DRAFT AS OF JULY 2021

A Closer Look at Reading Comprehension: Experimental Evidence from Guatemala

Daniel Rodriguez-Segura
University of Virginia – School of Education and Human Development

Available for comment and discussion only
Please do not cite or quote without author permission

Abstract: In spite of the relatively high literacy rates around the world, functional illiteracy is still rampant. Reading comprehension, or functional literacy, is a key foundational skill to progress into higher educational levels, and to later on reap the multifaceted benefits of literacy. However, the current literature in developing countries has focused more on emergent literacy skills than on reading comprehension and the developmental processes through which it improves. This paper provides two key contributions to the literature on reading comprehension in developing countries. First, I provide experimental evidence that an intervention in public primary schools in Tecpán, Guatemala, which provides teacher training, coaching on the delivery of a new and evidence-based instructional approach, and high-quality reading materials, is highly effective at improving reading comprehension levels, particularly among first graders. Secondly, I leverage the early literacy theory of the "Simple View of Reading", to empirically test hypotheses about the developmental processes through which reading comprehension improved in this context. In all, these results have important implications for the targeting of reading interventions in developing countries, as they highlight that reading comprehension is a multi-layered developmental process which requires special attention on each student's learning gaps in the building blocks of literacy.

Keywords: early literacy, reading comprehension,
Simple View of Reading, education in developing countries

Author's note: The author is grateful to Asociación COED, Rony Mejía, Joe Berninger, Katie Dawson, Ben Kelsey, and everyone else who worked on the implementation, collection, and facilitation of the data for this project. The author would also like to thank Beth Schueler, Susan Thacker-Gwaltney, Vivian Wong, Caroline Whitcomb, Isaac Mbiti, Andy de Barros, and Walter Herring for their helpful comments and encouragement. The author received IRB approval the University of Virginia, protocol number 3864. This trial was pre-registered at the AEA RCT (number AEARCTR-0006366) after the trial was completed but before analysis of the data. The views expressed herein are those of the authors and do not necessarily reflect the views of Asociación COED.

I. Introduction

Children in the developing world are enrolling in school at historically high levels, and yet their foundational literacy skills are remarkably weak. According to the World Bank, youth literacy rates[1] worldwide have increased from 80% in 1980, to 92% in 2019, painting a relatively positive picture about the state of literacy around the world. However, "literacy" as a metric of reading capacity tends to be very superficial, heterogeneously measured across countries and time, and often self-reported (Ortiz-Ospina and Belketian, 2018). Literacy rates as typically reported do not necessarily reflect the general population's ability to comprehend a text, and to take full advantage of the benefits of literacy and education more broadly (Evans and Hares, 2021). Instead, the World Bank recently estimated that over 1 of every 2 children in developing countries experience "learning poverty", or the inability to read and comprehend an age-appropriate text by the end of primary (Azevedo et al., 2021). This figure is likely to be exacerbated by the COVID-19 crisis, and it is estimated to come closer to 2 of every 3 children (Azevedo, 2020). Empirically, these assessment-based estimates of learning poverty come into stark contrast with more typical and superficial measures of literacy, and many countries are indeed overestimating the share of their population that is *functionally* literate. For instance, Guatemala, the country where this study takes place, estimates that 6% of its youth population and 18% of its adult population is illiterate, but 67% of its children are reported to experience learning poverty[2]. In Sub-Saharan Africa, the contrast is even starker: the youth illiteracy rate is reported to be 24%, but 87% of all children experience learning poverty.

The high number of children who are functionally illiterate in the developing world is not only a worrying figure for the state of literacy in these contexts, but also more broadly for the state of learning in general. The generally low levels of literacy achievement in developing countries are also mirrored in low levels of achievement in other subjects like math. For instance, 50% of all third graders in Uganda, and 69% of third graders in India cannot perform simple subtraction exercises (Uwezo, 2019; and ASER, 2020). Indeed, before children *read to learn* in other subjects, they must be able to *learn to read*, where improved reading comprehension is the key skill to transition between these two processes (Chall, 1996). In a broader sense, "learning poverty" can be understood through the traditional lens of "poverty traps" (such as those explored in Kraay and McKenzie, 2014), where "poverty begets poverty", and individuals need to get over a threshold (say, a minimum reading comprehension level in the case of learning poverty) so they can get on a virtuous cycle of learning. As such, pushing literacy outcomes from the current state of functional illiteracy

[1] World Bank Development Indicators: Literacy rate, youth total (% of people ages 15-24).

[2] World Bank Development Indicators: Literacy rate, adult total (% of people age 15 and above).

to higher levels of reading comprehension is a crucial first step to maximize both learning across the board while children are in school, and other benefits to literacy post-schooling.

In spite of the growing awareness regarding the crucial role that literacy skills play during a child's schooling experience, especially in contexts where illiteracy and learning poverty are the modal experiences, the development and economics literature has engaged little with the individual components of literacy, and the micro-level, developmental processes through which reading comprehension evolves. Even well-known studies that report effects on students' language outcomes (see, for example, Banerjee et al., 2016; Duflo et al., 2011; Glewwe et al., 2009; Mbiti et al., 2019; Muralidharan et al., 2019) tend to report outcomes either in terms of an aggregate standardized test, or individual skills with a "linear" conception of literacy (e.g., letter, word, paragraph, etc.). This stands in contrast with reporting more disaggregated and theory-driven literacy sub-skills such as those measured by instruments like the popular Early Grade Reading Assessment (EGRA). Among studies that report on specific sub-skills following the "linear" conception of literacy, the focus of these assessments is in fact on emergent skills (e.g., letter names and sounds, word decoding, or perhaps even fluency), with reading comprehension either not being tested, or tested last with only a few items. In a sense, this approach is justified given the extremely low literacy levels in many contexts, where it is assumed that if children do not reach acceptable decoding or reading fluency levels, reading comprehension should not yet be tested. However, as countries start to move towards higher levels of learning, more research and policy emphasis should be placed on reading comprehension. This is especially true given that it is not clear whether the same interventions that managed to improve earlier skills like decoding are as effective at improving reading comprehension (Catts, 2018). In fact, Catts (2018) mentions that it is a "[false] impression that comprehension is not all that different from decoding in terms of its complexity and malleability", as reading comprehension is indeed a "multidimensional cognitive activity and one of the most complex behaviors that we engage in on a regular basis". Therefore, as countries aim to improve functional literacy in developing countries, reading comprehension should be treated as a separate, and complex process that requires particular attention on how it develops and how it can be fostered for all children.

The current study serves a two-fold purpose. Firstly, it provides the impact evaluation of an early grade reading intervention in rural Guatemala. In particular, it examines whether a program that provides teacher training, coaching, and high-quality, context- and grade-appropriate reading materials is an effective driver of literacy outcomes, including reading comprehension. The study takes place in a rural department of Guatemala, where indigenous groups make up a very large share of the local population, and socioeconomic indicators are particularly weak, even within the Guatemalan context. Leveraging the need for a staggered rollout of the program due to oversubscription within

the implementing partner, entry into the program in 2018 is randomly assigned at the end of 2017. This setup allowed for a randomized controlled trial (RCT) among first and second graders in 2018, with a follow up in 2019, to understand whether this specific bundle of treatments can be effective at raising learning outcomes in a developing country.

The second purpose of this study is to understand more systematically and granularly how reading comprehension develops, particularly in this highly understudied context, by translating concepts from the early literacy literature and applying them to the data at hand. I leverage the "Simple View of Reading" (Gough and Turner, 1986; Hoover and Gough, 1990), whose central claim is that children need both decoding skills and oral language skills to then develop proper reading comprehension skills. Within the early literacy literature, the Simple View of Reading is a well-known framework, strongly backed by scientific evidence, to diagnose and understand reading comprehension development in children within the early literacy literature and policy (Language and Reading Research Consortium, and Chiu, 2018). I translate this framework into a classical economics model of "perfect complements", and test its predictions as such, to understand whether this is a valid framework for the design of interventions aiming to improve reading comprehension in developing contexts.

I find that, on average, the intervention was highly effective at improving reading comprehension outcomes, in the order of 0.43 SD or 0.5 school years of business-as-usual education. These effects were largely driven by first graders, who improved 0.8 school years of business-as-usual instruction relative to the control group. I also find effects on oral language proficiency, writing, and decoding of 0.29 SD, 0.32 SD, and 0.23 SD respectively. The increase in learning also spurred a reduction in grade repetition rates for first graders in the order of 3-12 p.p., suggesting a potential channel for fiscal savings after the implementation of a similar intervention. In terms of the literacy mechanisms through which reading comprehension improved, I find substantial support for the "Simple View of Reading" framework. I execute this task through the understanding of the reading comprehension production function as a perfect complements model with decoding and oral language as its inputs.

In all, this paper makes overall two key contributions to the existing literature. First of all, it contributes to the set of policy interventions that have been successful at raising reading comprehension outcomes in developing countries. Especially given the (at best) weaker outcomes that previous interventions which implemented only one of the three components of the Spark program in isolation (e.g. teacher training and coaching in Cilliers et al., 2020, or Yoshikawa et al., 2015, and in a meta-analysis in Stone et al. (2020), and textbooks in Glewwe et al., 2009; Shabarwal, et al., 2014), this evaluation suggests that the thoughtful combination of the three can be much more effective than any of these in isolation, in line with the growing literature on complementarities in educational inputs

(Mbiti et al., 2019). Secondly, by considering the development of literacy "sub-skills" as the mechanism for the observed growth in reading comprehension, this paper highlights the need to thoughtfully design interventions that aim to improve reading comprehension outcomes in a way that addresses each student's pedagogical constraint to improve their reading comprehension skills. Specifically, I advocate for a deeper understanding of reading comprehension as a complex process that integrates many emergent literacy skills, as opposed to a conception of reading comprehension emerging naturally and linearly after students achieve the more basic literacy levels.

This paper proceeds as follows: section II provides an overview of the context for this study, and the data available for the analyses in the following sections. Section III provides a description of the achievement levels and growth of students under business-as-education education in this context. Section IV describes the intervention, and Section V estimates the causal effects of this intervention on literacy skills. Section VI analyzes how the "Simple View of Reading" framework can be used to understand growth in reading comprehension levels, and Section VII concludes.

II. Intervention – The "Spark" Program

The specific intervention studied in this paper is the "Spark Program" by Asociación COED in Guatemala. The Spark Program has been in place for over 10 years, and has been rolled out to schools throughout different municipalities in Guatemala, as funding becomes available to COED. The intervention, as studied in this paper, focuses on developing children's literacy skills, particularly in the early grades, and is structured around three main components. The first component of the intervention entails three in-person, intensive teacher trainings per year over two years, spaced out throughout the school years. These trainings are carried out in small groups, and consist of interactive discussions and "learning by doing" of best practices to teach early literacy skills. Anecdotally, in the words of COED employees, the focus of these trainings is to move away from the predominantly antiquated practices in Guatemalan schools, which are not usually implemented with evidence-based frameworks in mind. For instance, these status quo practices might include excessive writing on the board so students simply copy from the board, the expectation that having materials with letters in the wall without major engagement would still yield learning improvements, and the lack of oral language stimulation. Instead, the trainings for the Spark program were designed to align with the recommendations put forth by evidence-based reports such as the National Reading Panel Report in the United States (NRP and NICHD, 2000).

Shortly after each training, teachers receive the second component of the intervention: an in-person visit from an instructional coach to observe how the practices taught during the trainings are implemented, and to then discuss other ways to further

strengthen each teacher's instruction. During the coaching visits, coaches also appraise implementation of activities through evidence of work performed in class, discussion with students, and a rapid review of the plans prepared by each teacher. The teacher trainings and coaching sessions are mandatory for all teachers teaching a focal grade (1-2) in a school that has adopted the Spark program.

As the third component of the intervention, COED provides schools with a set of high-quality, age-appropriate and culturally-sensitive, and curriculum-aligned reading materials, tailored to the Guatemalan curriculum for grades 1 and 2. Each teacher receives between 60 to 144 textbooks, depending on their grade and class size, aiming for an average of one book per two children after each training session. Importantly, teachers from grades 1 and 2 for this specific project got the exact same training and coaching, without any grade-level customization other than fine-tuning advice provided during the coaching and follow-up visits. Finally, the overall cost of the program is about $77 per student per year (at the current scale of the program), including all material and personnel expenses for the teacher trainings, coaching, and textbooks, but excluding any costs related to this impact evaluation. The specific breakdown of these per-student costs were as follows: USD 27.8 (36%) for books and school supplies, USD 11.6 (15%) for training-related expenses, USD 20.6 for coaching-related expenses (27%), and USD 17 for administrative, coordination and personnel costs (22%).

III. The "Simple View of Reading"

1. Theoretical framework

Beyond reporting the effects of this intervention on early literacy sub-skills, understanding more granularly how this improvement happened from an individual and developmental point of view is perhaps just as important for the design of future interventions. The main early literacy framework that I leverage to understand these developmental mechanisms is the "Simple View of Reading" (SVR). The SVR is a theory in the field of early literacy which predicts that in order for a child to develop reading comprehension (R) skills, they must possess both decoding[3] (D) and oral language[4] (L) skills. Evidence in favor of the SVR has accumulated for decades since it was first proposed by

[3] Gough and Tunmer (1986) provide a thorough discussion of what they mean by "decoding". Without giving a precise definition, the authors do set the boundaries for this skill. On one extreme, decoding can mean "sounding out" words (in the literal sense of "decoding", or the recreation of a word through the alphabetic correspondences of letters to sounds). On the other extreme, decoding may imply word recognition in the sense of "sight words", for which the pronunciation is immediately retrieved from memory without any need of sounding out the word. While the latter definition is more closely linked to skilled decoders that can accurately and silently read isolated words quickly, the former definition refers to the true ability of using the emergent literacy building blocks to decipher the sound of a word.

[4] "Oral language" (which Gough and Tunmer (1986) originally refer to as "listening comprehension", and Hoover and Gough (1990) refer to "linguistic comprehension") is the ability to "take lexical information and derive sentence- and discourse-level interpretations" (Language and Reading Research Consortium, 2015), or in plain words, to make sense of (oral) discourse. In a sense, this skill is similar in spirit to reading comprehension, with the exception that the reading comprehension process starts with written symbols (which also must be learned), as opposed to spoken words.

Gough and Tunmer (1986), and expanded by Hoover and Gough (1990). In the words of Catherine Snow, "few hypotheses in the field of literacy have proven as robust as the SVR. The basic claim has been confirmed in dozens of studies, and the dissent that has emerged does not threaten that central claim" (Snow, 2018). The classical version of this theory states that the relationship between decoding and oral language is "multiplicative" (not "additive"), meaning that if a child is very strong on one, but weak on the other, their reading comprehension skills will still be weak. Although some details about the theory are still up for scientific debate[5], much evidence suggests that indeed individual differences in decoding and oral comprehension can account for much of the variance in reading comprehension (Catts, 2018, Catts, et al., 2005; de Jong and Van Der Leij, 2002; Hoover & Gough, 1990). Similarly, previous work has indeed shown that oral language skills and decoding skills are separable constructs, and that they indeed predict later reading comprehension (Hjetland et al., 2020)[6]. In all, a large body of work provides strong support for the SVR as a general framework to understand the developmental processes behind reading comprehension.

