not allow me to explore this question in the present study. I leave this topic for the future research.

4.8 Appendix

4.8.A Financing of public education in Brazilian municipalities

Primary education expenditure in Brazil is mostly sub-national. In 2001, federal government spent overall R$3.5[89] billion in pre-school and primary education.[90] This is a small amount when compared to over R$30 billion spent by state and municipal governments on primary education (World Bank, 2002). Spending on education at state level is relevant, but it is mostly focused on secondary education, while primary education is largely provided and funded by municipal governments.

Current system of education expenditure in Brazilian municipalities is determined by the reforms to the Constitution of 1988 and the National Law of Education (LDB) approved in 1996. One of the most important reforms in education financing were the laws that established FUNDEF (Constitutional Amendment 14/96 and the Law 9424/96). The FUNDEF mechanisms works

[89] The exchange rate of Brazilian Real to US dollar in 2001 was R$61.409 to UD$1.

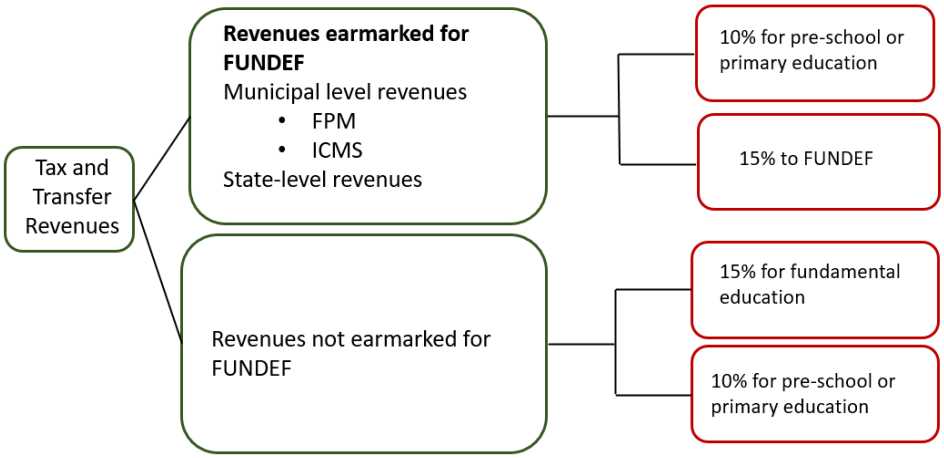
[90] Among other school programs, federal government finances textbooks (National Textbook Program), food (National Program for Nourishment in Schools), transportation (National Program for the Support of School Transportation, as well as the Path to School program), and resources for school maintenance and repairs (Direct Funding for Schools Program).

as follows: FUNDEF collects resources from state and municipal governments in a single fund dedicated exclusively to primary education. Each of the 26 states in Brazil has its own FUNDEF. The FUNDEF for each state receives 15% of designated sources of revenue at state and municipal level. Two sources of revenue are earmarked for FUNDEF at the municipal level, one of them is the FPM transfer.[91] The resources collected by FUNDEF in each state are divided by the number of primary education students enrolled in that state in the previous academic year and distributed to state and local governments according to the number of students enrolled in each system. Federal government tops up FUNDEF funds received by sub-national governments if the resources received are smaller than the preestablished minimum level.

At the same time, the article 212 of the Federal Constitution states that states (including the Federal District) and municipalities must apply at least 25% of the tax revenue (taxes and tax transfers) on the maintenance and development of education. Thus, FUNDEF does not include the entire 25% revenue sources earmarked for FUNDEF that municipalities have to spend on education. Each municipality also needs to spend at least 10% of the same sources of revenue on education. Consequently, after application of the compulsory 15% FUNDEF discount, municipality must spend on education at least 10% of each affected source of revenue, FPM transfer among them. Figure 4.8.A.1 explains the flow of resources that constitute FUNDEF and other expenditures on education.

[91] The second source of revenue tapped by FUNDEF at the municipal level is a tax on goods and services – ICMS - similar to the value added tax.

Figure 4.8.A.1: Diagram of minimum municipal resources for education



Notes: Adopted from World Bank (2002).

It is worth noting that figure 4.8.A.1 shows the minimum requirements for spending on education by local governments. There are no specific rules for the maximum resources that municipalities can allocate to this end.

4.8.B Balance test of time invariant pre-determined characteristics of municipalities in the main sample

Table 4.8.B.1: Balance tests of time-invariant and pre-determined characteristics of municipalities in the main sample

Variables	Area	Elevation	Latitude	Longitude	Distance to federal capital	Distance to state capital	Share black 1991	Income per capita 1991
Log theoretical FPM	1,929.805	-75.811	0.149	-0.216	80.433	-70.816	0.005	-0.002
	(1,218.760)	(72.510)	(0.510)	(0.545)	(55.298)	(52.675)	(0.014)	(0.010)
Observations	2,195	2,195	2,195	2,195	2,195	2,195	1,928	1,932

Notes: Estimates from reduced-form regressions of pre-determined municipal characteristics on the log of theoretical FPM transfer. All regressions include a third-order population polynomial and state fixed effects. Robust standard errors in parentheses are clustered at the municipal level. *, **, *** represent significance at the 10%, 5%, and 1% level, respectively. Estimation sample comprises municipalities with population between 6,793 and 50,940 inhabitants with test scores data in 2007 and spending and school inputs data in year 2006 for which reported spending on education is at least 25% of total revenue.

