


Education Experiments in Latin America: Empirical Evidence to Guide Evaluation Design

Evaluation Review
2024, Vol. 0(0) 1–32
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0193841X241241354
journals.sagepub.com/home/erx


**Steven Glazerman¹, Larissa Campuzano² , and
Nancy Murray²**

Abstract

Randomized experiments involving education interventions are typically implemented as cluster randomized trials, with schools serving as clusters. To design such a study, it is critical to understand the degree to which learning outcomes vary between versus within clusters (schools), specifically the intraclass correlation coefficient. It is also helpful to anticipate the benefits, in terms of statistical power, of collecting household data, testing students at baseline, or relying on administrative data on previous cohorts from the same school. We use data from multiple cluster-randomized trials in four Latin American countries to provide information on the intraclass correlations in early grade literacy outcomes. We also describe the proportion of variance explained by different types of covariates. These parameters will help future researchers conduct statistical power analysis, estimate the required sample size, and determine the necessity of collecting different types of baseline data such as child assessments, administrative data at the school level, or household surveys.

Keywords

statistical power, intraclass correlation, experimental design, education evaluation

¹Innovations for Poverty Action, Washington, DC, USA

²Mathematica Inc, Washington, DC, USA

Corresponding Author:

Larissa Campuzano, Mathematica Inc, 1100 First Street, NE, 12th Floor, Washington, DC 20002-4221, USA.

Email: LCampuzano@mathematica-mpr.com

Keywords

JEL Classifications Numbers:, C8, C9, I2

Many educational interventions need to be implemented at the school level; for example, professional development programs are often offered to all teachers in a school to ensure that the teachers' instruction is similar. In these cases, experimental evaluations conduct random assignment at the school level. Because groups of students clustered in schools are the unit of assignment, this design is referred to as a cluster-randomized control trial (RCT). Randomizing schools guarantees that the offer of the intervention is independent of observed and unobserved school and student characteristics, but because students, teachers, and other resources are not randomly allocated into schools, using the school as the unit of assignment creates a new problem: statistical imprecision. Generating precise estimates of model parameters requires a large number of clusters.

Recent studies have focused on how to design cluster-randomized studies that provide adequate precision (Spybrook & Kelcey, 2016; Westine et al., 2013). Researchers use statistical power calculations (Duflo et al., 2008) to estimate optimal sample sizes and thus can design an efficient experiment (Hedges & Hedberg, 2007; Schochet, 2008). Power calculations for cluster-randomized trials need information on the degree to which outcomes vary between schools versus within schools. The statistic used to summarize this is the intraclass correlation (ICC). Previous research has provided information on the values of ICCs for education in the United States (Hedges & Hedberg, 2007, 2013; Schochet, 2008; Shen et al., 2023), in Germany (Stallasch et al., 2021), in sub-Saharan Africa (Kelcey et al., 2016; Seidenfeld et al., 2023), and in a sample of 81 countries (Brunner et al., 2018). But this information is rarely available for other regions, so researchers often make assumptions that turn out to be incorrect. In this paper, we provide information on the possible values of ICCs for education interventions based on studies conducted in Latin America.

Researchers often collect baseline data or time-invariant background data to use as covariates in their impact estimation models to increase the precision of the study. Including covariates in the statistical model to estimate the impact reduces the error variance and increases the precision of the parameter estimates (Raudenbush et al., 2007). Deciding which covariates to include in the impact models should account for both precision gains and the cost of measuring (collecting data on) those covariates.

Evaluations of education interventions often include baseline assessments as covariates in the impact estimation models to increase the statistical precision. However, collecting these assessments can be costly. For example, child assessments may have to be administered individually. Even group-administered

cognitive tests require trained proctors and take valuable time away from instruction. Household surveys may provide information on parents' education or socio-economic information that could help increase the precision of the study. But collecting household information requires extra travel to communities in addition to school visits, and may require respondent payments to ensure sufficiently high response rates. Other forms of data may be less costly to collect, such as administrative data on student or school performance, but such data may not be available at the right level of disaggregation or may require considerable effort to obtain and clean for research purposes.

Empirical researchers need detailed information at the planning stage to forecast the degree to which a given type of baseline data will increase the precision of model parameters and to assess the cost-variance tradeoff of each type of data collection. The summary statistic that captures this information is the R^2 from a regression of the outcome of interest on covariates. Previous research has provided information on the precision gains of including different sets of covariates in the U.S. context (Bloom et al., 2007; Hedges & Hedberg, 2007; Kelcey & Phelps, 2013; Konstantopoulos, 2012; Raudenbush et al., 2007). This paper provides information on values of R^2 based on evaluations conducted in Honduras, Guatemala, Nicaragua, and Peru.