Although the SVR proposes decoding and oral language as the main inputs into reading comprehension skills, it does not necessarily propose that they are weighted equally, or that their respective contribution to reading comprehension is fixed throughout a child's developmental process (Catts, 2018; Catts et al., 2005; Language and Reading Research Consortium, 2015; Tilstra et al., 2009). Hoover and Gough (1990) propose that, if literacy is understood as the gap between a person's linguistic comprehension and reading comprehension, and if decoding skills are adequate enough to efficiently decode any word encountered, then "the limit on reading is the limit on linguistic comprehension". This suggests that indeed the importance of decoding skills decreases relatively to oral language comprehension with child development, in line with other studies such as Lonigan et al. (2017) find. The exact point when the shift in relative importance from decoding to oral language happens is still an empirical question, but broadly speaking, it happens "once decoding has become faster and more automatic, and the vocabulary, grammar, and discourse demands have increased" (Catts, 2018)[7].

The vast majority of the work that supports the SVR has been done through English-language assessments (Catts, 2018) in developed countries. Studying the SVR in other languages (and for students who may not be native speakers of this language, as it is the

[5] For example, the actual functional relationship between oral language and decoding, or the relative importance of each of the two inputs along the developmental process.

[6] However, the extent to which they are the *only* two inputs is also still up for debate. Some work has argued that reading fluency has a separate role within the SVR (Silverman et al., 2013), particularly in languages with semi-transparent orthographies like Spanish (Montesinos et al., 2016), but there is not as much empirical support for this claim as for the more classic SVR framework.

[7] Among English-speaking children in developed countries, this shift seems to happen around third or fourth grade (Catts, 2018; Catts et al., 2005; Language and Reading Research Consortium, 2015), although this inflection point may depend on the transparency of different orthographies and the instructional methods in different contexts (Joshi et al., 2015).

case here) is not trivial: different languages have different phonological and grammatical structures that might in turn affect how literacy skills are acquired. Hence, the relative contribution of decoding may strongly depend on the direct correspondence of grapheme-phonemes in a language, or the degree of familiarity needed with specific syllables, and exceptions to decode unknown words. In the context of this study, Spanish has a simple phonological structure with semi-transparent orthography (Meneses et al., 2017; Ardila and Roselli, 2014; Defior and Serrano, 2014), and hence the initial marginal returns to improving decoding skills may be higher than in languages with more opaque orthographies like English or French. Whether this also translates to higher initial contributions of decoding to reading comprehension within the SVR framework is an empirical question.

2. The SVR as a "perfect complements" model

Following the classical multiplicative proposition of the SVR, the SVR can be translated to the economics literature through the classical framework of "perfect complements". In particular, one can think about a simple "reading comprehension production function" with decoding (D) and oral language (L) as its inputs, and a production function $R(D, L) = \min\{\alpha D, \beta L\}$. I display this model graphically in *Figure 1* through three isoquants at different levels of reading comprehension production R (where $z > y > x$, and $\alpha D = \beta L$ denotes the expansion path).

Practically, this model provides a structure with testable hypotheses to understand how increases in reading comprehension levels might happen in this sample, and whether the SVR is a valid framework to explain these potential gains. These hypotheses relate the marginal product (MP) of each input, and the marginal return of technical substitution (MRTS), given a starting point (D_i, L_i) for student i. Specifically, this framework would predict that the magnitudes of the gains in reading comprehension would be differentially correlated with gains in decoding and oral language, depending on each student's starting point. For instance, in the case of someone for which $\alpha D > \beta L$ (such as point A in *Figure 1*), the $MP_d = 0$, the $MP_L = \beta$, and the MRTS (defined as MP_d over MP_L) would be zero. In plain words, if a student has much stronger decoding skills than oral language skills, then the added benefit to reading comprehension of strengthening their decoding skills even further (i.e., moving point A to the right) are predicted to be null, since this was not the input that was constraining this child's reading comprehension (graphically, any rightward movement of point A would still keep them on RC = y). Contrarily, if this child's oral language is improved by a given amount (i.e., moving point A upwards), their reading comprehension would in turn increase. The marginal rate of technical substitution being zero would mean in this case that there is no amount of strengthening this child's decoding skills that would improve their reading comprehension as much as improving their oral language skills even

by a little bit. These are all very strong predictions, but they serve as an initial framework to later test the implications of the SVR in this particular context.

IV. Study setup

a) Context

The experimental evaluation project was conducted in Guatemala, as part of the impact evaluation of the "Spark" literacy project ran by the local non-profit "Asociación COED" (COED). Guatemala is a middle-income Central American country with high levels of poverty and inequality. Approximately 6 of every 10 Guatemalans were classified as poor by local authorities (INE, 2014), and one in five were classified as extremely poor. These poverty levels coexist with high levels of illiteracy, as one in five adults is illiterate[8]. Broadly speaking, educational outcomes in Guatemala are weak, even compared to peer countries. For example, 19% of all adults in Guatemala are illiterate, compared to 6% on average in Latin America, and the youth illiteracy rate is 6 times the Latin American average. In a more granular definition of literacy, the World Bank estimates that 2 in 3 Guatemala children experience "learning poverty", meaning that they are not proficient readers even by the time they get to grade 6 (World Bank, 2019b). These learning deficits are aggravated by ethnic inequities within Guatemala. For instance, McEwan (2007) estimates that the reading gap in Spanish, the language of instruction, between indigenous and non-indigenous ("Ladino") students ranges between 0.8-1.0 standard deviations, even after controlling for socioeconomic status. This is particularly relevant for this intervention, as the municipality where the intervention takes place has a population who is 78% indigenous, with 97% of these being of the Kaqchikel ethnicity (INE, 2011).

This project was rolled out in the municipality of Tecpán, within the department of Chimaltenango (see *Figure 2*). The population of Chimaltenango is especially poor, even within the Guatemalan context: 21% of the population is severely food insecure, compared to the national average of 14% (INE, 2011). In terms of reading achievement, third graders in Chimaltenango place just below national average, ranking overall 16th out of all 23, across 22 Guatemalan departments, plus the capital (INE, 2018). In Appendix A, I give a thorough description of what "business-as-usual" early literacy development is like for schools in this sample (i.e., for control schools). Importantly, Tecpán also mimics the national disparities in academic achievement between indigenous and non-indigenous students, as *Figure 3* shows. Throughout primary school, the failing rate for Kaqchikel students is much higher than for non-indigenous students. Interestingly, high failing rates of grade 1 students in Central American contexts has been linked before (for instance in Costa Rica by Rodriguez-Segura, 2020) to the low reading achievement of some students.

[8] World Bank Development Indicators: Literacy rate, adult total (% of people age 15 and above).

b) Sampling of schools and students

The precise setting for this study was COED's planned expansion for their Spark program to the municipality of Tecpán in 2018. At the end of 2017, COED allowed public schools to express interest in joining the program. After schools signed up, the number of interested schools exceeded their logistical capacity for yearly rollout. Specifically, 15 schools expressed interest in implementing the Spark program, but COED only had capacity for approximately half of the schools. Therefore, COED decided to have a staggered rollout, and entry into the program in 2018 was randomly allocated for 7 schools at the end of 2017. In 2019, the rest of the 8 schools entered the program. Therefore, the sample consists of 15 schools, all of which entered the Spark program in either 2018 or 2019.

Among the schools that entered the experimental sample, COED randomly sampled at least 12 students per focus grade (first and second grade), which led to an average of 27.9 sampled students per school across both grades, for a total of 419 students across all 15 schools for 2018. For 2019, the sample consists of 450 students, approximately half in each grade. All students are sampled at baseline, and followed to the end of the year. The students in second grade in 2018, exited the program in 2019 and therefore a follow up was not possible. The students in first grade in 2018, were tracked to the best of COED's abilities into 2019 for baseline and endline measurement. Approximately 81% of the first graders in 2018 were surveyed again in 2019, and the rest of the 230 students making up the sample in second grade in 2019 were randomly selected to fill in the place of the students that COED was not able to locate[9]. In 2019, there was also a new cohort of first graders into the Spark program, although by 2019 Spark had expanded to all schools in the sample, leaving no pure control group in 2019.

V. Data

a) Data availability

I use several sources of data for the current study. The first source of data is detailed, item-level early literacy data for 678 unique grade 1 and 2 students between 2018 and 2019, across the 15 schools where the Spark program was rolled out, for the baseline and endline of each year. This dataset was collected through a hired independent third-party based at the "Centro de Investigaciones Educativas" from the Universidad del Valle in Guatemala. This early literacy assessment consists of an adaptation of the Early Grade Reading Assessment ("EGRA", RTI International, 2015), modified by the Guatemalan Ministry of Education to officially test literacy skills in early grades, naming the local adaptation of the test "Evaluación de Lectura en Grados Iniciales" or "ELGI". ELGI follows in spirit all the

[9] Within schools and experimental groups, I find no patterns of differences on observables between the first graders of 2018 compared to the first graders of 2019, in general or separately within experimental groups, or among the students that were selected to complete the sample and those that were originally selected to be in the sample.

EGRA elements, and was mostly adapted to use local vocabulary and examples (Del Valle et al., 2017). Following EGRA standards, ELGI does not change by grade, allowing researchers to directly compare outcomes across grades. Of all 419 students at baseline in 2018, endline ELGI outcomes are available for 329 students. As such, I present in section VI.6. a thorough discussion of how attrition may bias the main estimates of causal effect of the intervention on reading outcomes.

The second source of data was a survey that COED carried out for all students at baseline of both years. This survey asked for basic demographic information including gender, age, indigenous status, academic history, and school information. These data also include survey questions regarding their attitudes towards learning, literacy practices, and school in general at baseline and endline. The third source of data are some publicly available geolocated and administrative datasets on aspects such as population count surrounding a school, nightlights, enrollment, and remoteness. The goal of these data is to understand how the context for these school communities may compare amongst themselves, with the rest of the region, and within Guatemala as a whole. Finally, the fourth source of information is from codified teacher observations and teacher surveys. These are only available for the treatment and control schools at endline of 2018, and the quality of these data is relatively poorer than the other datasets, so this can only be used descriptively and not for any substantial analysis.

b) Outcomes

Since the main aim of the program is the development of early literacy skills, the main outcomes are also sub-skills of early literacy. As mentioned before, the main tool to measure literacy skills was the ELGI, a local adaptation of the EGRA exam. In particular, the ELGI test items can be aggregated into "literacy sub-skills" to understand changes in constructs like oral language, alphabetic principles, phonological awareness, decoding, reading fluency, reading comprehension, and writing skills[10]. For the purposes of analysis, all of these variables are standardized as percentages from 0-100, where a 100 means that a child scored all questions correct for this given skill. Leveraging the richness of the item-level data, I also create other potential outcomes of interest. First, I create a binary outcome for whether a child is in "early grade learning deprivation" (in the spirit of the "learning poverty" concept proposed by the World Bank), where a child is classified as learning deprived if they answer fewer than 80% of all reading comprehension questions correctly. Secondly, I create an outcome indicating whether a child is repeating the grade for each year. I also attempt to recreate ASER/Uwezo-like measures of a student's particular literacy level (e.g., "letter-level", "word-level", "sentence-level"), where a 1 means that a child

[10] For more details on how the items are aggregated into these larger categories, see Appendix B.

reached at least that level. The description of how these variables were created is also in Appendix B. Finally, I also use as outcomes the students' practices and attitudes towards literacy to understand whether the intervention also changed any of these measures. All of the outcomes described here are calculated for all four rounds of literacy data collection (i.e., baseline and endline for both 2018 and 2019).

c) Balance at baseline

To establish the validity of the randomization process, *Appendix Table 2* formally tests for baseline covariate balance across experimental groups. Given the richness of the baseline and survey data, I can perform an analysis of balance on baseline covariates to ensure that the randomization process yielded similar treatment and control groups, at least on observables. The sample is balanced along the lines of demographic characteristics, academic history, and academic attitudes and behaviors. However, I find some imbalances in baseline performance across experimental groups. Specifically, the treatment students are slightly higher performing than the control students. The process of randomization was carried by COED using statistical software, where schools where assigned a random number once and the sample was then divided into two groups (i.e., without checking for balance). Given the small number of schools, this sample lends itself to more extreme outcomes than larger samples. To quantify this, I use randomization inference to estimate the probability of getting a draw at least this extreme. Depending on the outcome observed, I estimate this possibility to be between 1 in 4 to 1 in 6, which seems reasonably likely. Furthermore, I do not see substantial patterns in school characteristics like enrollment, male-female pupil ratio, remoteness, urbanicity, elevation, or population surrounding the school. In sum, this imbalance will be accounted for using different empirical strategies as I describe in the following section, but I do not find systematic evidence to believe that this imbalance was due to anything other than an unlucky draw.

VI. Results

1. What is the effect of the intervention on early literacy skills?

The first key research question that I explore is whether the intervention had a causal effect on children's early literacy skills. To address this first question, I will test three models, for which I show results in *Table 1*. A major consideration of these models is that the randomization yielded a slightly imbalanced sample at baseline along pre-intervention achievement levels. Therefore, these models must all include a feature that controls for this imbalance, and does not spuriously attribute effects to the intervention that were actually due to the differences in baseline starting points[11]. I begin with a simple model ("model 1")

[11] Empirically, when I estimate these models without controls for baseline achievement, the (biased) coefficients of interest with the "causal" effect of the intervention are in the order of two to three times higher than those reported using baseline controls.

which estimates the intent-to-treat effect of attending a treated school in 2018 on achievement outcomes with the following ordinary-least squares regression. Specifically, for student i, grade j, and school k:

$$[1] \qquad \text{Endline score}_{ijk} = \beta_0 + \beta_1(\text{treatment}_k) + \beta_2(\text{baseline score}_{ijk}) \; + \mu_j + \varepsilon_{ijk}$$

Model 1 serves as my starting point to estimate the causal estimate of the intervention. This model includes the pre-intervention literacy level for each student i to control for any potential initial differences in performance. The model also includes grade-level fixed-effects (μ_j) to de-mean the outcomes, since students from both grades were given the same test instrument. Given the small number of clusters (15 schools), I cluster bootstrap standard errors at the school-level[12]. Alternatively, I also estimate the standard errors and significance levels using a randomization inference procedure, which I show in *Appendix Table 3*. In general, the results from using randomization inference are *less conservative* in terms of statistical significance than the results using cluster bootstrapped standard errors, so I decide to show the most conservative estimates in the main text. Similar to this model, "model 2" has all the same features as model 1 but includes a vector of individual-level covariates[13] in the following manner:

$$[2] \qquad \text{Endline score}_{ijk} = \beta_0 + \beta_1(\text{treatment}_k) + \beta_2(\text{baseline score}_{ijk}) \; + \mu_k \; + \mathbf{B}(\text{covariates})_{ijk} + \varepsilon_{ijk}$$

For the third model, I follow a different methodological approach through the estimation of a school fixed-effects model, which also leverages the 2019 data. This method is a different way to address the performance imbalance at baseline, by controlling for unobserved and time-invariant characteristics of the schools that may also be responsible for the imbalance at baseline. In other words, this model is alike to a panel-data estimation where units are compared pre- and post-, across treated and untreated units. In particular, pooling the baseline and endline data for both years, I can characterize if a specific grade by year combination has yet received treatment at any of the four periods of data collection, and code this as the "treated" variable. This variable indicates (=1) if a given grade and school have been yet treated, and it is 0 otherwise. For example, the students who were in grade 1 from the original treatment schools at endline in 2018 were not treated by the baseline of 2018, but were by the endline of 2019, and onwards. On the other hand, the

[12] There is relatively low variance in the size of the clusters in the sample. For 2018, the average number of students sampled per school was 28, but the standard deviation was only 12. Similarly, only 3 schools are more than one standard deviation above mean the mean, and none are one standard deviation below. For this reason, I use the cluster bootstrapped standard errors, as opposed to a wild cluster bootstrap. However, results are very similar if a wild bootstrap is used.