4.8.C Reduced form results for test scores in the sample of municipalities with state schools

Table 4.8.C.1: Reduced form estimates for test scores in municipal schools in the sample with state schools

	(1)	(2)
Combined 4th grade	0.21	0.33
	(0.15)	(0.16)**
Math 4th grade	0.18	0.28
	(0.14)	(0.14)*
Language 4th grade	0.22	0.34
	(0.16)	(0.16)**
Number of municipalities	1,091	1,004
Covariates	No	Yes

Notes: Estimates from reduced-form regressions of standardized *Prova Brasil* 4th grade test scores in municipal schools on log of theoretical FPM transfers in the analytical sample restricted to municipalities with state schools. All regressions include a third-order population polynomial and state fixed effects. Robust standard errors in parentheses are clustered at the municipal level. *, **, *** represent significance at the 10%, 5%, and 1% level, respectively.

Chapter 5: Conclusions and Policy Implications

This dissertation presents the results of three impact evaluations implemented in developing countries. The first two studies are large randomized controlled trials evaluating early childhood development interventions in Bolivia. The third study is a quasi-experimental evaluation of the education policy in Brazil. All three studies contribute to the growing body of development economics literature by providing rigorous causal evidence on the effectiveness of evaluated interventions. Given the focus of the studies on human capital development, the relevance of their results is also endorsed by the link between human capital development and productivity, which is crucial for sustainable growth in developing countries. The importance of generating convincing evidence on the effectiveness of social programs is particularly relevant in LAC countries facing structural challenges of social expenditure effectiveness and efficiency, which became especially salient in the wake of the economic downturn caused by the COVID-19 pandemic.

The first two studies, leveraging random treatment assignment, provide experimental evidence of the impact of two interventions implemented by the Bolivian Ministry of Health through the "Grow Well to Live Well" (GWLW) program. Both interventions were large randomized controlled trials implemented in disadvantaged communities in Bolivia. The first impact evaluation, presented in Chapter 2, focused on the program components financing improvements in childcare centers. The implemented interventions included provision of materials, development of guidelines and protocols of care, training and support to personnel through specialized facilitators and nutritionists. In addition, some childcare centers also benefited from improvements in physical infrastructure. The second impact evaluation presented in Chapter 3 focused on the home-visiting component of the program, which aimed at improving child development outcomes through weekly home visits following the Reach Up curriculum adapted for Bolivia. During these visits, trained community workers worked with caregivers offering guidance and counseling on parenting and early stimulation of children.

The general objective of both interventions was to generate a positive impact on child development outcomes. However, these outcomes were measured

only in the second study. The first study measured the effects of childcare centers improvements on the intermediate outcomes, specifically, childcare quality and infrastructure indicators. According to the estimations, the program had large and statistically significant effects on childcare quality in treated centers. For the main quality indicator, the results show an increase in quality by two standard deviations or 62% with respect to the control mean. The disaggregation of the overall quality score shows that the increase is driven by improvements in all dimensions of quality. The intervention also improved child-caregiver interactions and centers' infrastructure. The cost-effectiveness analysis shows that moving quality indicators in desired direction by implementing only soft program component is more cost-effective than by implementing infrastructure upgrades in addition to the soft component investments.

From a public policy perspective, these results are telling in several ways. First, they indicate that investments in improving curricular components, training of caregivers, and providing materials could be a cost-effective alternative of improving childcare centers' quality, which is a relevant insight for policy decisions in resource-constrained settings, such as those faced in developing countries. Second, the program results were achieved in deprived communities in childcare centers staffed by community mothers lacking formal childcare training. The fact that such remarkable results in process quality and child-caregiver interaction indicators were achieved show that the cascade training model implemented by the program could be scalable in the current operational context of the country, where municipalities are responsible for hiring personnel in childcare centers. Another relevant point raised in this study is the grim status quo of center-based childcare quality revealed in the indicators for the control group centers. For instance, only 10% percent of the control group centers had their building in good condition and all of them needed repairs. These and other indicators measured for this evaluation show that the business-as-usual of center-based childcare in disadvantaged communities in Bolivia is upsetting and requires immediate attention.

The second experimental evaluation presents the results of a home-visiting intervention on the intermediate outcomes and final child development impacts. The results show that the program improved child development in

several dimensions, including communication, fine motor, cognitive (problem-solving skills), socio-emotional dimension and the overall development score. The results for the intermediate outcomes show that the program most likely achieved impacts on child development through improvements in caregiver-child interaction and better stimulation practices. The endline survey also registered some improvements in health indicators. These results plausibly occurred because home visitors encouraged the use of primary healthcare services during visits and were perceived as medical professionals. All above-mentioned results were registered in the rural areas despite low program take-up. While the take-up was not perfect, rural households that received the intervention did so according to the program protocol. In urban areas the take-up reported in the endline survey was too low to observe any program effects.

An important policy implication arising from the results of this evaluation is that better parenting practices are a pathway to improving child development. This study corroborates the existing evidence that parenting matters. It is known that families shape children's skills through stimulation and daily interactions. Parents are the most important actors shaping children's environment, influencing their choices and lifetime decisions. Strengthening vulnerable families is a viable solution for bringing more opportunities to disadvantaged children. The more families are reached by programs like GWLW, the more children would get a chance to leave poverty and achieve better outcomes in adulthood. Scaling up development programs is always a challenging task, and even more so in developing countries with tight budget constraints and, oftentimes, weak institutional capacity. A more effective program scale-up could be achieved leveraging existing public services infrastructure used by other programs. This would allow to reduce per-child program cost and mitigate some operational challenges, such as hiring and training of program staff or program delivery monitoring. Another alternative could be bringing the program to more families through a different delivery model, for instance, group sessions with parents (parenting programs). Existing evidence suggests that these programs are highly effective and relatively inexpensive. In LAC, parenting programs are still rare, which means that this policy option has a lot of room for expansion to promote skills development in early childhood.