The goal of obtaining these parameters and conducting power analysis is to plan the optimal study design that allocates sample most efficiently between number of clusters (schools) and cluster size (number of students per school) as well as the total sample size. The researcher seeks to minimize costs subject to a constraint of being able to detect a given impact with some acceptable level of statistical power (Raudenbush, 1997; Shen & Kelcey, 2020). Once we set the power level (typically 80 or 90%) and minimum detectable effect size (MDES), and we input the marginal costs of collecting data per student and per school, we can compute the optimal sample size in terms of both students and schools.

The ICC and R^2 are critical inputs to this exercise, but such variance parameters might depend on the context. It helps to think of why the ICC or covariate model R^2 might be high or low. For example, in some rural areas, schools might be located in ethnic communities of different linguistic and cultural features that influence literacy in the national language. Such locations also might have different pay or working conditions that influence their ability to attract good school teachers or principals. In these cases, variation between schools may be high and so the ICC will be high, which makes it more challenging to conduct an efficient cluster-randomized trial, but a high R^2 (or stratification of schools on dominant ethnicity or language) might mitigate the problem.

Alternatively, an urban setting might have many schools that draw from a similar population of students and prospective teachers and have centrally determined policies like class size and budgets, resulting in low between-

school variance, and hence a low ICC. This type of study setting would permit the researcher to conduct the study with fewer schools and therefore lower costs. Whether the study sample is national or regional can also change the ICCs. [Hedges and Hedberg \(2013\)](#) compared ICCs for math and reading outcomes from national U.S. samples to samples by state and found that, for elementary grades, states have lower ICCs than national samples.

In this paper, we raise the issue of data collection costs to point out that some covariates are more costly to collect than others, so their contribution to statistical precision should be weighed against those costs. However, the discussion is illustrative, and it does not attempt to obtain an optimal sample that minimizes cost and attains the largest statistical power. A framework which can be used for optimal allocation is presented in [Shen & Kelcey, \(2020\)](#); [Shen et al., \(2023\)](#) present an example from an optimal design perspective.

Because the required parameters for planning a cluster-RCT may depend on the context, it is important for social scientists to share findings from as many outcomes and study contexts as possible for the benefit of future researchers. Several researchers have documented estimates of these key parameters for education interventions. Most of these published parameters are based on data from the United States ([Hedberg & Hedges, 2014](#); [Hedges & Hedberg, 2007, 2013](#); [Jacob et al., 2010](#); [Schochet, 2008](#)). [Kelcey et al. \(2016\)](#) reported ICC and multilevel R^2 results from RCTs in 15 sub-Saharan African countries, [Stallasch et al. \(2021\)](#) reported parameters for Germany, and [Brunner et al. \(2018\)](#) reported parameters for a sample of 81 countries. However, we are not aware of published estimates for Latin America. We contribute to the literature by providing estimates of key parameters used for power calculations for literacy outcomes such as fluency and reading comprehension, based on data from four studies which conducted cluster-RCTs in six sites in four countries in Latin America: Guatemala, Honduras, Nicaragua, and Peru.

The RCTs used in this study were funded by the United States Agency for International Development (USAID) as part of the Latin America and the Caribbean (LAC) reads initiative (LAC Reads), whose goal was to improve early grade reading in the region. The four studies, described in more detail in [Appendix B](#), tested different promising reading interventions in various contexts. The study populations varied, but all focused on children in early grades (1 through 3), and the duration of the reading interventions, described below, ranged from 12 to 36 months. The outcomes were typically reading comprehension and associated word-level skills such as reading vocabulary, fluency, and other basic literacy skills (such as phonemic awareness and decoding).

This paper is organized as follows. The next section describes the statistical model and parameters that drive precision and sample size requirements. We then discuss the data used for this study and the context in which those data

were collected. The last two sections present results and discuss implications for study design and concluding remarks.

Impact Estimation Model

The RCT design allows a straightforward estimation of average treatment effects. The following model captures the two levels of aggregation for a cluster-randomized trial for an outcome y_{ij} that varies at the student (i) and school (j) levels:

$$y_{ij} = \mu + \lambda' T_j + \alpha' x_{ij} + \theta' X_j + u_j + \varepsilon_{ij} \quad (1)$$

Here T_j is a vector of randomly assigned treatment indicators, which are, by construction, independent of all variables and the error term. Two of the studies had two treatment arms in addition to the omitted group (pure control schools, which received no additional intervention). The others had a simple treatment-control design, so T_j reduces to a scalar. x_{ij} is a vector of student-level covariates such as baseline test scores, age, gender, or household characteristics, and X_j is a vector of school-level covariates such as whether the school has electricity, type of water supply, type of waste disposal, painted walls, separate latrines for boys and girls, and where relevant, school means of the student characteristics such as average test scores. ε_{ij} is the student error term for student i in school j , which is assumed to be normal and independent and identically distributed (iid), with mean zero and variance $\sigma_{W_adj}^2$. u_j is the error term for school j , which is assumed to be normal iid, with mean zero and variance $\sigma_{B_adj}^2$. We have conducted all of our analyses without centering variables at the grand mean or cluster mean, so readers who use these results should consider this when specifying similar models in the design phase.