[13] These covariates include the student's age at baseline, indicator variables for their sex, whether they speak an indigenous language at home, whether this is their first time in grade 1, whether they had pre-primary education, whether they have a male teacher, whether their instruction is in Spanish, as well as an indicator for their classroom stream within their school.

students who were in grade 1 from the original control schools were only coded as treated by the endline of 2019. A nice feature of the data is that since no school had yet received treatment by the baseline of 2018, but every school had been exposed to treatment by the endline of 2019, all clusters have variation in treatment status, expanding the common support across the full sample. I specify this model as follows:

$$[3] \qquad \text{Endline score}_{ijkt} = \beta_0 + \beta_1(\text{treated}_{jkt}) + \lambda_k + \mu_i + \eta_t + \varepsilon_{ijkt}$$

Where the model includes school fixed-effects (λ_k) to absorb the unobserved time-invariant characteristics of schools, time period fixed-effects (η_t) to absorb the secular trends, and grade fixed-effects (μ_i) to de-mean the outcome, as all students were tested with the same instrument. The standard errors are again clustered at the school level.

I display the results stemming from models 1-3 on the main literacy outcomes in *Table 1*[14]. In general, the three models yield similar outcomes. The intervention had large, but noisy effects, particularly on areas like oral language, decoding, reading comprehension, and writing[15]. Nevertheless, these gains are indeed economically large: the increases in the reading comprehension score within the 8.3-10.7 range translate to 0.43-0.55 standard deviations (SD) or according to the growth patterns displayed in *Appendix Table 1*, the equivalent of 0.5-0.6 years of education with business-as-usual instruction. Benchmarking this estimate with other similar interventions within a review of early grade reading interventions in Graham and Kelly (2019), this effect size in reading comprehension is slightly above the mean Graham and Kelly report of 0.45 SD. However, this estimate was much lower in terms of average years of business-as-usual education than what these effects represented in the other contexts reviewed by Graham and Kelly, 2010 (2.86 years). This fact likely displays the even lower rates of business-as-usual learning in other contexts included in Graham and Kelly (2019). Similarly, the increases in oral language proficiency and writing translate to gain of 0.29 SD and 0.32 SD respectively. The oral language proficiency can be contrasted to a higher mean effect size in Graham and Kelly (2019) of 0.47 SD. The increase in decoding skills translates to a gain of 0.23 SD. There are smaller effects reading fluency and alphabetic principles. For example, using the measure of reading fluency of correct words per minute (cwpm), this translates to approximately 3.1 additional cwpm, again on the slightly lower end of the distribution in terms of similar studies included in Graham and Kelly (2019). Still, this measure seems to be particularly hard to move in economically meaningful ways: only one third of all program-language groups reported by Graham and Kelly (2019) reported effect sizes above 5 cwpm. There are no effects in

[14] I also use an inverse probability weighting model to check for the robustness of the results of models 1 and 2, which is shown in *Appendix Table 4*. In general, the results are very similar, and as expected, the precision of the estimates increases with the IPW model.
[15] This pattern is supported by the randomization inference levels of significance displayed in *Appendix Table 2*.

emergent literacy skills like phonological awareness, and rapid automatized naming across any of the models or specifications[16].

2. Did the effects vary by grade?

Given the different developmental stages in terms of literacy acquisition for children in grades 1 and 2, it would not be surprising if the intervention had heterogenous effects by grade. This is especially true given that teachers from both grades received the *same* training and coaching, although the materials were actually leveled by grade. To explore the question of heterogeneous effects by grade, I now run models 1-3 but including an interaction term between treatment status and a binary variable for grade 2. These results are displayed in *Table 2*, and the interaction term is reported on all three columns labeled "Treatment x second grade". This coefficient, although noisy, is large and consistently negative for most models and outcomes, suggesting differential effects by grade. I show in *Appendix Table 6* the result of running models 1-3 on grade 1 students exclusively, and in *Appendix Table 7* on grade 2 students exclusively[17]. The differences in effect sizes are striking, especially for reading comprehension, there is a statistically significant increase in 0.8 years of business-as-usual instruction for grade 1, while for grade 2, there is a statistically insignificant decrease of 0.07 years of business-as-usual instruction. For oral language I see a statistically significant increase in 4.6 years of business-as-usual instruction for grade 1, while for grade 2 I see a statistically insignificant increase of 0.27 years of business-as-usual instruction. For decoding, both grades increased 0.28 years of business-as-usual instruction, although the magnitude is much larger for grade 1, and statistically significant. The intervention also increased reading fluency differentially, although neither gain was as economically large as the gains in other sub-skills. In particular, the intervention increased cwpm for grade 1 students by 4.3, a 41% increase, but only 1.2 wpm for grade 2 students, a 4% increase. Overall, there is strong evidence that the intervention, as designed and implemented in this case, is much better suited to cater to the needs of grade 1 students than grade 2 students.

3. Does increased exposure to the program lead to higher gains in early literacy?

Given that the main intervention is much more effective for grade 1 — in fact, driving most of the gains — I also explore whether increased exposure to the intervention leads to even higher outcomes. While it could be the case that the treatment is just less effective for second graders, it could also be the case that it is only effective conditional on students being also exposed to the intervention as first graders. Hence, the question of exposure is relevant to understand not only what the right grade to target for future

[16] I do show in *Appendix Table 5* that the intervention also had some noisy, but positive increases in the extent to which children report reading and writing, and how often they report reading and writing at home.

[17] Note that the grade fixed-effects term (μ_i) will be omitted since there is no variation in grade when the sample is subset in this way.

iterations of the intervention design, but also the right dosage. I use the full 2018-2019 sample, and create a variable φ_{ijtkm} denoting whether each student has received 0, 1, or 2 years of intervention ("exposure"). As such, for student i, grade j, and school k, and for year m, I run model 4 as follows:

$$[4] \qquad \text{Endline score}_{ijk} = \beta_0 + \varphi_{ijtkm} + \beta_2(\text{baseline score}_{ijkm}) + \mu_{jm} + \varepsilon_{ijkm}$$

Where φ_{ijtkm} represents flexible exposure fixed effects. I show these results in the first two columns of *Table 3*, where each column tests for the difference between the given number of years of exposure and 0. Similarly, I run the same model but restricting to only students who were exposed to the program at least for one year (i.e. excluding students from control schools in 2018). These results explicitly test for the difference in literacy outcomes between those that received two years of intervention versus just one. Interestingly, most coefficients are negative, and some even achieve statistical significance. Taken together, the exposure analyses suggest that increased exposure at best did not provide additional benefits, and at worst hurt some children. This is an important consideration given that the current design of the program is for two years, but each year costs about USD 77 per child, 14% of what the government spends per primary school pupil on average. Therefore, achieving the right grade-level targeting and dosage of the program is key to maximize the literacy benefits it provides.

4. Did the gains in early literacy translate into lower grade repetition rates?

Grade repetition is a major policy problem in Guatemala. In 2018, 11.1% of all primary school students in Guatemala failed their respective grades, and 1 in 5 first graders students did. The most common reason to fail a grade is because a student has not met the minimum learning requirements. In similar contexts, grade repetition in grade 1 is strongly linked with not achieving the minimum reading standards to move on to other grades, as described by work like Rodriguez-Segura (2020) in Costa Rica, and anecdotal evidence in Guatemala. Importantly, when a student fails a grade, they need to enroll in the same grade the following year if they would like to continue their education. This has led to a primary education system where 9.5% of all students are currently repeating a grade, and 1 in 3 students are overage. Tecpán, the municipality where this study takes place is no exception to this challenge[18]. Furthermore, Tecpán follows the national pattern found in McEwan (2008) in terms of the large disparities between Kaqchikel and Ladino students: indigenous students in primary school are twice as likely to fail the grade compared to their non-

[18] Among all municipalities in Guatemala, Tecpán ranks 145th out of 329 municipalities in terms of lowest failing rates in grade 1, with a failing rate of approximately 17.8%, and a grade repetition rate of 7.8% in primary school, but up to 15.1% in grade 1. These statistics are also the best indication of how to contextualize the specific sample for this study within Tecpán: 31.2% of grade 1 students in the control schools were repeating a grade in 2019, twice as high as the municipality average.

indigenous counterparts, and they are 7 p.p. points more likely to fail first grade than their counterparts, as shown in *Figure 3*.

Given that the intervention had such large and positive effects on reading comprehension, I also explore whether these gains translated into lower grade repetition rates. For students who were in grade 1 in 2018, I examine whether they were repeating the grade or not in 2019. Since schools were slightly imbalanced along baseline performance, I run this regression controlling baseline reading fluency, and without controlling for it without major changes to the estimates. In particular, for student i, grade j, and school k in 2019:

$$[5] \qquad \text{Repeating grade at baseline in } 2019_{ijk} = \beta_0 + \beta_1(\text{treatment status } 2018_{jk}) + \beta_2(\text{baseline fluency}_{ijk}) + \varepsilon_{ijk}$$

These results are shown in the second row of *Table 4*, showing an 11 p.p. statistically significance difference, which in turn, serves as my upper estimate. Another way to try to control for the lower estimate is to run model 5 on 2018 data to understand whether there were any baseline differences in grade repetition rates, which I show in the first row and displays an imprecise baseline difference of 8 p.p. Therefore, the lower estimate of the effect on grade repetition is the difference between the 2019 and 2018 rows, roughly a 3 p.p difference[19]. To get a more precise estimate for the cohort most impacted by the intervention, I run this same analysis only for those children who were in grade 1 in 2018, which places the upper and lower bounds for this subsample at 13 and 11 p.p. Therefore, depending on the sample and the approach, the effect on grade repetition ranges from 3-13 p.p.

Even the lower bounds of my estimates have important policy implications. Given that the baseline grade repetition rate for grade 1 in this municipality is 15%, even a 3 p.p. decrease would cut grade repetition rate by 20%, or approximately 84 children out of the 2,783 in grade 1 in the municipality. In turn, a 13 p.p. decrease would cut grade repetition rates in grade 1 by 87%, or 362 children. If the current government expenditure per year per primary school student is USD 540, the total expenditure on grade 1 students in Tecpán is USD 1.50 million. Therefore, the rough potential savings in Tecpán from lowering grade repetition in grade 1 range from USD 45,360-195,480[20]. If the program costs about USD 77 per student per year, adding this program throughout all of Tecpán, roughly speaking, would immediately increase government spending per first grader to USD 617 (over 2,783 students in Tecpán), for a total of USD 1.72 million. However, if the total number of grade 1 students was eventually reduced by 84-362 children, the total costs range from USD 1.49-

[19] Notice that this is fairly equivalent to a difference-in-differences strategy, which yields the same lower bound estimate.

[20] The lower estimate comes from 84 students (-3. p.p.) times USD 540=USD 45,360, and the upper estimate comes from 362 students (-13 p.p.) times USD 540=USD 195,480.

1.67 million. The midpoint (USD 1.58 million) would only increase the education expenditure on first graders by 5.3%, and the lower point would slightly *decrease* it by -0.6% (~USD 9000). These positive downstream effects could lighten (or even fully offset) the fiscal load of the intervention through lower future enrollment and increased learning. A potential criticism of this approach is that there is not enough evidence to say whether this effect in grade repetition rates were driven solely by higher levels of learning, or by the intervention also changing schools' likelihood and attitudes towards students failing. However, given that student learning did increase, this potential causal chain does not seem as policy-relevant.

5. Did the intervention work better for any particular sub-group?

An additional policy-relevant question to is whether the observed gains were driven by any other sub-group, other than by first graders. Therefore, I run models 1 and 2, adding an interaction effect between treatment and baseline performance by sub-skill, and show this in *Table 5*. Examining the two columns labeled (Treatment x BL), I find some evidence that the treatment was somewhat more effective for students with lower baseline scores, as all the coefficients for "treatment x BL" were negative, and some achieved statistical significance. These findings are not very strong, but they help understand any potential bias in the results due to the initial imbalance. If one believes that the two groups had different potential outcomes due to their initial imbalance in performance, and if indeed low-achieving students benefited more from the intervention, then the causal estimates shown in *Table 1* could be under-estimates. I also explore other sub-group heterogeneity by other baseline characteristics, which I show in *Appendix Table 8*, but I do not find systematic evidence of other heterogenous effects.

6. How does attrition change the estimates?

The data contains 419 students with reading outcomes at baseline in 2018, but only 329 of these have endline outcomes. Among those with baseline and endline outcomes, there is a statistically significant difference (at the $p<0.10$ level) of 10 percentage points between treatment and control students, which prompts for a deeper investigation of what may have triggered the difference in response rates. In principle, there were two reasons for which COED did not collect data for a given student. First, if a student was not able to focus enough to take a test, none of their outcomes were recorded. By default, all students were able to focus enough to take the test at baseline, so these students have all been able to take the test at least once. The second way attrition could have happened is because the student was not found at endline. While some absence is expected for these two groups, there is no ex-ante reason to believe that absence happened differentially across experimental groups at endline. Therefore, this differential attrition across the two

experimental groups, if indeed due to the intervention, could be either due to a boost in children's ability to focus enough to take an exam or an improvement in school dropout and retention. In either case, the implication for the causal analysis on learning is that the control group will tend to be a more positively selected group by endline. To further support this, previous work also shows that reading proficiency is positively correlated with self-regulation, and years of schooling (Skibbe et al., 2019; Glick and Sahn, 2010). In all, this suggests again that the main estimates could be, if anything, underestimating the true causal parameters.

I also examine whether certain sub-groups are more likely to not have endline scores, which is shown in the first column of *Appendix Table 9*. I do not find strong or surprising patterns: younger students, and those who do not receive instruction in Spanish seem to have been more likely to have outcomes at endline. The second model in this same table explores whether the attrition varied by treatment status interacting with certain baseline characteristics. Again, no obvious patterns emerge regarding whether there were major differences between those attrited from the treatment and control groups separately.

In order to deal with attrition more rigorously, I show in *Appendix Table 10* three different bounding approaches to understand how the main effects shown in *Table 1* (shown again in the first column of *Appendix Table 10)* vary under different assumptions. The first two approaches follow Manski's bounds (Manski, 1990), as practically implementated by Atteberry et al. (2019). The first approach estimates model 1 by replacing the missing values at endline in the treatment group with the average values from the control group at endline from the control group for the lower bound estimate, and then with the same group's mean for the upper bound estimate. The second approach is similar to the first, but instead, it replaces the missing values with the median scores of each group. The third approach uses Lee's bounds (Lee, 2009), which trims about 12% of the sample. In sum, I find that the Manski's bounds contain the main estimates fairly well. Instead, I find that my main estimates are very close (if not slightly lower) to the lower bound of the Lee's bounding estimates. This is not surprising, given that the expected direction of the bias for the main causal estimates is to underestimate the true parameters. If one assumes that those for which the endline is missing are indeed negatively selected, and that treatment had some positive effect in the extent to which students are present at endline, then Lee's bounds do suggest that the main estimates in *Table 1* are, if anything, biased down.