The third study presented in this dissertation uses a quasi-experimental methodology to estimate the causal effects of educational spending on student achievement. An important contribution of this study is that it addresses a relevant policy question – whether more spending on education translates into better student achievement – using a rigorous impact evaluation methodology. This study is an example of a situation when implementing an experimental evaluation is not possible and an alternative quasi-experimental methodology is used to identify causal policy effects. The instrumental variables approach adopted in this study exploits a plausibly exogenous variation in the federal transfers to Brazilian municipalities. The results suggest that increasing educational spending can improve student achievement measured by standardized test scores. No evidence is found that traditional school inputs mediate the relationship between an increase in spending and changes in student achievement. Comparing the program costs with other interventions aiming at improving student achievement indicates that increasing spending on education is a relatively cost-effective way to improve student outcomes, although not the most efficient.

This study contributes to the ongoing policy debate on whether it is possible to achieve more learning by spending fewer resources. There is large evidence that almost all known programs and policies aiming at increasing students learning work. However, the cost of achieving learning results differs largely across programs and policies. While reducing class size or increasing school day is more expensive than tracking (an intervention in which children receive instruction according to their initial level), the implementation of tracking can be challenging. Arguably, this intervention could have unintended consequences such as stigmatization or segregation, ultimately limiting the opportunities of the most disadvantaged students. Improving student outcomes at lower costs requires a thorough understanding of program effects and precise estimates of costs and benefits. Generating this evidence requires a cross-disciplinary approach and coordination between many stakeholders. While challenging, this is probably the most likely way to find the better and less costly options to improve education. Finding public policy solutions to make learning more effective, reduce costs, or both, would help developing countries channel scarce resources in the right directions to improve human capital and foster development.

The studies presented in this dissertation also touched on some practical challenges and limitations of impact evaluations. One of them is the availability of data. Usually, the available administrative data is not sufficient for measuring all program effects. For instance, in the third study presented in this dissertation, it was not possible to find administrative data to measure non-traditional school inputs. In prospective evaluations designed ex-ante, a solution is usually implementing surveys. Surveys, however, are costly in terms of time and money. In fact, data collection often represents a large share in the cost of field experiments. In the case of the first study presented in this dissertation, the measurement of child development outcomes was left for future research. Nonetheless, the feasibility of this follow-up data collection largely dependents on the availability of monetary resources.

Another important aspect is the quality of the data. The data collected for evaluations needs to meet certain quality standards so that the estimated program effects are credible. In the case of the second study presented in this dissertation, a possible reason why the take-up in urban areas was so low is a mismatch in the intervention unit limits used by the program staff and the survey teams. Close coordination between the data collection team and program implementer is required to mitigate this kind of risk and ensure that the data is gathered and appropriately measured. The quality of an impact evaluation critically depends on the quality of collected data. Warranting data quality may be costly, but it is a worthy investment for getting credible and reliable program effect estimates.

Regarding future research, as mentioned earlier, the first study of the effects of center-based childcare did not measure final impacts on child development. The observed large increases in process and structural quality indicators give promising signals for the potential achievement of impacts on child development, which was left for future research. The second study measured the effects of home visits on child development right after the program finished. The follow-up evaluations will be relevant to show whether the program results in rural areas are sustained in the medium and long run. However, the continuity of this research agenda is conditional on the country context and availability of resources. The third study, due to lack of administrative data, did not manage to identify the intermediate outcomes through which additional spending gets to produce impacts on student

achievement. Addressing this gap is important to shed light on the program theory of change and understand the mediating pathways from educational resources to student outcomes.

An important opportunity for future research emerges from the discussion on finding new ways to achieve more with less resources. While the focus of this dissertation is skills development and fostering of human capital, this question applies to any area of social policy and development. Answering this question requires a rigorous evaluation of program effects and precise estimates of costs. Regarding estimation of the program effects, in the areas of studies presented in this dissertation, a large body of literature shows that many known programs and policies work. However, some relevant and popular policies implemented by governments still lack rigorous evidence. While it is not possible to randomize something that happened in the past, these interventions deserve attention and should be considered for generating rigorous evidence using techniques other than experimental evaluation.

Another critical aspect is the program cost. As it was discussed in this dissertation, a program can be highly effective, however, its implementation might require large amounts of resources, making a program a less efficient option in comparison to the alternatives. From the public policy perspective, a relevant parameter is program efficiency, which informs about program costs in addition to effects. Having reliable costs estimates is a necessary condition for informing choice decisions of policy alternatives. Consequently, any research agenda aiming at producing cost estimates or improving cost estimation methods would undoubtedly add value to evidence-based decision-making.

In sum, evaluating the effectiveness and efficiency of existing and prospective programs and policies is the best way to know what works. This would allow to generate valuable inputs for evidence-based policymaking and contribute to achieving better development results.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge. When should you adjust standard errors for clustering? 2017. No. w24003. National Bureau of Economic Research, 2017.

Abdelghafour, Nassima. 2017. “Randomized Controlled Experiments to End Poverty? A Sociotechnical Analysis.” Anthropologie & développement (46-47):235-262.

Achie, Fiona Gedeon. 2019. “Evidence in action: an anthropology of global poverty alleviation efforts.” PhD diss., McGill University.

Al-Maadadi, Fatima, and Atmane Ikhlef. 2015. “What Mothers Know About Child Development and Parenting in Qatar.” The Family Journal: Counseling and Therapy for Couples and Families 23(1): 65–73.