Variance Parameters

The parameters of interest for this paper are the ICC and the cluster- and individual-level R^2 terms. The ICC, as noted above, is the proportion of the total variance between schools.

$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \quad (2)$$

where σ_B^2 is the between-school variance and σ_W^2 is the within-school variance. When the between- and within-school variances are obtained from a model described in equation (1) without covariates, we refer to the ICC described in equation (2) as unconditional. The ICC could also be calculated with between-school and within-school variances obtained from a model described

in equation (1) with covariates ($\sigma_{B_adj}^2$ and $\sigma_{W_adj}^2$)—in this case we refer to the ICC as conditional.

The standard error of the ICC can be estimated following Kelcey et al. (2016):

$$SE(\widehat{ICC}) = \sqrt{\frac{2(1 - \widehat{ICC})^2(1 + (N - 1)\widehat{ICC})^2}{N(N - 1)J}} \quad (3)$$

where N is the total number of students in the sample and J is the total number of schools.

The R^2 terms at the cluster and individual levels can be calculated as follows:

$$\begin{aligned} R_{cl}^2 &= \frac{\sigma_B^2 - \sigma_{B_adj}^2}{\sigma_B^2} \\ R_{st}^2 &= \frac{\sigma_W^2 - \sigma_{W_adj}^2}{\sigma_W^2} \end{aligned} \quad (4)$$

where R_{cl}^2 is the proportion of variance at the school level explained by the covariates, and R_{st}^2 is the proportion of variance at the student level explained by the covariates. These quantities will vary with the explanatory power of the covariates included in the model as well as the context. Therefore, in this study we present results for different sets of covariates (different models) and for different contexts (targeted regions in four countries).

To assess a study's ability to detect policy-relevant effects, researchers calculate the minimum detectable impact (MDI), which is the smallest true impact for which there is a high probability of detection. The smaller the MDI, the greater the study's ability to detect an impact of policy-relevant magnitude. Both the ICC and R^2 affect the MDI of a cluster-randomized trial. Typically, researchers fix the calculation at 80% power (probability of rejecting the null, given that the null is false) and a two-tailed test with 5% significance level. To facilitate comparison across studies, researchers often report results in terms of effect sizes—impact divided by the standard deviation of the outcome. We can calculate minimum detectable *effect size* (MDES) by dividing the MDI by the standard deviation of the outcome. As Bloom (2005) has shown, the MDES can be written as follows:

$$MDES = M_{J-g-2} \sqrt{\frac{ICC(1 - R_{cl}^2)}{P(1 - P)J} + \frac{(1 - ICC)(1 - R_{st}^2)}{P(1 - P)nJ}} \quad (5)$$

where P is the proportion of groups assigned to the treatment group, J is the number of groups randomly assigned, n is the average number of individuals

in a group, g is the number of group-level covariates included, and M_{J-g-2} is a factor that depends on the statistical power level, the level of confidence of the test, and the degrees of freedom given by $J-g-2$. The other parameters have been defined above.

As is evident from equation (5), for a given experimental design, a higher ICC increases the MDES and therefore makes it harder to detect impacts. A larger R^2 at either level decreases the MDES, and therefore makes it easier to detect impacts. As the ICC gets larger, the second term under the radical becomes less important to the standard error (and the MDES), so the role of cluster size and the explanatory power of individual-level covariates become relatively less important for improving the design. Abstracting away from the role of covariates, the design effect—that is, the penalty associated with the clustering relative to an individual RCT—is proportional to $\sqrt{\text{ICC} + \frac{(1-\text{ICC})}{n}}$, which makes apparent the critical role of the ICC in planning clustered RCT study designs. Despite the central role of the ICC, the stakes for data collection, including potential costs, are high enough that the explanatory power of covariates deserves considerable attention also. While some other research has focused on selecting sample sizes to optimize cost given different costs of data collection at the group or individual level [Konstantopoulos \(2013\)](#), this paper focuses on the tradeoff between statistical power gains and the cost of collecting different types of data, for example, data that can be collected in schools versus data that require household visits.