7. How cost-effective is the intervention?

The full cost of the intervention is approximately USD 77 per student per year. For the purpose of cost-effectiveness analysis, I follow Kremer et al. (2013), and put estimates on a scale of effect size per USD 100. Given that the focus of this paper and ultimate goal of reading interventions is to achieve higher reading comprehension levels, I will use these

effect sizes as the main indicators. Starting with the average increase of 0.43 SD during year 1, this would yield a cost-effectiveness of 0.56 SD per USD 100. For grade 1, the grade for which the effect was the largest, this would instead translate to 0.90 SD per USD 100[21]. However, the intervention is currently designed as a two-year intervention, and the previous analysis suggests that increased exposure do not improve (at best) learning outcomes. Therefore, when considering the full intervention, the average cost-effectiveness would be cut down to 0.28 SD per USD 100 on average, and 0.45 SD per USD 100 for students first exposed to the program in grade 1. I bound the cost-effectiveness of this intervention using the lower estimate of 0.28 SD per USD 100 and the upper estimate 0.90 SD per USD 100. To benchmark this estimate, I use the comprehensive review by McEwan (2015), which cites the cost-effectiveness of 26 different interventions. The two bounds of this cost-effectiveness estimates are among the middle of the cost-effectiveness distribution, trending towards the more expensive side. This stands in sharp contrast with the distributional position of the absolute effect size found above: in a review of effect sizes in international education, Evans and Yuan (2020) report that even an effect of 0.43 SD would be between the 80th and 90th percentile among effect sizes for reading across all studies in general, and also across studies with samples as small as the one in this paper. Given these findings, clearly targeting the intervention at grade 1 students, or improving upon the effectiveness in grade 2, could significantly increase the cost-effectiveness of the program to match its high level of overall effectiveness. Finally, from the point of view of the full educational system, the cost of implementing this intervention is likely an overestimate, since it could yield further downstream savings due to lower grade repetition rates, as previously discussed.

VII. Developmental mechanisms behind changes in reading comprehension

I leverage the perfect complements model of "Simple View of Reading" in Section III and the fine-grain, item-level outcome data to conduct exploratory analyses regarding how different "inputs" or early literacy skills improved in order to unlock reading comprehension changes of the magnitude of those shown in the previous sections. This is also an initial attempt to explicitly weave the literatures of early literacy, and policy and economics closer together, given the evident thematic overlap but relative siloed state of the two. I do so by presenting the components of early childhood literacy as the essential inputs in a perfect

[21] Although I cannot express the main effects of this using the control group standard deviation at baseline because the standard deviation is 0, I pinpoint an estimate of 0.69 SD for grade 1 using the equivalent increases in years of education. In particular, the intervention had an effect of 0.5 additional years of education for the full sample, and 0.8 for grade 1 students, for a ratio of 1.6. Using this ratio, if the main effect in SD was 0.43, the effect in SD for grade 1 students was 0.69 SD.

complements production function and deriving testable empirical predictions from this model, placing early literacy theory front and center within an economics framework[22].

3. Empirical test of the SVR

I test the SVR framework by exploring whether the marginal products of decoding and oral language comprehension on reading comprehension vary depending on each student's starting point (i.e., what they are initially constraint by). First, I create "change" (or growth) variables for decoding (D), oral language (L) and reading comprehension (R) by defining each variable X as $\Delta X_{ijkt}=X_{ijk(\text{endline t})} - X_{ijk(\text{baseline t})}$. These change variables show the baseline-to-endline change in each of the three skills for each child, for each year they appear in the data with endline outcomes. For ease of joint interpretation, I then transform each of these change variables into z-scores using the moments of the control group[23].

Using each student's baseline score in each sub-skill $X_{ijk(\text{baseline t}}$, I then classify students into four groups. By grade, I label each student as "high" or "low" in decoding and oral language respectively (where each threshold is the baseline median), for a total of four "baseline groups". If the model holds, the predictive power of ΔL on ΔR, and ΔD on ΔR would vary by group in line with the predictions of the model. For instance, for the group with low oral language but high decoding, I would expect the change in oral language to be correlated with a larger change in the change in reading comprehension, as oral language (and not decoding) was the limiting input for this specific group. *Figure 4* offers a graphical representation of how these groups would fit into the SVR framework as understood through the less of perfect complements.

Having classified students based on their most limiting input at baseline, I use the following statistical model separately for each baseline group, in the spirit of estimating cross-elasticities in the economics literature, to jointly recover the predictive power of ΔD and ΔL:

$$[6] \qquad \Delta R_{ijkt} = \beta_0 + \beta_1(\Delta D_{ijkt}) + \beta_2(\Delta L_{ijkt}) + X_{ijkt} + \varepsilon_{ijk}$$

[22] As a starting point, I check whether empirically the interaction between the post-levels of decoding and oral language is indeed a predictor of a child's post-level reading comprehension, which I find to be the case at the $p<0.000$ level, whether the model controls for baseline levels of achievement or not. In fact, as Gough and Turner (1986) predict, and Hoover and Gough (1990) execute, the inclusion of the interaction term between the two inputs significantly improves the fit of the model in this sample, with the R^2 going from 0.61 to 0.72.

[23] Using a simple linear model like before with these change variables, the interaction between the changes in decoding and oral language is not statistically significant nor does it add significant precision to the model in terms of R^2, like it did in the case of levels. This is the case whether the model controls for baseline performance or not, and whether the results are given by grade or not. Crucially, this result is indicative of a more complex process underlying reading comprehension development than a purely additive relationship between the simultaneous change in both these inputs, as the classic "multiplicative" understanding of the SVR suggests.

Where X_{ijkt} is the baseline performance for D, L, and R[24], and standard errors are bootstrapped and clustered at the school-level. Since ΔR, ΔD, and ΔL have all been standardized, the coefficients β_1 and β_2 should have similar interpretations. In the full sample, a 1 SD change in decoding is correlated with a 0.51 SD change in reading comprehension, and a 1 SD change in oral language is correlated with a 0.37 SD change in reading comprehension, for a ratio of the decoding power of decoding being 1.38 times as large as the predictive power of oral language.

The results of model 6 are displayed separately for each of the four baseline groups in *Table 6*. As predicted by the perfect complements framework of the SVR, the limiting input was significantly more predictive of a change in reading comprehension for the groups that had only one constraint. In particular, if oral language was the limiting input, then a 1 SD change in oral language was 1.6 times (1 divided by the 0.62 from *Table 6*) more predictive of a change in reading comprehension than a 1 SD change in decoding, as well as it was more precise from a statistical point of view. Similarly, when decoding was the limiting input, a 1 SD change in decoding was 2.8 times more predictive of a change in reading comprehension than a 1 SD change in oral language, as well as it was more precise from a statistical point of view. When none of the inputs was particularly "problematic" (i.e., either if they were both low, or both high), the predictive power of decoding hovered around 1.5-1.6 times larger than that of oral language, consistent with a perfect complements model where these two groups would hover around the kink along the expansion path. In general, these results also agree with the heightened importance of decoding relative to oral language in the initial stages of reading comprehension development discussed above. It is only for students with high decoding skills that oral language plays a more important role in predicting gains in reading comprehension, as Catts (2018) also predicts.

In terms of limitations of this framework, the empirical approach does not mimic identically the perfect complements model, as the returns to the non-constraining inputs in the groups with only one constraint are non-zero (i.e., even for students with strong decoding skills, an improvement in decoding is still positively correlated with increases in reading comprehension increases). Secondly, as discussed before, the actual ratio of the returns to decoding and returns to oral language is likely to vary by language, development stage, and age, as both of these skills likely achieve ceiling effects at different points (e.g., one can only learn 27 letter names, along with the finite number of syllables that these allow in Spanish, whereas the depth of a child's vocabulary can always be deepened further). Students in this sample were mostly still strengthening their emergent skills, and hence

[24] The results are largely the same without controlling for baseline performance in each of the variables. However, my preference is to keep them in the model to acknowledge that the tangible implications of one unit of growth in any of these sub-skills may differ depending on the stage along the developmental process of the child, with the added benefit of increased statistical precision.

decoding was still playing a large role in their reading comprehension. As a suggestive piece of evidence, when the analysis for *Table 6* is recreated only with grade 2 students, the ratios and conclusions remain qualitatively the same, except for the high/high group (i.e., the highest performers in the whole sample), for which changes in decoding are less than half as predictive of changes in reading comprehension as changes in oral language. Therefore, putting a definitive number to the fixed proportions needed for these two skills seems a more context-dependent endeavor, worthy of more exploration in different data sets. An additional limitation is that the experimental design did not explicitly set out to investigate these patterns, so any evidence presented here should be taken as suggestive evidence, instead of definitive evidence, of the complementarity of these skills in the improvement of reading comprehension skills. In spite of these limitations, these results indeed support the SVR, understood in its multiplicative form, as a powerful vehicle to understand the developmental mechanisms and sub-skill improvements that led to large gains in reading comprehension. This finding adds to the body of evidence that supports the SVR as a valid framework to understand early reading comprehension, particularly in a highly understudied context and language.

4. Implications for some classifications of literacy achievement

An even deeper implication of this analysis is the understanding of early literacy development as a layered process, rather than strictly linear – in line with the early literacy literature. Instead of conceptualizing a student's reading skills as moving from smaller units (e.g., letter or syllable) to larger units (e.g., words or sentences), I leverage the Simple View of Reading to illustrate the need to focus interventions and measurements on literacy sub-skills that need to be developed concurrently. On one hand, the extreme resource constraints in many developing countries, high pupil-teacher ratios, and wide intra-classroom variability in achievement have led to policymakers and researchers to explore better ways to group children so that instruction can better cater to their levels and needs. Initiatives that actively group and track children like "Teach at the Right Level" (TaRL) have achieved large gains in many contexts, and are seen as serious policy options to improve learning within these resource-constraint contexts. Similarly, assessments to diagnose learning at a large-scale scale such as ASER and Uwezo have been crucial to detect the worldwide learning crisis, and the dire state of schooling across many countries. On the other hand, both TaRL-like interventions and these large-scale assessments tend to classify students along a linear literacy spectrum such as "letters", "words", "short sentences/simple paragraphs", or stories" (ASER Centre (n.d.), Teaching at the Right Level (n.d.)), and this may curtail the proper placement and effectiveness of these interventions.

To underscore how a more layered approach that takes into account crucial literacy sub-skills may be a better diagnostic tool, I use the detailed item-level data to recreate the

type of linear outcomes that the ASER/Uwezo type of test yields on each of the observations in the 2018 data. I then run model 1 on each of these outcomes, as shown in *Appendix Table 11*. From this point of view, the intervention seems to have been effective at improving letter and syllable knowledge, as well as comprehension/story reading skills, but not the middle skills. It is not clear how to reconcile these findings with 1) active policy recommendations about how to design early literacy programs, 2) why the current intervention yields these results, and 3) the previously shown results that the intervention is more effective for first graders, and to an extent for students with weaker baseline performance. These murkier findings using the "linear classification" of literacy contrast with a more theory-based analysis of emergent literacy sub-skills (in particular, oral language and decoding) and how these come together to build higher order skills like reading comprehension. In all, these findings highlight the need for programs aiming to improve reading comprehension in developing countries to either track and tailor instruction depending on each student's baseline levels of oral language and decoding skills, or if this is logistically unfeasible, to include activities that strengthen both skills so that each child can receive at least some instruction aimed their specific limiting input.

VIII. Discussion

The impact evaluation of the Spark intervention in Guatemala provides two key contributions to the literature. First of all, it provides an example of an intervention that is suitably tailored to the context upon which local governments can model similar initiatives to boost learning in similar rural areas of Guatemala and beyond. This model also provides suggestive evidence of the need for complementary and coherent parts to literacy interventions. For example, a meta-analysis on what works to improve early grade literacy in Latin America by Stone et al., (2020) finds no evidence that, on average, teacher training in isolation had a positive effect on reading outcomes. Similarly, the evidence on teacher coaching in developing countries is limited and mixed to a certain extent (Cilliers et al., 2020). The provision of textbooks and class materials by themselves have not proven effective in other contexts, particularly if there is no context-relevant tailoring of the materials or a steady and reliable source for them (McEwan, 2015; Glewwe et al., 2009; Shabarwal, et al., 2014). However, a recent body of work such as Mbiti et al. (2019) and Kerwin and Thornton (2020) suggests the need for complementary features of interventions to really improve learning. In fact, all early literacy interventions reviewed by Graham and Kelly (2019) include teacher training and many other features such as coaching or provision of materials, and these interventions had for the most part, positive significant effects.

The second contribution offered by this paper is the intertwining of an early literacy framework, the Simple View of Reading, and the more classical impact evaluation techniques to explain how reading comprehension development happens, particularly in this

context. In spite of the large volume of work validation the SVR in English, only one previous study attempts to replicate this framework in Spanish (Montesinos et al., 2016), with a much smaller sample consisting of students from an advantaged background in Peru. This extension is valuable not only from the point of view of creation of knowledge, but also from the more practical point of view of intervention design. While current tracking and grouping interventions in developing countries have displayed strong learning gains, these interventions typically rely on "linear" classifications of students. This classification system implicitly assumes that what needs to be addressed is the unilateral difficulty of instruction, not a particular constraint across separate sub-skills. For instance, a non-indigenous first grader who did not go to pre-primary education may have a high level of oral language in Spanish but low levels of decoding skills. This student may be placed in the same category (say, the "syllable-level") as an indigenous student who may have attended pre-primary education with a good command of alphabetic and syllabic principles but weak Spanish skills. Their constraints are clearly different but a linear classification may place them in the same group with the same prescribed instruction. In this example, it is clear that this is neither helpful from a pedagogical point of view, or from a policy-design point of view. As Catts (2018) mentions, "the implications of this line of research is that if we are going to adequately identify children at risk for the full range of reading disabilities, early screening protocols need to include measures of oral language as well as decoding-related predictors" (Catts, 2018; referencing also Catts et al., 2016; and Foorman et al., 2009).

Clearly identifying each student's constraining skills to achieve higher level of reading comprehension is especially important in contexts where at least a share of the students is not receiving instruction in their native language, and hence oral language skills may be a crucial constraint beyond just the "linear" difficulty of instruction. In other words, in places like Tecpán, where a large share of students does not receive instruction in the main language spoken at home, attention to the multi-faceted components of reading comprehension is key. Importantly, this is in fact the modal experience around the world: it has been estimated that more than half of the global population first learn to read in a second language (McBride et al., 2017). While decoding instruction is needed in these contexts, program design also needs to emphasize the oral and linguistic comprehension skills of children, particularly when some students are starting out their education with a linguistic disadvantage. For instance, in South Africa, grade 3 students performed between 0.3-0.7 SD worse in literacy simply because they were tested in English as opposed to their first language (Spaull, 2016). Therefore, even if mother-tongue instruction were not feasible or desirable from the point of view of the national language policy, and if it is indeed true that the motivation behind literacy interventions is often "compensatory" to enhance language skills in children who are demographically at risk (McBride et al., 2017), then

special emphasis should be placed on helping these students catch up in the language of instruction so they can eventually develop proper reading comprehension abilities.