Anderson, Kaitlin, Gema Zamarro, Jennifer Steele, and Trey Miller. 2021. “Comparing Performance of Methods to Deal With Differential Attrition in Randomized Experimental Evaluations.” Evaluation Review 45(1-2): 70-104.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. “Identification of causal effects using instrumental variables.” Journal of the American statistical Association 91(434): 444-455.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. Mostly harmless econometrics. Princeton university press.

Araujo, M. C., M. Rubio-Codina, and N. Schady. 2021. “70 to 700 to 70,000: Lessons from the Jamaica Experiment.” Working Paper 1230. Inter-American Development Bank, Washington, DC.

Araujo, Maria Caridad, Lazarte, Fabiola., Rubio-Codina, Marta, and Norbert Shady. 2016. Home visiting at scale: the impact evaluation of Cuna Mas. Power Point presentation at the “The Early Years: Child Well-Being and the Role of Public Policy” Conference. The British Academy, London, England.

Araujo, Maria Caridad, Florencia Lopez Boo, Rafael Novella, Sara Schodt, and Romina Tomé. 2015. The quality of Centros Infantiles del Buen Vivir in Ecuador. Policy Brief IDB-PB-248, Inter-American Development Bank, Washington DC.

Arnett, Jeffrey. 1989. "Caregivers in Day-Care Centers: Does Training Matter?" Journal of Applied Developmental Psychology 10(4): 541–52.

Attanasio, Orazio, Sarah Cattan, and Costas Meghir. 2021. Early Childhood Development, Human Capital and Poverty. No. w29362. National Bureau of Economic Research.

Attanasio, Orazio P., Camila Fernández, Emla OA Fitzsimons, Sally M. Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina. 2014. "Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial." BMJ(349): g5785.

Attanasio, Orazio, Costas Meghir, and Emily Nix. 2020. "Human capital development and parental investment in India." The Review of Economic Studies 87(6): 2511-2541.

Baker, Michael, Jonathan Gruber, and Kevin Milligan. 2019. "The long-run impacts of a universal child care program." American Economic Journal: Economic Policy 11(3): 1-26.

Baker, Michael, Jonathan Gruber, and Kevin Milligan. 2015. Non-cognitive deficits and young adult outcomes: The long-run impacts of a universal child care program. National Bureau of Economic Research Working Paper Series No. w21571.

Bancalari, Antonella, Gastón Gertner, and Sebastian Martinez. 2016. ¿Quién Se Conecta? Estimación de La Propensión a La Conexión Al Alcantarillado En Áreas Peri-Urbanas de Bolivia. Technical Note No. 1075. Inter-American Development Bank, Washington, DC.

Banerjee, Abhijit, and Esther Duflo. 2011. Poor economics: A radical rethinking of the way to fight global poverty. Public Affairs.

Bastos, Paulo, and Odd Rune Straume. 2016. “Preschool education in Brazil: does public supply crowd out private enrollment? ” World Development 78: 496-510.

Bedregal, Paula, Gaston Gertner, Julia Johannsen, and Sebastian Martinez. 2016. Centros Infantiles En Bolivia: Atención, Infraestructura y Calidad de Servicios de Desarrollo Infantil. Technical Note No. 1187. Inter-American Development Bank, Washington, DC.

Behrman, Jere R, Yingmei Cheng, and Petra E Todd. 2004. “Evaluating Preschool Programs When Length of Exposure to the Program Varies: A Nonparametric Approach.” Review of Economics and Statistics 86(February): 108–32.

Berlinski, Samuel, and Norbert Schady. 2015. “The Early Years: Child Well-being and the Role of Public Policy,” IDB Publications (Books), Inter-American Development Bank, number 7259.

Bernal, Raquel, Orazio Attanasio, Ximena Peña, and Marcos Vera-Hernández. 2019. “The effects of the transition from home-based childcare to childcare centers on children's health and development in Colombia.” Early childhood research quarterly, 47:418-431.

Bernal, Raquel. 2015. “The impact of a vocational education program for childcare providers on children's well-being.” Economics of Education Review, 48:165-183.

Bernal, Raquel, and Camila Fernández. 2013. “Subsidized childcare and child development in Colombia: effects of Hogares Comunitarios de Bienestar as a function of timing and length of exposure.” Social Science & Medicine 97: 241-249.

Bos, Johannes, Akib Khan, Saravana Ravindran, and Abu Shonchoy. 2021 “Early Childhood Human Capital Formation at Scale.” Available at SSRN 3906697.

Bouguen, Adrien, Deon Filmer, Karen Macours, and Sophie Naudeau. 2018. “Preschool and parental response in a second best world evidence from a school construction experiment.” Journal of human Resources 53(2): 474-512.

Bouguen, Adrien, Deon Filmer, Karen Macours, and Sophie Naudeau. 2013. Impact evaluation of three types of early childhood development interventions in Cambodia. The World Bank.

Brollo, Fernanda, Tommaso Nannicini, Roberto Perotti, and Guido Tabellini. 2013. “The political resource curse. ” American Economic Review 103(5): 1759-96.

Busso, Matías, Julián Cristia, Diana Hincapié, Julián Messina, and Laura Ripani, eds. 2017. Learning better: Public policy for skills development. Inter-American Development Bank.

Caldwell, B., and R. Bradley. 2001. HOME inventory and administration manual (3rd edition), University of Arkansas for Medical Sciences and University of Arkansas at Little Rock.