Models Tested

We used data from cluster-RCTs conducted in Latin America to estimate key parameters for a cluster-level design. We estimated ICCs for cluster-RCTs conducted in six sites in four countries. We also estimated R^2 for different sets of covariates. Calculating R^2 s for different sets of covariates allows us to assess how much the MDES would change if we collected different types of data.

The models simulate various combinations of having or not having a pretest at baseline, a baseline household survey, a school observation, or school administrative data. We take advantage of the fact that the studies collected different combinations of data. For example, two of the studies collected household data because the study focus included community engagement. Other studies did not collect household data but collected school-level achievement for more than one year before the intervention started. These different data sets allow us to assess how much outcome variance would be explained by different combinations of covariates.

We first estimated the model with only the treatment indicator as a covariate. This is what we refer to as the unadjusted model and used to obtain

estimates for the ICC. We then added different sets of student-level variables and/or school-level variables to assess how much variance was explained by each set of covariates, and we obtained the other two key parameters, R_{cl}^2 and R_{st}^2 . Table 1 summarizes the covariates included in each model. For example, models 3 and 5 require a baseline household survey. Models 7, 8, and 9 add school administrative data. In the case of models 8 and 9, we are testing whether such administrative data can substitute for a baseline child assessment, which could save considerable resources if such a strategy yielded adequate statistical precision. We assume that student age and gender are always known. We made this assumption because all the studies discussed collected student age and gender. However, these two variables make a small contribution to explaining the variance, so the results for models that do not include these two covariates would not differ much from the results presented here.

Data

The data used in this study were obtained from studies conducted as part of the LAC Reads initiative. Below we describe the data and sample for each study, but it is worth noting what they had in common:

- They were all clustered-RCTs with schools or communities as the unit of randomization. Outcomes were measured at the student level.
- They focused on early grade reading as the outcome, so the population is students in grades 1 to 3.
- They are school-level interventions. The interventions differed but all aimed to improve reading.
- They were implemented roughly from 2014 to 2017, with intervention durations of 12–36 months.
- The evaluations were funded by USAID, with evaluation design and implementation conducted by Mathematica in partnership with local research and data collection partners.

Interventions and Contexts

Each of the studies examined a different intervention and had slightly different characteristics, as shown in Table 2. The interventions included teacher supports (training, coaching, and mentoring), curriculum materials, after-school (or before-school) enrichment programs, and community-level supports for literacy-promoting activities such as mobile book banks, reading fairs, and classes for parents. The interventions were implemented in diverse contexts, including urban, rural, highlands, and coastal/island communities.

Table 1. Covariates Included in Each Model.

Characteristic	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Treatment indicator(s)	X	X	X	X	X	X	X	X	X
Student pretest		X	X	X	X	X			
Student age, gender			X	X	X	X	X	X	X
Student household characteristics			X		X				
School means of pretest from previous cohort						X	X	X	X
School infrastructure				X	X	X			X

Table 2. Description of the Six Sites in Which Cluster-RCTs Were Conducted.

Characteristic	Amazon Reads		Leer Juntos		PRI	Espacios para Crecer
	Peru site 1	Peru site 2	Guatemala	Peru		
Intervention name	Amazonia Lee		Leer Juntos, Aprender Juntos		Honduras	Nicaragua
Intervention description	Teacher training, coaching, and curriculum materials		Teacher training, coaching, community reading activities		Educación promising reading intervention Student assessment data support	Afterschool enrichment
Geography	Rural and peri-urban, Amazon rain forest region		Rural, mountainous region		Large city, small city, small town, and rural	Urban and rural coastal and island
Grade of cohort at endline	2nd		3rd		3rd	1st to 5th
Language of instruction	Spanish	Spanish	Mixed Spanish-K'iche'	Mixed Spanish-Quechua	Spanish	Mostly Spanish, some Kriol, Miskitu, Ulwa
Number of clusters (schools except for Nicaragua)	200	75	150	147	180	139
Number of students in the data collection sample	8	10	10	10	35	8
Source of data	Campuzano et al. (2018)		Lugo Gil et al. (2019a)	Lugo Gil et al. (2019b)	Liuzzi et al. (2019)	Bagby et al. (2019)

PRI, Promising Reading Intervention; RCT, randomized controlled trial.

They were also linguistically diverse, including predominantly Spanish-speaking communities as well as those where indigenous languages were commonly spoken at home. [Appendix B](#) provides more detailed information on the interventions and contexts.

Outcomes

For three of the four studies, the outcomes were measured using the widely used Early Grade Reading Assessment (EGRA). The EGRA captured reading comprehension, phonemic awareness, familiar word reading, decoding, and oral reading fluency (RTI International, 2016). Reading comprehension was measured as the percentage of questions answered correctly after reading a passage. Decoding accuracy was measured as the number of pseudo-words read correctly in 1 minute. Fluency was measured as the number of real words read correctly in 1 minute.