In all, promoting higher levels of reading comprehension in developing countries is likely to be a pivotal policy step to translate the early gains in emergent literacy skills into larger economic, cognitive, and social gains. Reading comprehension, as the unique skill that takes children from "learning to read" to "reading to learn," needs to play a central role in helping countries escape learning poverty traps. However, reading comprehension is a highly complex task that involves the engagement of many other emergent literacy sub-skills at once. As such, specific knowledge about how it develops, and the type of interventions that best promote it, is needed in order to spend resources effectively in this area of education. These effective designs, in turn, require policymakers and researchers to thoughtfully keep in mind the underlying developmental processes for each child, the relative importance of different skills along the full literacy acquisition process, and the specific constraints in terms of literacy skills that may be holding back some sub-populations so they can be explicitly addressed by these interventions.

IX. References

Abadzi, H. (2011). Reading Fluency Measurements in EFA FTI Partner Countries: Outcomes and Improvement Prospects. GPE Working Paper Series on Learning, No. 1.

Ardila, A. and Rosselli, M. (2014). Spanish and the characteristics of acquired disorders in reading and writing. Estudios de Psicología, 35(3), 502–518. doi:10.1080/02109395.2014.965453.

ASER. (2020). ASER 2019 – Rrural. Annual Status of Education Report (Rural): 'Early Years'.

ASER Centre. (n.d.). ASER Centre: Evidence for Action. Retrieved December 3, 2020, from http://www.asercentre.org/p/50.html

Atteberry, A., Bassok, D., & Wong, V. C. (2019). The Effects of Full-Day Prekindergarten: Experimental Evidence of Impacts on Children's School Readiness. Educational Evaluation and Policy Analysis, 41(4), 537–562. https://doi.org/10.3102/0162373719872197

Azevedo, J. P. (2020). Learning Poverty: Measures and Simulations. World Bank Policy Research Working Paper. No. 9588

Azevedo, J.P., Goldemberg, D., Montoya, S., Nayar, R., Rogers, H., Saavedra, J., Stacy, B.W., William. (2021). Will Every Child Be Able to Read by 2030? Defining Learning Poverty and Mapping the Dimensions of the Challenge. World Bank Policy Research Working Paper. No. 9588

Banerjee, A. V., Banerji, R., Berry, J., Kannan, H., Mukerji, S., & Walton, M. (2016). Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2846971

Catts, H. W. (2018). The Simple View of Reading: Advancements and False Impressions. Remedial and Special Education, 39(5), 317–323. https://doi.org/10.1177/0741932518767563

Catts, H. W., Hogan, T., & Adlof, S. (2005). Developmental changes in reading and reading disabilities. In H. Catts & A. Kamhi (Eds.), Connections between language and reading disabilities (pp. 25–40). Mahwah, NJ: Lawrence Erlbaum.

Catts, H. W., Nielsen, D., Bridges, M., & Liu, Y. (2016). Early identification of reading comprehension difficulties. Journal of Learning Disabilities, 49, 451–465.

Chall, J. S. (1996). Stages of reading development (2nd ed). Harcourt Brace College Publishers.

De Jong, P. F., & Van Der Leij, A. (2002). Effects of phonological abilities and linguistic comprehension on the development of reading. Scientific Studies of Reading, 6, 51–77.

Defior, S. and Serrano, F. (2014). Diachronic and synchronic aspects of Spanish: the relationship with literacy acquisition. Estudios de Psicología, 35(3), 450–475. doi:10.1080/02109395.2014.974422.

Del Valle, M., Cotto, E., & Mirón, R. (2017). Evaluación de lectura inicial en estudiantes de segundo primaria. Dirección General de Evaluación e Investigación Educativa, Ministerio de Educación. Disponible en red: http://www.mineduc.gob.gt/digeduc. https://www.mineduc.gob.gt/digeduca/documents/investigaciones/2018/RESULTADOS_ELGI.pdf

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya. American Economic Review. 101(5), 1739-74.

Durand, V. N., Loe, I. M., Yeatman, J. D., & Feldman, H. M. (2013). Effects of early language, speech, and cognition on later reading: A mediation analysis. Frontiers in Psychology, 4. https://doi.org/10.3389/fpsyg.2013.00586

Evans , D. K., Yuan, (2020). How Big Are Effect Sizes in International Education Studies? CGD Working Paper 545. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/how-big-are-effect-sizes-international- education-studies

Evans, D., & Hares, S. (2021). Should Governments and Donors Prioritize Investments in Foundational Literacy and Numeracy? CGD Working Paper 579.

Foorman, B. R., Torgesen, J. K., Crawford, E., & Petscher, Y. (2009). Assessments to guide reading instruction in K-12: Decisions supported by the new Florida system. Perspectives on Language and Literacy, 35, 13–19.

Francis, D. J., Kulesz, P. A., & Benoit, J. S. (2018). Extending the Simple View of Reading to Account for Variation Within Readers and Across Texts: The Complete View of Reading (CVR i). Remedial and Special Education, 39(5), 274–288. https://doi.org/10.1177/0741932518772904

Glewwe, P., & Kremer, M. & Moulin, S. (2009). Many children left behind? Textbooks an test scores in Kenya. American Economic Journal: Applied Economics, 1(1), 112-135. https://doi.org/10.1257/app.1.1.112

Glick, P., & Sahn, D. E. (2010). Early Academic Performance, Grade Repetition, and School Attainment in Senegal: A Panel Data Analysis. The World Bank Economic Review, 24(1), 93–120. https://doi.org/10.1093/wber/lhp023

Gough, P. B., & Tunmer, W. E. (1986). Decoding, Reading, and Reading Disability. Remedial and Special Education, 7(1), 6–10. https://doi.org/10.1177/074193258600700104

Gove, A., & Cvelich, P. (2010). Early reading: Igniting education for all (A report by the early grade learning community of practice) Research Triangle Park, NC: Research Triangle Institute.

Graham, J., & Kelly, S. (2019). How effective are early grade reading interventions? A review of the evidence. Educational Research Review, 27, 155–175. https://doi.org/10.1016/j.edurev.2019.03.006

Hjetland, H. N., Brinchmann, E. I., Scherer, R., Hulme, C., & Melby-Lervåg, M. (2020). Preschool pathways to reading comprehension: A systematic meta-analytic review. Educational Research Review, 30, 100323. https://doi.org/10.1016/j.edurev.2020.100323

Hoover, W. A., & Tunmer, W. E. (2018). The Simple View of Reading: Three Assessments of Its Adequacy. Remedial and Special Education, 39(5), 304–312. https://doi.org/10.1177/0741932518773154

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. Reading and Writing, 2(2), 127–160. https://doi.org/10.1007/BF00401799

Instituto Nacional de Estadística (INE). (2011). Resultados del Censo Nacional. https://www.ine.gob.gt/sistema/uploads/2014/02/26/L5pNHMXzxy5FFWmk9NHCrK9x7E5Qqvvy.pdf

Instituto Nacional de Estadística (INE). (2011). Resultados del Censo Nacional. https://www.ine.gob.gt/sistema/uploads/2014/02/26/L5pNHMXzxy5FFWmk9NHCrK9x7E5Qqvvy.pdf

Instituto Nacional de Estadística (INE). (2018). Base Educación Formal. [Dataset] https://www.ine.gob.gt/ine/estadisticas/bases-de-datos/educacion/

Joshi, M., Ji, X., Breznitz, Z., Amiel, M., & Yulia, A. (2015). Validation of the simple view of reading in Hebrew—A semitic language. Scientific Studies in Reading, 19, 243–252.

Kerwin, J. T., & Thornton, R. L. (2020). Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. The Review of Economics and Statistics, 1–45. https://doi.org/10.1162/rest_a_00911

Kraay, Aart, and David McKenzie. 2014. "Do Poverty Traps Exist? Assessing the Evidence." Journal of Economic Perspectives, 28 (3): 127-48.

Kremer, M., B. Conner, and R. Glennerster (2013). The challenge of education and learning in the developing world. ScienceMag 340.

Kucirkova, N., Snow, C. E., Grøver, V., & McBride, C. (Eds.). (2017). The Routledge international handbook of early literacy education: A Contemporary Guide to Literacy Teaching and Interventions in a Global Context. Routledge.

Language and Reading Research Consortium. (2015). Learning to read: Should we keep things simple? Reading Research Quarterly, 50, 151–169.

Language and Reading Research Consortium, & Chiu, Y. D. (2018). The Simple View of Reading Across Development: Prediction of Grade 3 Reading Comprehension From Prekindergarten Skills. Remedial and Special Education, 39(5), 289–303. https://doi.org/10.1177/0741932518762055

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. Review of Economic Studies, 76(3), 1071–1102. https://doi.org/10.1111/j.1467-937X.2009.00536.x

Lonigan, C. J., Burgess, S. R., & Schatschneider, C. (2018). Examining the Simple View of Reading With Elementary School Children: Still Simple After All These Years. Remedial and Special Education, 39(5), 260–273. https://doi.org/10.1177/0741932518764833

Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. (2014). Improving early-grade literacy in east africa: experimental evidence from Kenya and Uganda: Improving Early-Grade Literacy in East Africa. Journal of Policy Analysis and Management, 33(4), 950–976. https://doi.org/10.1002/pam.21782

Manski, C. F.(1990). “Nonparametric Bounds on Treatment Effects”. American Economic Review, Papers and Proceedings, 80: 319–323

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, incentives, and complementarities in education: experimental evidence from Tanzania. The Quarterly Journal of Economics, 134(3), 1627-1673. https://doi.org/10.1093/qje/qjz010

McBride, C., Snow, C. Kucirkova, N., Grøver, V. (2017). Old and new: reflecting on the enduring key issues in early literacy. In Kucirkova, N., Snow, C. E., Grøver, V., & McBride, C. (Eds.). The Routledge international handbook of early literacy education: A Contemporary Guide to Literacy Teaching and Interventions in a Global Context. (pp. 124-138) Routledge.

McEwan, P. J. (2015). Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. Review of Educational Research, 85(3), 353–394. https://doi.org/10.3102/0034654314553127

McEwan, P. J., & Trowbridge, M. (2007). The achievement of indigenous students in Guatemalan primary schools. International Journal of Educational Development, 27(1), 61–76. https://doi.org/10.1016/j.ijedudev.2006.05.004

Meneses, A., Rodino, A. M., Mendive, (2017). Lessons from Costa Rica and Chile for Early Literacy in Spanish-speaking Latin American countries. In Kucirkova, N., Snow, C. E., Grøver, V., & McBride, C. (Eds.). The Routledge international handbook of early literacy education: A Contemporary Guide to Literacy Teaching and Interventions in a Global Context. (pp. 124-138) Routledge.

Montesinos, M. M. T., Aguado, G., & Ripoll, J. C. (2016). Validation of the Simple View of Reading in Spanish. https://doi.org/10.13140/RG.2.1.1196.8246

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. American Economic Review, 109(4), 1426–1460. https://doi.org/10.1257/aer.20171112

National Reading Panel (U.S.), & National Institute of Child Health and Human Development (U.S.). (2000). Report of the National Reading Panel: Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. [Bethesda, Md]: U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, National Institute of Child Health and Human Development.

Orozco, H., Santisteban, T., Agreda, C., & Avendaño, N. (n.d.). Guatemala: Perfil del país y análisis de actores clave en lectoescritura inicial. Programa de Capacidades LAC READS. https://lacreads.org/sites/default/files/documents/guatemala-stakeholderanalysis-es.pdf

Ortiz-Ospina, E., & Belketian, D. (2018). How is literacy measured?. Our World in Data. https://ourworldindata.org/how-is-literacy-measured

RTI International. 2015. Early Grade Reading Assessment (EGRA) Toolkit, Second Edition. Washington, DC: United States Agency for International Development.

Rodriguez-Segura, D. (2020). Strengthening early literacy skills through social promotion policies? Intended and unintended consequences in Costa Rica. International Journal of Educational Development, 77, 102243. https://doi.org/10.1016/j.ijedudev.2020.102243

Sabarwal, S.; Evans, D. K.; Marshak, A. 2014. The Permanent Input Hypothesis: The Case of Textbooks and (No) Student Learning in Sierra Leone. Policy Research Working Paper;No. 7021. World Bank Group, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/20339 License: CC BY 3.0 IGO.

Silverman, R. D., Speece, D. L., Harring, J. R., & Ritchey, K. D. (2013). Fluency Has a Role in the Simple View of Reading. Scientific Studies of Reading, 17(2), 108–133. https://doi.org/10.1080/10888438.2011.618153

Snow, C. E. (2018). Simple and Not-So-Simple Views of Reading. Remedial and Special Education, 39(5), 313–316. https://doi.org/10.1177/0741932518770288

Stone, R., Hoop, T., Coombes, A., & Nakamura, P. (2020). What works to improve early grade literacy in Latin America and the Caribbean? A systematic review and meta-analysis. Campbell Systematic Reviews, 16(1). https://doi.org/10.1002/cl2.1067

Teaching at the Right Level. (n.d.). Teaching at the Right Level (TaRL): Assessment. Retrieved December 3, 2020, from https://www.teachingattherightlevel.org/the-tarl-approach/assessment/

Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. Journal of Research in Reading, 32, 383–401.

Uwezo (2019) Are our Children Learning? Uwezo Uganda Eighth Learning Assessment Report. Kampala: Twaweza East Africa

World Bank. (2019a). Ending Learning Poverty : What Will It Take?. World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/32553 License: CC BY 3.0 IGO.

World Bank. (2019b). Guatemala: Learning Poverty Brief. http://pubdocs.worldbank.org/en/640231571223409894/LAC-LCC2C-GTM-LPBRIEF.pdf

World Bank. (2020). Atlas of Sustainable Development Goals: Learning poverty: children's education in crisis. https://datatopics.worldbank.org/sdgatlas/goal-4-quality-education/

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M., Weiland, C., Arbour, M. C. (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. Developmental Psychology, 51(3), 309–322.