Campbell, Frances A., Barbara H. Wasik, Elizabeth Pungello, Margaret Burchinal, Oscar Barbarin, Kirsten Kainz, Joseph J. Sparling, and Craig T. Ramey. 2008. “Young adult outcomes of the Abecedarian and CARE early childhood educational interventions.” Early Childhood Research Quarterly 23(4): 452-466.

Campbell, Frances, Gabriella Conti, James J. Heckman, Seong Hyeok Moon, Rodrigo Pinto, Elizabeth Pungello, and Yi Pan. 2014. “Early childhood investments substantially boost adult health.” Science 343(6178): 1478-1485.

Card, David, and A. Abigail Payne. 2002 “School finance reform, the distribution of school spending, and the distribution of student test scores. ” Journal of public economics 83(1): 49-82.

Carneiro, Pedro, Oswald Koussihouèdé, Nathalie Lahire, Costas Meghir, and Corina Mommaerts. 2015. Decentralizing education resources: school grants in Senegal. No. w21063. National Bureau of Economic Research.

Celhay, Pablo, Sebastian Martinez, and Cecilia Vidal. 2018. Socioeconomic Gaps in Child Development: Evidence from a National Health and Nutrition Survey in Bolivia. IDB Working Paper No. 00949.

Coleman, James S., Ernest Campbell, Carol Hobson, James McPartland, Alexander Mood, Frederick Weinfeld, and Robert York. 1966. “The coleman report.” Equality of Educational Opportunity: 1-32.

Colwell, Nicole, Rachel A. Gordon, Ken Fujimoto, Robert Kaestner, and Sanders Korenman. 2013. “New Evidence on the Validity of the Arnett Caregiver Interaction Scale: Results from the Early Childhood Longitudinal Study-Birth Cohort.” Early Childhood Research Quarterly 28(2): 218–33.

Cristia, Julian. Improving Skills in Childhood: A Cost-Effective Approach (pp. 145-171). 2017. In Ed. Busso, Matías, Julián Cristia, Diana Hincapie, Julián Messina, and Laura Ripani. “Learning Better: Public Policy for Skills Development. ” Inter-American Development Bank.

Cunha, Flávio, Irma Elo, and Jennifer Culhane. 2013. Eliciting maternal expectations about the technology of cognitive skill formation. No. w19144. National Bureau of Economic Research.

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik Muralidharan, and Venkatesh Sundararaman. 2013. “School inputs, household substitution, and test scores.” American Economic Journal: Applied Economics 5(2): 29-57.

Del Bono, Emilia, Marco Francesconi, Yvonne Kelly, and Amanda Sacker. 2016. “Early maternal time investment and early child outcomes.” The Economic Journal 126(596): F96-F135.

Donovan, Kevin P. 2018. “The rise of the randomistas: on the experimental turn in international aid.” Economy and Society 47(1):27-58.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007.“Chapter 6 Using Randomization in Development Economics Research: A Toolkit.” In Handbook of Development Economics. Volume 4, , ed. T. Paul Schultz and John A. Strauss, 3895–3962. Elsevier.

Dragoset, Lisa, Jaime Thomas, Mariesa Herrmann, John Deke, Susanne James-Burdumy, Cheryl Graczewski, Andrea Boyle, Rachel Upton, Courtney Tanenbaum, and Jessica Giffin. 2017. “School Improvement Grants: Implementation and Effectiveness. NCEE 2017-4013.” National Center for Education Evaluation and Regional Assistance.

Engle, Patrice L., Lia CH Fernald, Harold Alderman, Jere Behrman, Chloe O'Gara, Aisha Yousafzai, Meena Cabral de Mello et al. 2011. “Strategies for reducing inequalities and improving developmental outcomes for young children in low-income and middle-income countries.” The Lancet 378(9799): 1339-1353.

Egert, Franziska, Verena Dederer, and Ruben G. Fukkink. 2020. “The impact of in-service professional development on the quality of teacher-child interactions in early education and care: A meta-analysis.” Educational Research Review 29: 100309.

Fernald, Lia CH, Patricia Kariger, Melissa Hidrobo, and Paul J. Gertler. 2012. “Socioeconomic Gradients in Child Development in Very Young Children: Evidence from India, Indonesia, Peru, and Senegal.” Proceedings of the National Academy of Sciences 109 (Supplement_2): 17273–80.

Fiene, Richard J. 1984. Child Development Program Evaluation Scale COFAS. Washington DC: Children’s Monitoring Consortium.

Fiorini, Mario, and Michael P. Keane. 2014. “How the allocation of children’s time affects cognitive and noncognitive development.” Journal of Labor Economics 32(4): 787-836.

Fort, Margherita, Andrea Ichino, and Giulio Zanella. 2020. “Cognitive and noncognitive costs of day care at age 0 - 2 for children in advantaged families.” Journal of Political Economy 128(1): 158-205.

Francesconi, Marco, and James J. Heckman. 2016. “Child development and parental investment: Introduction.” The Economic Journal 126(596): F1-F27.

Frongillo, A. Edward, Sara M. Sywulka, and Patricia Kariger. 2003. “UNICEF Psychosocial Care Indicators Project. Final report to UNICEF.” Mimeo, Cornell University.

Gamse, Beth C., Robin Tepper Jacob, Megan Horst, Beth Boulay, and Fatih Unlu. 2008. “Reading First Impact Study. Final Report. NCEE 2009-4038.” National Center for Education Evaluation and Regional Assistance.

García, Jorge Luis, Heckman, James J., and Stefano Mosso. 2021. The Lasting Effects of Early Childhood Education on Promoting the Sills and Social Mobility of Disadvantaged African Americans. No. w 29057. National Bureau of Economic Research.