In this paper, we focus on two outcomes: reading comprehension and reading fluency. For Nicaragua, the EGRA was adapted because of the age range of the children, to include test components from Dynamic Indicators of Reading Success, and to focus on only decoding and reading comprehension. The children in the Nicaragua study were also assessed on oral comprehension (for children who could not decode) and oral reading fluency.

The three studies that used the EGRA administered it at both baseline and endline, although the baseline versions were modified to adapt to children's lower expected level of literacy skills before being exposed to any intervention (including more than a year of school itself in most cases). In the case of the *Leer Juntos, Aprender Juntos* study in Peru and Guatemala, a screener was used to determine which language to use in administering the baseline assessment, since it was translated into Quechua and K'iche', the predominant indigenous languages in those two study sites, respectively.

The Honduras study did not administer the EGRA. Instead, the outcome was a nationally standardized Spanish language reading test developed to align with the National Basic Curriculum. It included an overall reading score and a subscore for reading comprehension. Baseline administrative records were not available at the student level in Honduras, so the study used school-level data. We present the Honduras results separately from the three studies that used the EGRA.

Covariates

Household surveys were administered for three of the study sites, in Guatemala, Peru, and Nicaragua, because those evaluations included at least one main outcome that required household data. Two of the studies, one in Peru and the Honduras study, did not include a household survey, so they are

excluded from some of the R^2 calculations below. The household characteristics in Nicaragua had almost no variation (98% of the households had the same household characteristics); therefore, we also excluded Nicaragua from those comparisons. Some of the variables collected from household surveys were whether the student attended preschool or kindergarten before first grade, the number of people in the household, parents' education, family income, and so on. [Table A3](#) in [Appendix A](#) lists the household variables included in each model.

School infrastructure data were collected from a principal interview, which recorded school characteristics in Honduras. The other studies included school visits with data collected on school infrastructure characteristics such as classroom materials, water supply, availability of working latrines, separate latrines for boys and girls, library, kitchen, outdoor space, computers, printers, music room, health center, and physical hazards at schools. We tried to include similar infrastructure characteristics in all the models, but not all the studies had the same infrastructure data. [Table A2](#) in the [Appendix](#) presents a list of the infrastructure variables included in each model. Other school- or teacher-level data were not collected consistently across the projects; therefore, only infrastructure data were included as covariates. Furthermore, school infrastructure may be correlated with other school resources and/or other factors such as the capacity of the principal or the involvement of the community, which could affect student achievement. Hence, we assume school infrastructure serves as a proxy for school resources and other factors that could affect student achievement.

Results

Below we present results for ICCs and R^2 at the student and school levels for reading comprehension and reading fluency.

ICC Results

The ICC analysis illustrates how homogeneous or heterogeneous the schools were in each study site. ICCs ranged from 0.04 to 0.28, depending on the setting and the outcome. In [Table 3](#), we present the ICC for each setting for the two main outcomes of interest: reading fluency and reading comprehension. We also show the number of schools and students and the grade at which the outcome was measured in each study.

Most of the variance in these outcomes occurs within schools rather than between them. This is good news for planning cluster-randomized trials, suggesting that the number of clusters does not have to be as large as it would be if the ICC were higher. Nevertheless, there is considerable variation. Nicaragua had the lowest ICCs, possibly because the cluster was a community

Table 3. ICC and Standard Error (SE) for Fluency and Reading Comprehension, by Study Site.

	Amazon Reads				Leer Juntos				PRI		Espacios para Crecer	
	Peru site 1		Peru site 2		Guatemala		Peru		Honduras		Nicaragua	
	ICC	SE	ICC	SE	ICC	SE	ICC	SE	ICC	SE	ICC	SE
Fluency	0.15	0.01	0.25	0.03	0.28	0.02	0.18	0.02	n.a.	n.a.	0.06	0.01
Reading comprehension	0.08	0.01	0.17	0.02	0.28	0.02	0.20	0.02	0.21	0.02	0.04	0.00
Number of schools ^a	200		70		150		147		180		139	
Number of students in the analysis sample	1642		740		1338		1022		6375		1106	
Grade	2nd		2nd		3rd		3rd		3rd		1st–5th	

ICC, intraclass correlation; n.a., not available; SE, standard error.

^aIn Nicaragua, clusters are communities not schools.

instead of a school and it included students from several grades. Two regions of Peru, the Leer Juntos site (Andahuaylas) and site 1 of the Peru Amazon Reads study (San Martín), had lower ICCs for reading fluency (0.18 and 0.15, respectively) compared to Guatemala and site 2 of Peru's Amazon Reads study (Department of Ucayali), where we found larger ICCs (0.28 and 0.25, respectively). The two sites with ICCs higher than 0.2 are in rural areas, with schools relatively far apart, and in the Guatemala study site different indigenous communities attend different schools.