X. Figures

Figure 1: the Simple View of Reading modeled through a perfect complements production function

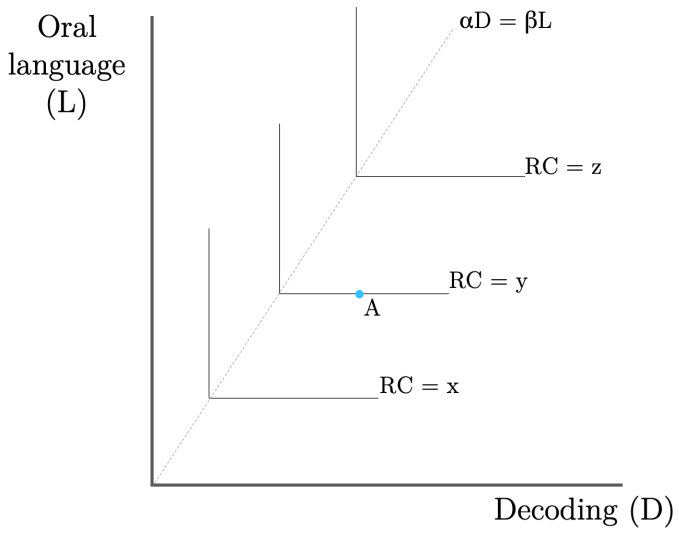


Figure 2: geographic location of schools in the sample

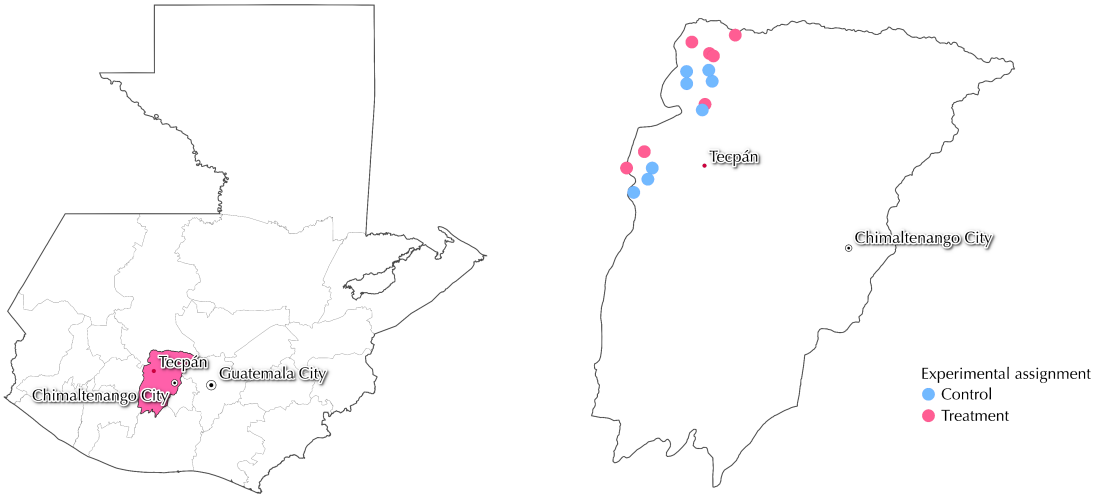


Figure 3: failing rate nationally (left), and in Tecpán (right), by grade and ethnicity

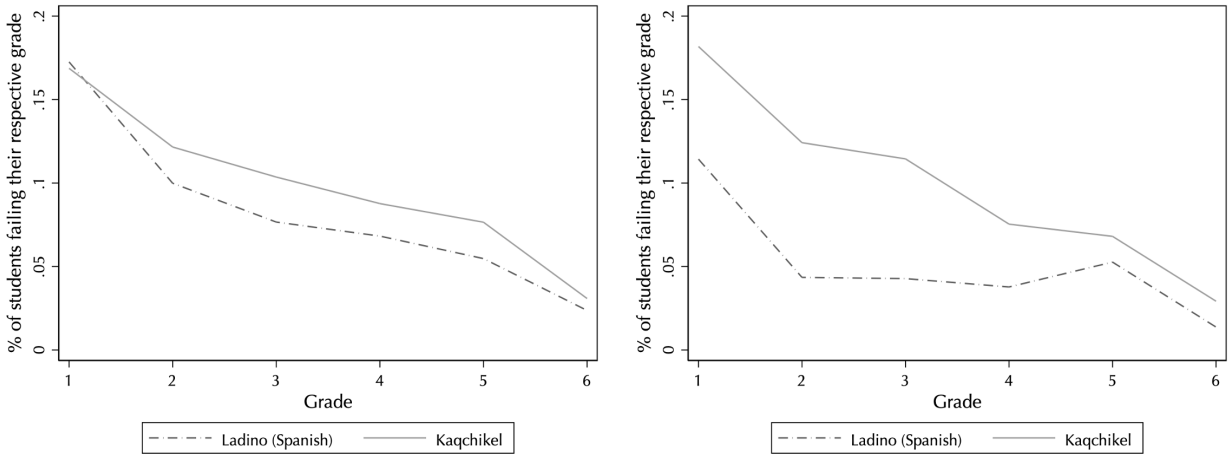
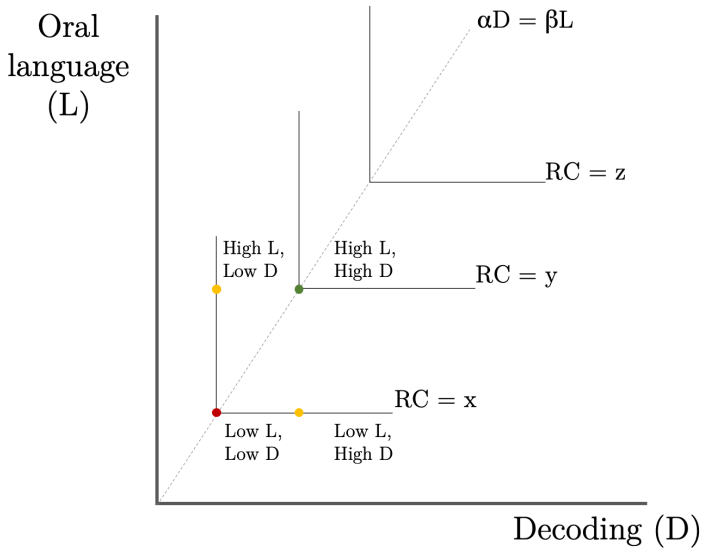


Figure 4: theoretical location of the four baseline groups on reading comprehension isoquants



XI. Tables

Table 1: regression results of the causal effect of treatment on literacy sub-skills

	Mean and SD at baseline	[1]	[2]	[3]
Oral language	49.52	6.21*	5.06	6.08**
	(21.19)	(3.11)	(3.33)	(2.27)
Alphabetic principles	27.85	2.68	1.06	5.2*
	(21.08)	(3.53)	(3.16)	(2.84)
Decoding	25.17	5.79*	4.47	7.83**
	(28.45)	(2.75)	(2.86)	(2.99)
Phonological awareness	38.23	1.32	-1.15	1.58
	(27.97)	(4.21)	(4.08)	(3.26)
Rapid automatized naming	22.70	2.85	2.11	3.75
	(20.85)	(2.47)	(2.13)	(2.47)
Reading fluency	11.72	3.73	2.47	1.36
	(18.94)	(2.28)	(1.76)	(2.22)
Reading comprehension	8.40	10.72	9.03	8.25*
	(19.44)	(6.21)	(5.18)	(4.34)
Writing	22.88	7.42*	5.69	6.02*
	(33.40)	(4.19)	(4.59)	(3.09)
Observations	329	329	321	1568
Control for baseline		Y	Y	School fixed-effects
Demographic control		N	Y	
Years		2018	2018	

Notes: the sample for models 1 and 2 consists of all observations with endline outcomes in 2018, while the sample for model 3 consists of all observations with endline outcomes for 2018 and 2019. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Table 2: regression results of the causal effect of treatment on literacy sub-skills, interacting treatment status with grade

	Mean and SD at baseline	[1]			[2]			[3]		
		Treatment x second grade	Treatment	Second grade	Treatment x second grade	Treatment	Second grade	Treatment x second grade	Treatment	Second grade
Oral language	49.52	-5.66	8.67**	12.67***	-8.48	8.60*	12.38***	1.39	1.69	2.53
	(21.19)	(4.88)	(3.59)	(4.09)	(5.60)	(4.20)	(3.57)	(3.37)	(2.45)	(3.22)
Alphabetic principles	27.85	-1.20	3.19	-2.16	-3.18	2.36	-3.43	-7.24**	4.26	29.47***
	(21.08)	(3.91)	(4.60)	(3.87)	(4.46)	(4.13)	(3.73)	(2.59)	(3.49)	(1.75)
Decoding	25.17	-5.36	8.05	-12.52**	-6.46	7.10	-10.61**	-26.14***	14.28***	37.81***
	(28.45)	(6.31)	(5.33)	(4.51)	(5.56)	(5.11)	(4.65)	(5.16)	(4.73)	(3.31)
Phonological awareness	38.23	0.43	1.14	1.18	-3.32	0.22	6.71	-11.26**	2.01	30.96***
	(27.97)	(8.05)	(6.44)	(6.86)	(6.52)	(5.44)	(5.94)	(4.91)	(5.15)	(2.58)
Rapid automatized naming	22.70	-4.65	4.81	0.85	-6.44	4.74	1.32	-8.12**	4.70	29.49***
	(20.85)	(4.48)	(4.25)	(3.53)	(4.23)	(3.51)	(3.78)	(2.77)	(2.93)	(2.22)
Reading fluency	11.72	-6.46	6.43	-1.52	-9.02*	6.10	0.93	-1.99	3.82	29.85***
	(18.94)	(5.45)	(4.02)	(2.71)	(4.61)	(3.97)	(2.91)	(4.82)	(3.27)	(3.05)
Reading comprehension	8.40	-20.15**	18.88*	12.39**	-22.71**	17.92**	16.21**	-7.29	10.98	27.68***
	(19.44)	(7.40)	(8.80)	(4.49)	(7.77)	(7.63)	(6.27)	(6.79)	(6.62)	(2.87)
Writing	22.88	-14.69	13.57	5.25	-16.1*	12.19	11.28	-22.64***	12.98**	47.4***
	(33.40)	(10.75)	(8.04)	(8.52)	(9.15)	(7.94)	(7.32)	(7.09)	(5.15)	(3.58)
Observations	212	329	329	329	321	321	321	1568	1568	1568
Control for baseline			Y			Y				
Demographic control			N			Y				
Years			2018			2018			School fixed-effects	

Notes: the sample for models 1 and 2 consists of all observations with endline outcomes in 2018, while the sample model 3 consists of all observations with endline outcomes for 2018 and 2019. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Table 3: regression results of the causal effect of increased exposure to the treatment

		[4]	[4]
	1 year	2 years	2 years
Oral language	8.49***	2.86	-3.16
	(2.53)	(2.66)	(3.35)
Alphabetic principles	3.55	-5.63**	-8.50***
	(2.91)	(2.23)	(1.9)
Decoding	4.37	0.79	-2.75**
	(2.70)	(2.27)	(1.23)
Phonological awareness	-2.48	-16.38***	-15.58***
	(4.14)	(2.92)	(3.49)
Rapid automatized naming	2.04	-6.37**	-7.58***
	(2.70)	(2.24)	(1.67)
Reading fluency	2.94*	-0.49	-2.09*
	(1.61)	(1.49)	(0.97)
Reading comprehension	9.18*	0.37	-4.19
	(4.54)	(5.22)	(4.33)
Writing	8.60**	6.92*	2.80
	(3.87)	(3.51)	(2.87)
Observations	699	699	543
Control for baseline		Y	Y
Demographic control		N	N
Sample		Full sample	Only those exposed 1 or 2 years

Notes: the sample for model 4 consists of all observations with endline outcomes in 2018 and 2019. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Table 4: regression results of the causal effect of the intervention on grade repetition

	Baseline means and SD	[1]	[2]	Observations
Only 2018 sample	0.28	-0.08	-0.08	418
	(0.45)	(0.05)	(0.05)	
Only 2019 sample	0.33	-0.11***	-0.11**	450
	(0.47)	(-0.11)	(-0.11)	
Cohort 1 - 2018	0.18	-0.02	-0.02	233
	(0.39)	(-0.02)	(-0.02)	
Cohort 1 - 2019	0.32	-0.13**	-0.13**	216
	(0.47)	(0.03)	(0.04)	
Control for baseline		N	Y	

Notes: cohort 1 refers to students who were in grade 1 in 2018. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Table 5: regression results of the causal effect on literacy sub-skills, interacting treatment status with baseline performance

	[1]			[2]		
	Treatment x BL	Treatment	BL	Treatment x BL	Treatment	BL
Oral language	-0.06	9.83	0.73***	-0.01	5.65	0.67***
	(0.09)	(5.81)	(0.07)	(0.07)	(5.29)	(0.07)
Alphabetic principles	-0.10	5.80	0.93***	-0.12	4.80	0.91***
	(0.11)	(5.92)	(0.10)	(0.13)	(6.26)	(0.07)
Decoding	-0.20*	11.20*	0.87***	-0.18	9.35	0.81***
	(0.10)	(5.79)	(0.09)	(0.11)	(5.63)	(0.08)
Phonological awareness	-0.19	9.59	0.75***	-0.18*	6.24	0.70***
	(0.11)	(7.79)	(0.13)	(0.09)	(6.15)	(0.09)
Rapid automatized naming	-0.10	5.33	0.93***	-0.12	5.25	0.89***
	(0.09)	(3.95)	(0.07)	(0.11)	(4.68)	(0.09)
Reading fluency	-0.13	5.45	1.24***	-0.13	4.22	1.20***
	(0.11)	(3.45)	(0.07)	(0.10)	(3.16)	(0.07)
Reading comprehension	-0.37**	14.90**	1.02***	-0.34**	12.84**	0.89***
	(0.14)	(6.72)	(0.12)	(0.13)	(5.77)	(0.10)
Writing	-0.16	11.51	0.76***	-0.12	8.82	0.66***
	(0.11)	(7.53)	(0.08)	(0.10)	(6.95)	(0.07)
Observations		329			321	
Demographic control		N			Y	

Notes: Models 1 and 2 use all observations with endline outcomes in 2018. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Table 6: predictive power on reading comprehension improvements of changes in decoding and oral language, by baseline constraints

		Oral language					
		Low			High		
		Δ in decoding	Δ in oral language	Ratio coefficients Δ D/Δ L	Δ in decoding	Δ in oral language	Ratio coefficients Δ D/Δ L
Decoding	Low	0.29***	0.19**	1.54**	0.86***	0.30*	2.84*
		(0.06)	(0.06)	(0.54)	(0.08)	(0.14)	(1.38)
			n=214			n=142	
		Both skills are constraint			*Decoding is the main constraint*		
	High	0.29**	0.44***	0.65***	1.15***	0.70***	1.64***
		(0.10)	(0.08)	(0.23)	(0.12)	(0.07)	(0.24)
			n=119			n=224	
		Oral language is the main constraint			*No skill is a major constraint*		

Notes: all models use pooled 2018 and 2019 data. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Appendix A: Business-as-usual education in this context

As a way to contextualize this intervention and the results later discussed, I qualitatively describe the status quo learning profiles of students in this context, using the students in the control group during the first year of data collection. I complement this description with *Appendix Table 1*, which shows the baseline, endline, and growth scores by sub-skill and grade, using the outcomes described in the previous section. In this sense, *Appendix Table 1* also serves as a tool to understand what an average "year of schooling" (by sub-skill) means in this context. In turn, this then allows for more standardized comparisons of the regression coefficients across sub-groups, and also with other studies.