Gertner, Gaston, Julia Johannsen, and Sebastian Martinez. 2016. Perfil de desarrollo infantil temprano en la población elegible para visitas domiciliarias en Bolivia. Inter-American Development Bank Technical Note 1142, Washington D.C.

Gertler, Paul, James J. Heckman, Rodrigo Pinto, Susan M. Chang, Sally Grantham-McGregor, Christel Vermeersch, Susan Walker, and Amika Wright. 2021. Effect of the Jamaica Early Childhood Stimulation Intervention on Labor Market Outcomes at Age 31. No. w29292. National Bureau of Economic Research.

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel MJ Vermeersch. 2016. Impact evaluation in practice. World Bank Publications.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M. Chang, and Sally Grantham-McGregor. 2014. "Labor market returns to an early childhood stimulation intervention in Jamaica." Science 344(6187): 998-1001.

Gibbons, Stephen, Sandra McNally, and Martina Viarengo. 2017. "Does additional spending help urban schools? An evaluation using boundary discontinuities." Journal of the European Economic Association 16(5): 1618-1668.

Gigliotti, Philip, and Lucy C. Sorensen. 2018. "Educational resources and student achievement: Evidence from the Save Harmless provision in New York State." Economics of Education Review 66: 167-182.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. "Many children left behind? Textbooks and test scores in Kenya." American Economic Journal: Applied Economics 1(1): 112-35.

Gordon, Nora, and Emiliana Vegas. 2004. "Education Finance Equalization, Spending, Teacher Quality and Student Outcomes: The Case of Brazil's FUNDEF. Education Sector." Human Development Department, Latin America and the Caribbean Region, World Bank, Washington, DC.

Grantham-McGregor, Sally M., Christine A. Powell, Susan P. Walker, and John H. Himes. 1991. "Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican Study." The Lancet 338(8758): 1-5.

Grantham-McGregor, Sally, and Susan Walker. 2015. "The Jamaican early childhood home visiting intervention." Kingston: Bernard van Leer Foundation.

Grantham-McGregor, Sally, and Joanne A. Smith. 2016. “Extending the Jamaican early childhood development intervention.” Journal of Applied Research on Children: Informing Policy for Children at Risk 7(2), Article 4.

Gupta, Nabanita Datta, and Marianne Simonsen. 2010. “Non-cognitive child outcomes and universal high quality child care.” Journal of Public Economics 94(1-2): 30-43.

Guryan, Jonathan. 2001. Does money matter? Regression-discontinuity estimates from education finance reform in Massachusetts. No. w8269. National Bureau of Economic Research.

Haddad, Mônica A., Ricardo Freguglia, and Cláudia Gomes. 2017. “Public Spending and Quality of Education in Brazil. ” The Journal of Development Studies 53(10): 1679-1696.

Hamadani, Jena D., Syed N. Huda, Fahmida Khatun, and Sally M. Grantham-McGregor. 2006. “Psychosocial stimulation improves the development of undernourished children in rural Bangladesh.” The Journal of nutrition 136(10): 2645-2652.

Hamadani, Jena D., Fahmida Tofail, Afroza Hilaly, Syed N. Huda, Patrice Engle, and Sally M. Grantham-McGregor. 2010. “Use of family care indicators and their relationship with child development in Bangladesh.” Journal of health, population, and nutrition 28(1): 23.

Hanushek, Eric A. 2003. “The failure of input-based schooling policies. ” The economic journal 113(485): F64-F98.

Hanushek, Eric A. 2006. “School Resources.” In Handbook of the Economics of Education, Vol. 2, edited by E. Hanushek and F. Welch, 865–908. Amsterdam: North-Holland.

Harding, David J., Lisa Sanbonmatsu, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, Matthew Sciandra, and Jens Ludwig. 2021. “Evaluating Contradictory Experimental and

Nonexperimental Estimates of Neighborhood Effects on Economic Outcomes for Adults." Housing Policy Debate 1-34.

Harms, Thelma, Debby Cryer, and Clifford Richard M. 2006. Infant/Toddler Environment Rating Scale, Revised Edition - ITERSR, Teachers College Press.

Heckman, James J., Lance Lochner, and Christopher Taber. 1998. "Explaining Rising Wage Inequality: Explorations With A Dynamic General Equilibrium Model Of Labor Earnings With Heterogeneous Agents," Review of Economic Dynamics,v1(1,Jan), 1-58.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2010. A new cost-benefit and rate of return analysis for the Perry Preschool Program: A summary. National Bureau of Economic Research Working Paper Series No. w16180.

Herbst, Chris M., and Erdal Tekin. 2010. "Child care subsidies and child development." Economics of Education review 29(4): 618-638.

Hoddinott, John, John A. Maluccio, Jere R. Behrman, Rafael Flores, and Reynaldo Martorell. 2008. "Effect of a nutrition intervention during early childhood on economic productivity in Guatemalan adults." The lancet 371(9610): 411-416.

Hyman, Joshua. 2017. "Does money matter in the long run? Effects of school spending on educational attainment." American Economic Journal: Economic Policy 9(4): 256-80.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." Econometrica 62(2): 467-475.

Imbens, Guido W., and Donald B. Rubin. 2015. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

Izquierdo, Alejandro, Carola Pessino, and Guillermo Vuletin, eds. 2018. Better spending for better lives: how Latin America and the Caribbean can do more with less. Vol. 10. Inter-American Development Bank.