In this study, ICCs for reading ranged from 0.04 to 0.28. Kelcey et al. (2016) found ICCs between 0.08 and 0.60 for sub-Saharan Africa using national samples in grade 6. One explanation of why ICCs are lower in this study is that we used local samples instead of national samples and previous studies have found that local ICCs are lower than national ICCs (Hedges & Hedberg, 2013). This study's estimates are similar to findings from Hedges and Hedberg (2013), where ICCs ranged from 0.06 to 0.22 for grade 3 state level samples. Our estimates are also similar to the ranges found in Schochet (2008).

R^2 Results

As discussed above, including covariates in the regression model can reduce the residual variance of the outcome and increase statistical precision. Covariates with stronger linear association with the outcome will be better at increasing statistical precision. Bloom et al. (2007) and Hedges and Hedberg (2007) provided information on the precision gains of including covariates such as pretests or demographic information in the U.S. context, and Stallasch et al. (2021) did so for the German context. In this section, we present the estimates of the amount of variance explained at both the group (school) and individual (student) level for different sets of covariates, according to the models presented above for the Latin American contexts reflected in our studies. Table 4 presents the estimated R^2 at the group level and the individual level, using reading fluency as the outcome for the different models and in each region.¹

Student Pretest Explains More Variance Than Other Data Collected. Student pretests explain 37% of the variance at the group level and 32% of the variance at the individual level for reading fluency in the Peru Leer Juntos site. In contrast, in the Amazon Reads site 1, the pretest explains 76% of the variance at the group level and 39% of the variance at the individual level. The percentage of variance explained in all other regions lies between these two cases. (We excluded Honduras from this discussion because there were no student-level baseline data, no household survey, and the test measured general reading, not fluency; see Table A1 in the Appendix). This result is

Table 4. R^2 at School Level, R^2_{cl} , and Student Level, R^2_{st} , for Reading Fluency, by Study Site.

Model	Student-level covariates			School-level covariates		Amazon Reads				Leer Juntos				Espacios para Crecer (EPC)	
	Pretest	Age and sex	Household characteristics	Previous cohort pretest	Infrastructure	R^2_{cl}	R^2_{st}	Peru site 1	Peru site 2	Guatemala	Peru	R^2_{cl}	R^2_{st}	R^2_{cl}	R^2_{st}
2	X					0.40	0.37	0.76	0.39	0.56	0.29	0.37	0.32	0.64	0.61
3	X	X	X			n.a.	n.a.	n.a.	n.a.	0.59	0.33	0.40	0.35	n.a.	n.a.
4	X	X			X	0.66	0.37	0.93	0.39	0.67	0.32	0.46	0.32	n.a.	n.a.
5	X	X	X		X	n.a.	n.a.	n.a.	n.a.	0.69	0.33	0.50	0.35	n.a.	n.a.
6	X	X		X	X	0.72	0.37	0.95	0.39	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
7	X	X		X		0.51	0.37	0.78	0.39	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
8		X		X		0.36	0.02	0.35	0.02	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
9		X		X	X	0.59	0.02	0.77	0.02	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Note. Individual-level variables are not centered at the group mean level.
EPC, *Espacios para Crecer*, n.a. not available.

consistent with previous research, which found that including a pretest as a covariate can explain two-thirds of between-group variance and up to half of within-group variance (Bloom et al., 2007; Hedges & Hedberg, 2007).

Including Household Characteristics in Regression Models Adds Little to Variance Explained When Pretest Is Included. Model 3 includes household information in addition to the pretest. In the Peru Leer Juntos site, with student pretest, age, gender, and household variables we can explain 40% of the variance at the school level and 35 at the student level, which is similar to what can be explained with only a pretest (37% and 32%). Similarly, in Guatemala, adding household characteristics to the regression that already has pretest data did not increase the R^2 by much; the school-level R^2 changed from 56 to 59% and the student-level R^2 changed from 29 to 33%. This result is similar to previous research, which found that demographic information does little to increase power once a pretest is accounted for (Hedges & Hedberg, 2007). Collecting household data is not common in U.S. education studies, but some studies in international contexts collect household data so it is important to assess how much power will be gained by collecting these data. Given the high cost of collecting household data, doing so to provide contextual information may not be worthwhile, and studies should prioritize collecting pretest information.