The first skill that students were tested on was their comprehension of oral language. On average, incoming first grade students were able to understand 81% of all oral instructions. However, there was a 10-percentage point gap between indigenous and Ladino students (74% vs. 84%). This gap likely represents that at the beginning of 2018, 41% of all students report speaking primarily a language other than Spanish at home. Students for which first grade was their first formal education experience, and that speak primarily a non-Spanish language at home, only understood 45% of all oral instructions, displaying a clear lack of familiarity with Spanish, the testing language, and language of instruction for 86% of the sample. In terms of alphabetic principles, students start grade 1 with very little knowledge, likely reflecting that only about 23% of all students in the sample went to kindergarten. For instance, students were only able to recognize on average 8% of all letter names coming into grade 1, and 24% coming out of grade 1. By the time they leave grade 2, they are able to recognize 55% of all letter grades. For letter sounds, the situation is similar: by the end of grade 1 students only knew 34% of all letter sounds, and 52% by the end of grade 2. Interesting, the rate at which students know the name of each letter varies substantially, across letters but also across cases of the same letter. The letter "a" is recognized in both its upper- and lower-case by 9 in 10 students. The letter "w" is known by about 1 in 4 students in both cases. However, a letter like "i" is known by 9 in 10 in its lower-case form, but fewer than half know it in its upper-case form.

Broadly speaking, the decoding skills for this sample were almost inexistent by the time they come into grade 1. However, by the time they start grade 2, they can read 27% of familiar words and 23% of "nonsense words" (23%). By the end of grade 2, these rates go to 44% and 34% respectively, displaying a widening gap between sight words (instinctive knowledge of familiar words) and decoding skills in isolation regardless of whether each word has a real meaning or not. Much like the differences in alphabetic knowledge of different letters, the rate at which students knew familiar words varied, even for words with similar letters and number of syllables. For instance, at the end of grade 1 and for two-syllable words, 61% of control students knew the word "casa" (house), but only 31% knew the word "pollo" (chicken). The situation is similar for three syllable words: 41% knew the

word “bonito” (pretty), but only 18% knew the word “banano” (banana). These differential rates in letter and word recognition also serve as a cautionary note in the design of rapid literacy assessments with only a few items in similar contexts (for instance, Medición Independiente de Aprendizajes in México), as the inclusion of certain letters or words may yield very different estimates of children's emergent literacy levels.

In terms of the three higher order skills (reading fluency, comprehension and writing), students came into grade 1 with virtually no achievement in these areas either. However, for reading fluency, by the time they finished grade 2, students were reading 31 correct words per minute (cwpm). These rates are low in general and even within Guatemala, but fairly on par or higher than other assessments in developing countries. For example, grade 2 students in a nationally representative in Guatemala sample in 2013 scored 48 cpwm (Del Valle et al., 2017). Abadzi, mentions that to achieve proper reading comprehension levels, students should read about 30 cwpm by the end of grade 1, and 45-60 cwpm by the end of grade 2. However, Abadzi finds that this is not the case in the vast majority of studies across 29 developing countries included in their report (Abadzi, 2011). In the same way, Graham and Kelly (2019) report that other early grade reading interventions have worked with populations comprised of 40% zero-word readers in second grade in Liberia and Yemen, and over 80% in Nepal or Malawi. In this sample, the share of zero-word readers is barely under 10%. Of the 25 program-language means reported by Graham and Kelly (2019), only 3 were above the 31 cwpm by endline in second grade that the control group in my sample achieves.

In terms of reading comprehension, following standard EGRA and ELGI procedures, students were asked to read a short passage, and were then asked simple comprehension questions. By the end of the year, students in grade 1 in this full sample scored 2.7 out of 11 questions right (24.4%), and students in grade 2 scored 5.5 questions right out of 11 (50.2%). By the end of grade 2 among the full sample, only 4% of all grade 2 students got all questions right. For perspective, the average Guatemalan grade 2 student got 7.9 questions right on the same assessment in 2013, and 36% of students got all questions right (Del Valle et al., 2017). However, these rates are higher than what was found in other samples in developing countries[25] (Gove and Cvelich, 2010). Finally, in terms of writing, students without the intervention were able to write no words at the beginning of grade 1, but were able to write almost 3 words out of 6 by the time they started grade 2. By the end of grade 2, they were able to write on average 4 full words.

[25] For example, the share of students scoring above 80% in reading comprehension by grade 2 was 28% in this sample, twice as high as the highest sample (Kenya, Central Provice, Gikuyu) reported in Gove and Cvelich (2010).

Appendix B: Creation of outcomes

Given the fine-grain, item-level data available for the reading outcomes, I need to aggregate and standardize these data to then use them in meaningful analyses. Each item belongs to a specific, pre-determined section, and these sections are then aggregated into the eight broader literacy skills tested. This aggregation process followed the conventional ELGI aggregation of items displayed in *Appendix Table 0* below.

Appendix Table 0: data available for each child at baseline and endline

Skill	**Sub-test**
Oral Language	Section 1: Comprehension of oral instructions
	Section 7: Listening comprehension
Alphabetic Principles	Section 2.1: Letter name recognition
	Section 3.1: Letter sound recognition
Decoding	Section 5.1: Reading short words
	Section 6: Speed for reading nonsense words
Phonological Awareness	Section 4.1: Initial phoneme identification
	Section 4.2: Phoneme segmentation
Rapid Automatized Naming	Section 2.2: Speed for identifying letter
	Section 3.2: Speed for identifying letter sounds
Reading Fluency	Section 5.2: Speed for reading familiar words
	Section 8.1: Speed for reading a passage
Reading Comprehension	Section 9: Reading comprehension
Writing (Dictation)	Section 10: Writing (Dictation)

Within each section, I first express each outcome as a percentage of correct answers out of all questions asked in this section. For skills that involve more than two sections (as outlined in *Appendix Table 0*), such as oral language or alphabetic principles, the total score for the skill is the average across both percentages for each section. The measure of oral language is comprised of 11 questions about comprehension of oral instructions such as "take that pencil, point it at the door, and then give it to me" and 10 listening comprehension questions. These questions are about a short oral story told by the assessor, are in the form of "who were the characters in this story?". The alphabetic principles skill is measured through 60 questions about letter names like "what is the name of this letter [showing a printed letter "A"]?" and 60 questions about the sound of each name. The decoding skills are measured by asking children to read 10 short words such as "dos" (two), and 70 nonsense words like "foba" (which does not mean anything). For phonological awareness, the first section was the percentage of initial phonemes identified out of 10 words. The second section was the percentage of phonemes correctly separated among 10 possible

words, for example "/c/l/i/m/a" in "clima" (weather). For rapid automatized naming, the two sections were the percentage number of letters named during a timed test of 60 seconds, and the percentage number of letter sounds recognized during a timed test of 60 seconds

The reading fluency is measured by first asking students to read 70 familiar words like "pollo" (chicken), and then through the more standardized reading fluency measure "correct words per minute" (cwpm), measured as the number of correctly read words, over the time spent on the exercise, divided by 60 to standardize the time unit as minutes. The reading comprehension skill was measured by showing students a short, written story, and asking them 11 simple comprehension questions such as "who was looking for the cat?". Finally, for the writing section, students were asked to write 6 words which were dictated to them ("mi escuela es limpia y bonita"/"my school is clean and cute"). The number of correct words was then recorded. Examples of what students wrote as partially incorrect answers were "Mi escuela lipia i bonita" (3 correct words), "mi es cuela es limpio y vonitas" (2 correct words), and "mipseue" (no correct words), showing the different degrees of phonological translation into print that exist in the sample.

For the recreation of the ASER/Uwezo outcomes, I used the sections of the test which broadly resembled the most those questions included in the ASER/Uwezo tests. I also followed a similar classification rule: if a student achieves more than half within each level, then I mark them as having achieved that level. In particular, if a student got more than half of all letter names recognized, they achieve the "letter" level. If a student read more than half of all short words (such as "dos") right, they achieve the "syllable" level. If they read more than half of all familiar words (such as "casa"), they achieve the "word" level. If they achieve over a 0% but below a 50% in reading comprehension, they achieve the "sentence" level, and if they scored over "50" in reading comprehension, they achieve they "story" level. I then take the higher ASER/Uwezo level achieved by a student and categorize them as such. Finally, I also create a "early grade learning deprivation" outcome in the spirit of the "learning poverty" concept most aligned with the Sustainable Development Goals 4.1.1a. In particular, a student was classified as "learning deprived if they scored less than 80% in the reading comprehension task".

Appendix C: Additional tables and figures

Appendix Table 1: student growth in control schools in 2018

	Both grades			Grade 1			Grade 2		
	Baseline (BL)	Endline (EL)	EL-BL	Baseline (BL)	Endline (EL)	EL-BL	Baseline (BL)	Endline (EL)	EL-BL
Oral language	49.52	55.29	5.77	47.74	49.55	1.81	51.84	63.75	11.91
	(21.19)	(22.62)		(20.63)	(22.9)		(21.78)	(19.47)	
Alphabetic principles	27.85	44.61	16.76	14.42	34.27	19.85	45.37	59.87	14.50
	(21.08)	(24.41)		(9.9)	(19.79)		(18.82)	(22.63)	
Decoding	25.17	46.98	21.81	5.67	38.52	32.85	50.61	59.48	8.87
	(28.45)	(27.43)		(9.78)	(25.77)		(24.42)	(25.09)	
Phonological awareness	38.23	66.35	28.12	22.71	56.56	33.85	58.48	80.79	22.31
	(27.97)	(32.25)		(16.93)	(32.07)		(26.59)	(26.81)	
Rapid automatized naming	22.70	39.80	17.10	9.67	28.80	19.13	39.70	56.03	16.33
	(20.85)	(24.76)		(8.6)	(18.24)		(19.87)	(24.29)	
Reading fluency	11.72	26.52	14.80	0.44	14.28	13.84	26.43	44.60	18.17
	(18.94)	(25.72)		(1.11)	(13.67)		(21.05)	(28.62)	
Reading comprehension	8.40	26.34	17.94	0.00	14.27	14.27	19.37	44.16	24.79
	(19.44)	(30.74)		(0)	(22.6)		(25.72)	(32.61)	
Writing	22.88	43.70	20.82	2.36	28.14	25.78	49.64	66.67	17.03
	(33.4)	(35.79)		(8.17)	(29.49)		(34.94)	(31.82)	
Observations	212			120			92		

Notes: the sample consists of all children in control schools at baseline and endline in 2018. Standard deviations are shown in parenthesis.

Appendix Table 2: Balance between treatment and control students at baseline

		Control group	T-C
Student demographics	Grade	1.43	0.02
		(0.50)	(0.04)
	Age at baseline	7.87	0.08
		(1.10)	(0.17)
	Female	0.52	-0.06
		(0.5)	(0.05)
	Speaks indigenous language at home	0.45	-0.08*
		(0.50)	(0.04)
	First time in first grade?	0.86	0.00
		(0.35)	(0.03)
Academic history	Went to kindergarten	0.23	-0.03
		(0.42)	(0.04)
	Went to pre--K	0.56	0.00
		(0.50)	(0.06)
	Went to ECE	0.75	0.04
		(0.43)	(0.06)
	Is currently repeating a grade	0.28	-0.08
		(0.45)	(0.05)
	Has a male teacher	0.13	0.08
		(0.33)	(0.14)
	Receives instruction in Spanish	0.83	0.07*
		(0.37)	(0.04)
Attitudes and behaviors	Reads stories in class	0.85	-0.01
		(0.36)	(0.04)
	Likes reading	0.88	-0.06
		(0.32)	(0.04)
	Likes writing	0.89	0.02
		(0.32)	(0.03)
	Was assigned hw this week	0.96	-0.03
		(0.19)	(0.03)
	Received help at home with hw	0.70	-0.14***
		(0.46)	(0.04)
	Has read at home this week	0.86	0.03
		(0.34)	(0.05)
	Has written at home this week	0.95	-0.01
		(0.22)	(0.03)
Academic achievement at baseline	Oral language	49.52	13.91**
		(21.19)	(5.21)
	Alphabetic principles	27.85	5.81
		(21.08)	(3.40)
	Decoding	25.17	6.31
		(28.45)	(4.79)
	Phonological awareness	38.23	10.22*
		(27.97)	(4.85)
	Rapid automatized naming	22.70	7.17*
		(20.85)	(3.57)
	Reading fluency	11.72	5.29
		(18.94)	(3.26)
	Reading comprehension	8.40	8.85**
		(19.44)	(3.15)
	Writing	22.88	9.49*
		(33.4)	(5.34)
	Observations		212

Notes: results obtained from regressing each covariate on an indicator variable for treatment schools. Sample consists of all children at baseline in 2018. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Appendix Table 3: regression results of the causal effect of treatment on literacy sub-skills, using randomization inference

	Both grades		Grade 1		Grade 2	
	[1]	[2]	[3]	[4]	[5]	[6]
Oral language	6.21**	5.06**	8.28**	7.81*	3.27	-0.07
	(3.77)	(3.16)	(4.54)	(4.87)	(4.17)	(3.46)
Alphabetic principles	2.68	1.06	3.06	1.57	2.02	-1.15
	(3.56)	(3.48)	(4.92)	(5.19)	(2.88)	(3.04)
Decoding	5.79*	4.47*	9.29*	8.61*	2.55	1.90
	(3.74)	(3.43)	(6.42)	(6.66)	(2.14)	(2.08)
Phonological awareness	1.32	-1.15	0.01	-1.75	1.89	-3.06
	(4.38)	(4.13)	(6.86)	(7.13)	(4.31)	(5.89)
Rapid automatized naming	2.85*	2.11	4.68	4.27	0.20	-1.72
	(2.44)	(2.31)	(4.17)	(4.44)	(1.23)	(1.71)
Reading fluency	3.73**	2.47	6.21*	6.18*	-0.01	-1.64
	(2.27)	(2.34)	(4.25)	(4.45)	(2.24)	(2.22)
Reading comprehension	10.72*	9.03*	19.68**	18.69**	-1.84	-5.80
	(6.49)	(6.46)	(10.27)	(10.98)	(4.79)	(6.22)
Writing	7.42*	5.69	14.76*	13.33*	-1.53	-2.19
	(4.73)	(5.01)	(9.49)	(8.95)	(4.37)	(4.29)
Observations	329	321	190	185	139	136
Control for baseline	Y	Y	Y	Y	Y	Y
Demographic control	N	Y	N	Y	N	Y

Notes: the sample for models 1 and 2 consist of all observations with endline outcomes in 2018, while the sample for model 3 consists all observations with endline outcomes for 2018 and 2019. Standard errors and significance estimated through randomization inference, where each outcome is simulated and re-estimated 500 times. Standard errors are shown in parenthesis. Within each individual regression, the standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Appendix Table 4: comparison of treatment effects estimated through a regular regression model and inverse probability weighting (IPW) model

	[1]		[2]	
	OLS	IPW	OLS	IPW
Oral language	6.21**	6.54***	5.06	5.25**
	(2.63)	(1.59)	(2.99)	(1.81)
Alphabetic principles	2.68	2.43	1.06	1.08
	(2.82)	(1.56)	(3.12)	(1.54)
Decoding	5.79*	5.72**	4.47	4.57**
	(2.94)	(2.04)	(3.18)	(2)
Phonological awareness	1.32	1.39	-1.15	-1.03
	(3.82)	(2.49)	(3.68)	(2.6)
Rapid automatized naming	2.85	2.29	2.11	1.66
	(2.31)	(1.64)	(2.36)	(1.72)
Reading fluency	3.73*	3.68**	2.47	2.22
	(2.11)	(1.43)	(1.95)	(1.55)
Reading comprehension	10.72	10.09***	9.03*	7.95**
	(6.16)	(2.82)	(4.5)	(3.09)
Writing	7.42*	7.22**	5.69	5.57*
	(4.14)	(2.9)	(4.87)	(2.86)
Observations	329	329	321	321
Control for baseline	Y	Y	Y	Y
Demographic control	N	N	N	Y