Jackson, C. Kirabo, Rucker C. Johnson, and Claudia Persico. 2016. "The effects of school spending on educational and economic outcomes: Evidence from school finance reforms." The Quarterly Journal of Economics 131(1): 157-218.

Jackson, C. Kirabo, Cora Wigger, and Heyu Xiong. 2018. Do School Spending Cuts Matter? Evidence from the Great Recession. No. w24203. National Bureau of Economic Research.

Johannsen, Julia, Sebastian Martinez, Cecilia Vidal, and Anastasiya Yarygina. 2019a. Evaluación de impacto del programa de desarrollo infantil temprano Crecer Bien para Vivir Bien en Bolivia: modalidad centros infantiles. Inter-American Development Bank Technical Note 1792, Washington D.C.

Johannsen, Julia, Sebastian Martinez, Cecilia Vidal, and Anastasiya Yarygina. 2019b. Evaluación de impacto del programa de desarrollo infantil temprano Crecer Bien para Vivir Bien en Bolivia: modalidad visitas domiciliarias. Inter-American Development Bank Technical Note 1790, Washington D.C.

Johnson, Rucker C., and C. Kirabo Jackson. 2017. Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. No. w23489. National Bureau of Economic Research.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman. 2001. "Moving to opportunity in Boston: Early results of a randomized mobility experiment." The Quarterly Journal of Economics 116(2): 607-654.

Kerstjens, Jorien M., Arend F. Bos, Elisabeth MJ ten Vergert, Gea de Meer, Phillipa R. Butcher, and Sijmen A. Reijneveld. 2009. "Support for the global feasibility of the Ages and Stages Questionnaire as developmental screener." Early human development 85(7): 443-447.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental analysis of neighborhood effects." Econometrica 75(1): 83-119.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018. "School finance reform and the distribution of student achievement." American Economic Journal: Applied Economics 10(2): 1-26.

LaLonde, Robert J. 1986. "Evaluating the econometric evaluations of training programs with experimental data." The American economic review (1986): 604-620.

Leroy, Jef L., Paola Gadsden, and Maite Guijarro. 2012. "The impact of daycare programmes on child health, nutrition and development in developing countries: a systematic review." Journal of development effectiveness 4(3): 472-496.

Leuven, Edwin, Mikael Lindahl, Hessel Oosterbeek, and Dinand Webbink. 2007. "The effect of extra funding for disadvantaged pupils on achievement." The Review of Economics and Statistics 89(4): 721-736.

Litschig, Stephan. 2012. "Are rules-based government programs shielded from special-interest politics? Evidence from revenue-sharing transfers in Brazil." Journal of public Economics 96(11-12): 1047-1060.

Litschig, Stephan, and Kevin M. Morrison. 2013. "The impact of intergovernmental transfers on education outcomes and poverty reduction." American Economic Journal: Applied Economics 5(4): 206-40.

Lopez Boo, Florencia, María Caridad Araujo, and Romina Tomé. 2016. How is Child Care Quality Measured?: A toolkit. Inter-American Development Bank.

Lopez Boo, Florencia, Marta Dormal, and Ann Weber. 2019. "Validity of four measures of child care quality in a national sample of centers in Ecuador." PloS one 14(2): e0209987.

Loughran, David S., Ashlesha Datar, and M. Rebecca Kilburn. 2008. “The Response of Household Parental Investment to Child Endowments.” Review of Economics of the Household 6(3): 223-42.

Lozoff, Betsy, Elias Jimenez, and Julia B. Smith. 2006. “Double burden of iron deficiency in infancy and low socioeconomic status: a longitudinal analysis of cognitive test scores to age 19 years.” Archives of pediatrics & adolescent medicine 160(11): 1108-1113.

Lucas, Robert E. 2015. “Reflections of new growth theory, human capital and growth.” In American Economic Review, Papers and Proceedings, 105:85-88.

MacPhee, David. 1981. Manual for the knowledge of infant development inventory. University of North Carolina; Unpublished manuscript.

Martinez, Sebastian, Julia Johannsen, Gaston Gertner, Jorge Franco, Ana B. Perez Exposito, Rosario M. Bartolini, Irma Condori et al. 2018. “Effects of a home-based participatory play intervention on infant and young child nutrition: a randomised evaluation among low-income households in El Alto, Bolivia.” BMJ global health 3(3): e000687.

Martorell, Paco, Kevin Stange, and Isaac McFarlin Jr. 2016. “Investing in schools: capital spending, facility conditions, and student achievement.” Journal of Public Economics 140: 13-29.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2018. Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. No. w24876. National Bureau of Economic Research.

McCrary, Justin. 2008. “Manipulation of the running variable in the regression discontinuity design: A density test.” Journal of econometrics 142(2): 698-714.

Miller, Corbin L. 2017. The Effect of Education Spending on Student Achievement: Evidence from Property Values and School Finance Rules. Unpublished paper.

Nahar, Baitun, M. I. Hossain, J. D. Hamadani, T. Ahmed, S. N. Huda, S. M. Grantham-McGregor, and L. A. Persson. 2012. "Effects of a community-based approach of food and psychosocial stimulation on growth and development of severely malnourished children in Bangladesh: a randomised trial." European Journal of Clinical Nutrition 66(6): 701.

Naudeau, Sophie, Sebastian Martinez, Patrick Premand, and Deon Filmer. 2011. Cognitive Development among Young Children in Low-Income Countries. In H. Alderman (Ed.), No Small Matter: The impact of poverty, shocks, and human capital investments in early childhood development, (pp. 9–50).

Neidell, Matthew J. 2000. Early parental time investments in children's human capital development: effects of time in the first year on cognitive and non-cognitive outcomes. UCLA Department of Economics Working Paper 806.