School Infrastructure Information Explained a Substantial Portion of School-Level Variance. Model 4 shows that, if we account for infrastructure information in addition to the pretest, we can explain around 10 percentage points more in Guatemala (school level R^2 goes from 0.56 to 0.67) and in Peru (from 0.37 to 0.46) and even more in the other regions in Peru (school level R^2 goes from 0.40 to 0.72 in Amazon Reads site 1 and from 0.76 to 0.93 in site 2). Given the low cost of collecting these additional data when visiting schools to collect student pretest data, it is worth collecting infrastructure information and including it in our regressions as covariates. In the U.S. context, school infrastructure may not have the same explanatory power as in the international context, where there may be more variation among schools.

School-Level Pretest Data Explain Less Variance Than Student Pretests. Often a student-level pretest is not feasible to collect. One possible solution is to use school-level average pretest scores for a previous cohort at the same school. For example, an intervention may target first-grade students, but randomization occurs before students actually enroll, so there is no time to test them before they learn of their treatment status. In these cases, researchers may wish to use the school average scores of first graders from the end of the school year before the experiment began. These average scores would absorb some of the chance differences in school-level baseline skills, but would not be measured for the actual students in the

study. Therefore, at best, they would capture school factors or factors associated with attending the school, such as motivation and support of the parents. This is appropriate to the extent that these factors are stable over time, or the students in the older cohort are siblings of the students in the study. In the U.S. context, administrative data are widely available so using school-level data may be more feasible than collecting pretest data for the study. But in international settings, administrative data may not be available so the study may need to collect its own data.

Using pretest information at the school level reduced outcome variance but not as much as using student-level pretest data. Model 7 presents the case where school average scores for a cohort of first-grade students who attended the schools one year before the intervention started are used as covariates in addition to a student-level pretest. In Amazon Reads site 1, the school mean pretests on the previous cohort added very little to what the student pretest already explained (the school-level R^2 went from 0.76 to 0.78). In contrast, in Amazon Reads site 2 the school-level R^2 went from 0.40 to 0.51. If we had collected only pretest data from the previous cohort (model 8), the percentage of variance we could have explained at the school level would be 0.36 in Amazon Reads site 1 and 0.35 in Amazon Reads site 2. This is less than half of what we can explain with a student pretest (model 2), so the precision gains were larger with a student pretest than with a school pretest on the previous cohort.

These results differ from Bloom et al. (2007). In their samples, school-level pretests are as effective as student-level pretests. However, in the Peru study, the student-level pretest corresponds to scores of the students attending the prior year (first grade) and the school-level pretest corresponds to the scores a previous cohort attained the prior year (first grade). In contrast, in Bloom et al. (2007), the school-level pretest corresponds to the scores a previous cohort attained in the same grade as when the post-test was measured (second grade). Hence, it is not surprising that our school-level data (previous grade rather than outcome from a previous cohort) is not as effective in reducing the outcome variance as the Bloom et al. (2007) data (same grade as outcome from a previous cohort).

Furthermore, as we can see in model 8, age and gender explain only 2% of the variance at the student level. This is similar to findings from previous research (Bloom et al., 2007; Hedges & Hedberg, 2007).

School-Level Pretest Data on Previous Cohort in Same Grade Reduced Variance Slightly More Than School-Level Pretest Data on Same Cohort in Previous Grade. In Honduras, we collected administrative data at the school level, which allowed us to compare models where only school-level data were available. Table A1 in the Appendix presents the R^2 s and MDES for these cases. Test scores for third graders at the school level were used as the outcome in all

models. In model 2, school average first-grade test scores for the same cohort were used as baseline data. In model 7, school average third-grade scores for a previous cohort that was not exposed to the intervention were used as baseline information. The R^2 is 0.17 for the model with school average first-grade pretest and 0.23 for the model with school average third-grade pretest. Hence, it seems that a school-level pretest on the same cohort in previous grades increases power slightly less than a pretest on a previous cohort but in the same grade as the outcome. We might have expected more precision gains using the pretest for the same cohort, but perhaps because there is large variation on reading in first grade this pretest was not more useful than the third-grade pretest on a different cohort. Bloom et al. (2007) also found that precision gains reduced as the time lag between the pretest and the post-test increases.

Translating Results Into MDES

One way to understand the role of the parameters presented above for ICC and R^2 for the school and student levels is to calculate the MDES that would result from each set of variance parameters. We used a sample size of 45 schools in the treatment group and 45 schools in the control group, and a cluster size of 10 students per school. We chose a sample size (90 schools) that was within the range of the studies we used for this exercise (70–200). The actual sample size is illustrative to show the relationships among the MDES and the sets of covariates used in each specification. Those relationships would be similar for different numbers of clusters used in the design. Similarly, the ICC would affect these exact numbers, but for realism we used the realized ICCs from each of the RCTs conducted. The MDES are shown in Table 5.