Notes: Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Appendix Table 5: regression results of the causal effect of treatment on attitudes and behaviors

	Mean and SD at baseline	[1]	[2]	[3]
Likes reading	0.88	0.10	0.09	0.05
	(0.32)	(0.06)	(0.07)	(0.04)
Likes writing	0.89	0.11	0.10	0.05
	(0.32)	(0.07)	(0.07)	(0.05)
Was assigned hw this week	0.96	0.10*	0.10*	0.06
	(0.19)	(0.06)	(0.05)	(0.04)
Received help at home with hw	0.70	-0.04	-0.05	0.00
	(0.46)	(0.06)	(0.06)	(0.05)
Has read at home this week	0.86	0.14**	0.13**	0.10**
	(0.34)	(0.05)	(0.06)	(0.04)
Has written at home this week	0.95	0.08	0.07	0.09**
	(0.22)	(0.05)	(0.04)	(0.04)
Observations	212	419	404	1738
Control for baseline		Y	Y	
Demographic control		N	Y	School fixed-effects
Years		2018	2018	

Notes: the sample for models 1 and 2 consists of all observations with endline outcomes in 2018, while the sample for model 3 consists of all observations with endline outcomes for 2018 and 2019. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Appendix Table 6: regression results of the causal effect of treatment on literacy sub-skills, for grade 1

	Mean and SD at baseline	[1]	[2]	[3]
Oral language	47.74	8.28**	7.81*	5.19
	(20.63)	(3.00)	(3.95)	(3.67)
Alphabetic principles	14.42	3.06	1.57	4.69
	(9.90)	(4.86)	(4.73)	(5.00)
Decoding	5.67	9.29*	8.61	6.40
	(9.78)	(5.17)	(5.88)	(5.81)
Phonological awareness	22.71	0.01	-1.75	-2.92
	(16.93)	(5.57)	(6.15)	(6.28)
Rapid automatized naming	9.67	4.68	4.27	5.60
	(8.60)	(3.64)	(4.01)	(4.11)
Reading fluency	0.44	6.21	6.18*	4.80
	(1.11)	(3.92)	(3.2)	(3.96)
Reading comprehension	0.00	19.68**	18.69**	12.69*
	(0.00)	(6.72)	(7.58)	(7.02)
Writing	2.36	14.76	13.33*	9.95
	(8.17)	(9.47)	(7.10)	(7.66)
Observations	120	190	185	829
Control for baseline		Y	Y	
Demographic control		N	Y	School fixed-effects
Years		2018	2018	

Notes: the sample for models 1 and 2 consists of all observations with endline outcomes in 2018, while the sample model 3 consists of all observations with endline outcomes for 2018 and 2019. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Appendix Table 7: regression results of the causal effect of treatment on literacy sub-skills, for grade 2

	Mean and SD at baseline	[1]	[2]	[3]
Oral language	51.84	3.27	-0.07	3.48*
	(21.78)	(3.77)	(3.34)	(1.88)
Alphabetic principles	45.37	2.02	-1.15	-0.39
	(18.82)	(2.58)	(3.41)	(2.08)
Decoding	50.61	2.55	1.9	0.61
	(24.42)	(2.1)	(2.07)	(1.88)
Phonological awareness	58.48	1.89	-3.06	-0.86
	(26.59)	(4.58)	(5.39)	(2.77)
Rapid automatized naming	39.70	0.2	-1.72	-1.98
	(19.87)	(1.44)	(2.24)	(1.94)
Reading fluency	26.43	-0.01	-1.64	-3.41***
	(21.05)	(2.22)	(2.74)	(0.9)
Reading comprehension	19.37	-1.84	-5.8	-2.41
	(25.72)	(5.72)	(4.1)	(4.1)
Writing	49.64	-1.53	-2.19	-3.51
	(34.94)	(4.32)	(4.78)	(2.25)
Observations	92	139	136	739
Control for baseline		Y	Y	
Demographic control		N	Y	School fixed-effects
Years		2018	2018	

Notes: the sample for models 1 and 2 consists of all observations with endline outcomes in 2018, while the sample model 3 consists of all observations with endline outcomes for 2018 and 2019. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Appendix Table 8: regression results of the causal effect of treatment on literacy sub-skills, interacting treatment status with baseline characteristics

	[1]			[2]			[3]			[4]			[5]		
	Tmnt x age	Tmnt	Age	Tmnt x Female	Tmnt	Female	Tmnt x Indigen	Tmnt	Indigen	Tmnt x Male teacher	Tmnt	Male teacher	Tmnt x Inst Span	Tmnt	Instruction Spanish
Oral language	-2.77	27.84*	1.45	0.49	5.92	-0.60	-3.14	7.47*	3.40	2.13	5.47**	6.58**	-0.25	6.44	1.83
	(1.71)	(14.91)	(1.78)	(4.00)	(4.31)	(3.13)	(4.81)	(4.19)	(3.95)	(4.07)	(2.47)	(2.23)	(7.06)	(6.61)	(5.80)
Alphabetic principles	1.26	-7.10	-0.28	4.78	-0.05	-7.20***	-0.06	2.51	-2.84	-2.21	2.87	3.46	3.51	-0.50	0.12
	(1.68)	(15.67)	(1.17)	(3.54)	(3.94)	(1.68)	(2.67)	(3.81)	(1.77)	(6.63)	(3.32)	(5.08)	(4.10)	(4.28)	(3.43)
Decoding	-0.77	11.74	-0.36	0.70	4.97	-6.33**	1.09	5.31	-1.13	-9.74	7.17*	7.45	-0.26	6.04	-1.72
	(2.09)	(18.85)	(1.80)	(4.99)	(3.84)	(2.65)	(4.48)	(3.87)	(2.86)	(5.69)	(3.64)	(4.90)	(6.07)	(4.59)	(5.86)
Phonological awareness	0.30	-1.06	-1.01	1.78	0.01	-6.91	7.72	-1.99	-6.90**	1.81	0.94	0.51	-2.36	3.42	2.79
	(2.64)	(23.34)	(1.84)	(6.71)	(5.75)	(4.82)	(5.24)	(4.54)	(2.67)	(11.16)	(3.53)	(7.29)	(5.23)	(4.55)	(4.57)
Rapid automatized naming	-1.14	11.71	0.53	6.89*	-0.88	-7.78***	-0.11	2.79	-1.51	-4.22	3.48	2.72	0.30	2.58	-0.69
	(1.90)	(16.4)	(1.42)	(3.36)	(2.81)	(2.21)	(3.17)	(2.86)	(2.44)	(3.46)	(2.57)	(2.76)	(4.71)	(4.54)	(4.21)
Reading fluency	-2.03	19.44	0.25	5.95	0.45	-7.30***	-1.48	4.33	0.84	-2.94	4.17	2.09	-0.23	3.95	-1.32
	(1.41)	(11.95)	(1.22)	(3.58)	(3.38)	(1.99)	(3.39)	(2.48)	(2.11)	(4.74)	(2.62)	(2.99)	(7.21)	(6.51)	(4.41)
Reading comprehension	-5.93**	56.61**	-0.09	9.81	5.68	-6.33	-1.13	11.01	-2.62	4.41	9.71	2.53	-3.42	13.78	-1.20
	(2.32)	(21.67)	(2.3)	(6.37)	(6.66)	(3.73)	(8.21)	(7.56)	(5.17)	(14.21)	(5.54)	(9.12)	(9.15)	(8.31)	(6.49)
Writing	-4.44	41.93	1.43	5.79	4.02	-9.63***	1.11	6.68	-4.97	-2.81	8.14	-2.44	3.37	4.40	-1.80
	(2.89)	(25.18)	(2.53)	(3.97)	(4.38)	(3.03)	(5.26)	(4.37)	(4.19)	(13.75)	(5.52)	(9.49)	(12.47)	(12.75)	(5.33)

Notes: the sample for all models consists of all observations with endline outcomes in 2018. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Appendix Table 9: balance table with two models exploring differential attrition

		[1]	[2]		
			Treatment x		
		Variable	Variable	Treatment	Variable
Student demographics	Grade	0.07*	-0.04	-0.04	0.09
		(0.04)	(0.09)	(0.13)	(0.06)
	Age at baseline	0.07***	0.06	-0.55	0.04*
		(0.02)	(0.04)	(0.33)	(0.02)
	Female	-0.04	0.06	-0.13	-0.08
		(0.04)	(0.08)	(0.09)	(0.07)
	Speaks indigenous language at home	0.05	-0.02	-0.09	0.05
		(0.05)	(0.10)	(0.07)	(0.07)
	First time in first grade?	-0.14	-0.13	0.01	-0.08
		(0.08)	(0.15)	(0.12)	(0.13)
Academic history	Went to kindergarten	0.03	-0.21***	-0.05	0.13***
		(0.04)	(0.06)	(0.04)	(0.03)
	Went to pre--K	-0.01	-0.02	-0.08	-0.01
		(0.05)	(0.09)	(0.07)	(0.06)
	Went to ECE	-0.06	-0.16	0.03	0.02
		(0.06)	(0.10)	(0.10)	(0.04)
	Is currently repeating a grade	0.09**	0.11	-0.12**	0.03
		(0.04)	(0.12)	(0.05)	(0.05)
	Has a male teacher	-0.02	0.00	-0.10	-0.01
		(0.06)	(0.11)	(0.07)	(0.10)
	Receives instruction in Spanish	-0.19***	0.32**	-0.36***	-0.3***
		(0.06)	(0.11)	(0.08)	(0.05)
Attitudes and behaviors	Reads stories in class	0.02	0.08	-0.16	-0.02
		(0.05)	(0.12)	(0.11)	(0.08)
	Likes reading	-0.09*	-0.07	-0.04	-0.06
		(0.04)	(0.1)	(0.13)	(0.08)
	Likes writing	-0.14**	-0.03	-0.08	-0.13
		(0.05)	(0.14)	(0.17)	(0.08)
	Was assigned hw this week	0.13***	-0.05	-0.05	0.14
		(0.04)	(0.09)	(0.11)	(0.09)
	Received help at home with hw	-0.01	0.00	-0.11	-0.02
		(0.03)	(0.06)	(0.06)	(0.04)
	Has read at home this week	-0.05	0.03	-0.13	-0.05
		(0.06)	(0.12)	(0.12)	(0.10)
	Has written at home this week	0.04	0.09	-0.19	-0.01
		(0.07)	(0.18)	(0.18)	(0.16)
Academic achievement at baseline	Oral language	0.00***	0.00	-0.02	0.00*
		(0.00)	(0.00)	(0.07)	(0.00)
	Alphabetic principles	0.00	0.00	-0.14*	0.00
		(0.00)	(0.00)	(0.07)	(0.00)
	Decoding	0.00	0.00	-0.13*	0.00
		(0.00)	(0.00)	(0.06)	(0.00)
	Phonological awareness	0.00	0.00	-0.11	0.00
		(0.00)	(0.00)	(0.08)	(0.00)
	Rapid automatized naming	0.00	0.00	-0.13	0.00
		(0.00)	(0.00)	(0.09)	(0.00)
	Reading fluency	0.00	0.00	-0.12*	0.00
		(0.00)	(0.00)	(0.07)	(0.00)
	Reading comprehension	0.00	0.00	-0.11	0.00
		(0.00)	(0.00)	(0.06)	(0.00)
	Writing	0.00	0.00	-0.11	0.00
		(0.00)	(0.00)	(0.07)	(0.00)
	Observations	419	419	419	419

Notes: the first model [1] displays the result of regressing each covariate on a binary variable indicating whether each student is missing their endline scores. The second model [2] shows the same model as [1], including an interaction effect between each covariate and the treatment status. Standard errors are shown in parenthesis. The sample for all models consists of all students present at baseline in 2018. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Appendix Table 10: regression results bounding estimated causal effects with three different approaches

	Main estimate	Manski's Bounds [1]		Manski's Bounds [2]		Lee's Bounds [3]	
		LB	UP	LB	UP	LB	UP
Oral language	6.21**	1.86	8.80***	1.19	9.17***	12.93***	20.50***
	(2.57)	(2.69)	(2.15)	(2.42)	(1.86)	(2.65)	(2.72)
Alphabetic principles	2.68	0.14	3.12	-0.08	3.27	2.77	11.94***
	(3.47)	(2.36)	(2.85)	(2.39)	(2.43)	(2.67)	(3.16)
Decoding	5.79*	1.98	5.97*	2.25	4.05	5.95*	16.44***
	(2.92)	(2.62)	(3.05)	(2.50)	(2.40)	(3.01)	(3.69)
Phonological awareness	1.32	-0.48	2.89	-1.88	2.38	4.44	14.76***
	(3.91)	(3.60)	(3.75)	(3.18)	(3.24)	(4.17)	(3.36)
Rapid automatized naming	2.85	0.21	3.77	-0.64	4.58*	4.06	13.45***
	(2.21)	(2.02)	(2.36)	(1.96)	(2.27)	(3.47)	(3.63)
Reading fluency	3.73*	0.5	4.18	0.69	5.17**	1.95	13.61***
	(2.11)	(2.3)	(2.71)	(2.64)	(2.15)	(3.59)	(3.47)
Reading comprehension	10.72	4.25	11.47*	0.89	16.21**	10.44**	23.02***
	(7.36)	(4.67)	(5.48)	(4.62)	(6.43)	(3.90)	(4.61)
Writing	7.42*	2.48	8.02*	-1.75	12.30**	7.28	20.68***
	(4.05)	(3.47)	(4.06)	(4.68)	(4.53)	(4.26)	(4.89)
Observations	329	419	419	419	419	419	419

Notes: all models use only 2018 data. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

Appendix Table 11: regression results of the causal effect of treatment on ASER-like literacy sub-skills

	Mean and SD at baseline	[1]	[2]	[3]
Letter level	0.40	0.12**	0.09	0.16**
	(0.49)	(0.05)	(0.05)	(0.06)
Syllable level	0.38	0.10*	0.08*	0.14**
	(0.49)	(0.05)	(0.04)	(0.06)
Word level	0.13	0.13	0.11	0.09
	(0.34)	(0.09)	(0.09)	(0.06)
Sentence level	0.12	0.13	0.11	0.10
	(0.32)	(0.09)	(0.07)	(0.06)
Story level	0.08	0.15*	0.13	0.11*
	(0.28)	(0.08)	(0.10)	(0.06)
Learning poverty	0.98	-0.12	-0.10	-0.07
	(0.14)	(0.07)	(0.06)	(0.05)
Observations	212	329	321	1568
Control for baseline		Y	Y	School fixed-effects
Demographic control		N	Y	
Years		2018	2018	

Notes: the sample for models 1 and 2 consists of all observations with endline outcomes in 2018, while the sample for model 3 consists of all observations with endline outcomes for 2018 and 2019. Standard errors are shown in parenthesis. Standard errors were bootstrapped and clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01