Noboa-Hidalgo, Grace E., and Sergio S. Urzua. 2012. "The effects of participation in public child care centers: Evidence from Chile." Journal of Human Capital 6(1): 1-34.

OECD. 2021a. Social spending (indicator). doi: 10.1787/7497563b-en (Accessed on December 23, 2021)

OECD. 2021b. Official development assistance (ODA) from members of the OECD's Development Assistance Committee (DAC). https://www.oecd.org/dac/financing-sustainable-development/development-finance-data/. (Accessed on December 23, 2021).

OECD. Education at a glance 2018: OECD indicators. OECD Publishing, Paris, France, 2018.

Özler, Berk, Lia CH Fernald, Patricia Kariger, Christin McConnell, Michelle Neuman, and Eduardo Fraga. 2018. “Combining pre-school teacher training with parenting education: A cluster-randomized controlled trial.” Journal of Development Economics 133: 448-467.

Papke, Leslie E. 2005. “The effects of spending on test pass rates: evidence from Michigan. ” Journal of Public Economics 89(5-6): 821-839.

Pradhan, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Armida Alishjabana, Arya Gaduh, and Rima Prama Artha. 2011. “Improving education quality through enhancing community participation: Results from a randomized field experiment.” In in Indonesia’, Mimeo, VU University Amsterdam.

Price, Joseph. 2010. The effect of parental time investments: Evidence from natural within-family variation. Unpublished manuscript, Brigham Young University.

Raudenbush, Stephen W., Sean F. Reardon, and Takako Nomi. 2012. “Statistical analysis for multisite trials using instrumental variables with random coefficients.” Journal of research on Educational Effectiveness 5(3): 303-332.

Rosero, José, and Hessel Oosterbeek. 2011. “Trade-offs between different early childhood interventions: Evidence from Ecuador.” (Tinbergen Institute Discussion Paper No. 11- 102/3). Amsterdam.

Rossi, Peter H., Mark W. Lipsey, and Gary T. Henry. 2018. Evaluation: A systematic approach. Sage publications.

Rubio-Codina, Marta, Orazio Attanasio, Costas Meghir, Natalia Varela, and Sally Grantham-McGregor. 2015. “The Socio-Economic Gradient of Child Development Children 6-42 Months in Bogota.” Journal of Human Resources 50(2): 464-83.

Rubio-Codina, Marta, Orazio Attanasio, and Sally Grantham-McGregor. 2016. “Mediating pathways in the socio-economic gradient of child

development: Evidence from children 6 - 42 months in Bogota." International journal of behavioral development 40(6): 483-491.

Ruhm, Christopher, and Jane Waldfogel. 2012. "Long-term effects of early childhood care and education." Nordic Economic Policy Review 1(1): 23-51.

Rutstein, Shea O., and Kiersten Johnson. 2004. "The DHS wealth index. DHS comparative reports no. 6", 1-71.

Saretsky, Gary.1972. "The OEO PC experiment and the John Henry effect." The Phi Delta Kappan 53(9): 579-581.

Schady, Norbert, Jere Behrman, Maria Caridad Araujo, Rodrigo Azuero, Raquel Bernal, David Bravo, Florencia Lopez Boo, …, and Renos Vakis. 2015. "Wealth gradients in early childhood cognitive development in five Latin American countries." Journal of Human Resources, 50(2): 446–463.

Sedgwick, Philip, and Nan Greenwood. 2015. "Understanding the Hawthorne effect." Bmj 351.

Squires, Jane, Diane D. Bricker, and E. Twombly. 2009. Ages & stages questionnaires. Baltimore: Paul H. Brookes, 2009.

Teele, Dawn Langan, ed. 2014. Field experiments and their critics: essays on the uses and abuses of experimentation in the social sciences. Yale University Press.

Vegas, Emiliana, and Chelsea Coffin. 2015. "When education expenditure matters: An empirical analysis of recent international data." Comparative Education Review 59(2): 289-304.

Walker, Susan P., Susan M. Chang, Marcos Vera-Hernández, and Sally Grantham-McGregor. 2011. "Early childhood stimulation benefits adult competence and reduces violent behavior." Pediatrics 127(5): 849-857.

Walker, Susan P., Susan M. Chang, Novie Younger, and Sally M. Grantham-McGregor. 2010. "The effect of psychosocial stimulation on cognition and

behaviour at 6 years in a cohort of term, low-birthweight Jamaican children." Developmental Medicine & Child Neurology 52(7): e148-e154.

Walker, Susan P., Sally M. Grantham-McGregor, Christine A. Powell, and Susan M. Chang. 2000. "Effects of growth restriction in early childhood on growth, IQ, and cognition at age 11 to 12 years and the benefits of nutritional supplementation and psychosocial stimulation." The Journal of pediatrics 137(1): 36-41.

Walker, Susan P., Christine Powell, Susan M. Chang, Helen Baker-Henningham, Sally Grantham-McGregor, Marcos Vera-Hernandez, and Florencia López-Boo. 2015. Delivering parenting interventions through health services in the Caribbean: Impact, acceptability and costs. No. IDB-WP-642. IDB Working Paper Series.

World Bank. 2002. Municipal Education. Resources, incentives, and results. Volume II: Research Report. Report No 244413-BR. The World Bank, December 20, 2002.

Yoshikawa, Hirokazu, Diana Leyva, Catherine E. Snow, Ernesto Treviño, M. Barata, Christina Weiland, Celia J. Gomez et al. 20015. "Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes." Developmental psychology, 51(3):309.