Without including any covariates (model 1), assuming 90 schools and 900 students our studies can detect effect sizes of around 0.30. The highest MDES is 0.35 for Guatemala and the lowest is 0.29 for Peru, Amazon Reads site 2.

Including pretests in our regression models (model 2), allows us to detect smaller effect sizes. For example, we can detect effect sizes of 0.25 in Guatemala and 0.22 in Peru, Amazon Reads site 2. In Peru, Amazon Reads site 1, adding the pretest allows us to detect an effect size of 0.19; this is a reduction of 44%, which is considerable given that without covariates we could detect only a 0.34 effect size. These reductions are consistent with the finding in Bloom et al. (2007) and slightly larger in some cases.

Table 5. MDES for Reading Fluency, by Site and Model.

Student-level covariates					MDES				
					School-level covariates		Amazon Reads		EPC
Pre-test	Age and sex	Household characteristics		Previous cohort pretest	Infrastructure characteristics	Peru site 1	Peru site 2	Guatemala	Peru
1						0.29	0.34	0.35	0.30
2	X					0.22	0.19	0.25	0.24
3	X	X				n.a.	n.a.	0.24	n.a.
4	X	X			X	0.19	0.15	0.22	0.23
5	X	X	X		X	n.a.	n.a.	0.22	n.a.
6	X	X		X	X	0.18	0.14	n.a.	n.a.
7	X	X		X		0.21	0.19	n.a.	n.a.
8	X	X		X		0.25	0.29	n.a.	n.a.
9	X	X		X	X	0.22	0.21	n.a.	n.a.
ICC						0.15	0.25	0.28	0.18

Note. We assumed the same sample size for all the sites and models. MDES were calculated for 90 schools (45 in treatment group and 45 in control group) and 900 students (10 students per school).
EPC, *Espacios para Crecer*; ICC, intraclass correlation; MDES, minimum detectable effect sizes; n.a, not available.

Including household characteristics in the regression models in addition to the pretest (comparing model 3 to model 2) does not change the MDES in the Leer Juntos Peru site and it decreases it by only 0.01 in Guatemala. Hence, collecting household variables may not be a cost-efficient way of increasing statistical power if a pretest has already been collected.

Including school infrastructure characteristics in the regression models (model 4 compared to model 2) reduced the MDES. We found the lowest reduction in the Leer Juntos Peru site, where the MDES went from 0.24 to 0.23, and the largest reduction in Peru Amazon Reads site 2 where the MDES decreased from 0.19 to 0.15. Given the low cost of collecting school infrastructure characteristics once the schools are visited for pretest, this may be a cost-efficient way of increasing statistical power.

Including a student pretest in the regression models attains higher statistical power than including a school average pretest on a previous cohort (model 2 compared to model 8). In Peru Amazon Reads site 2, including a school mean pretest, age, and gender (model 8) allowed us to detect an effect size of 0.25, and including a pretest for the students allowed us to detect an MDES of 0.22. The difference in site 1 was larger, with a decrease from 0.29 to 0.19, which is considerable. Given that administrative data are not widely available for many low- and middle-income countries, collecting pretest data in the sample of students who will be the target of the study and conducting longitudinal analysis may be a better alternative than collecting data on a previous cohort and conducting cross-sectional analysis.

The models that include all student- and school-level covariates, (models 5 and 6) attain the lowest MDES. For example, in the case of Peru site 2, we can detect an effect size of 0.14 if we include all covariates, compared to 0.34 without including covariates, which is a 57% reduction.

Conclusions and Implications for Experimental Design

This paper documents the range over which key parameters for designing education cluster-RCTs vary for several diverse study sites in Latin America and compares them to previous findings from the United States, in particular to findings from [Bloom et al. \(2007\)](#) and [Hedges and Hedberg \(2007\)](#). We found that ICCs for reading comprehension and fluency were as low as 0.04 for communities that included students of all elementary grades, and as high as 0.28 for schools in rural areas and those in which different indigenous communities attend different schools. These ranges are similar to what

researchers have found in studies in the United States. Our findings can be used to plan future studies by comparing the context of planned study sites to the contexts included in this paper to find the best match. The samples in our studies were in regions smaller than a state. Our data included only public schools and schools located in regions that were performing below the national averages. Future studies to be conducted in rural, low-income regions may find the results for Guatemala Leer Juntos and Peru Amazon Reads site 2 more relevant. Studies in urban and peri-urban areas of low-income regions may find the results from Peru Leer Juntos and Peru Amazon Reads site 1 more relevant. Rural areas seem to have higher ICCs than urban and peri-urban areas in our studies because there is more variation between schools in rural areas.

The findings in this paper also inform which type of baseline data provides the greatest gains in terms of statistical power. For each type of covariate, we recommend repeating the power calculation using tools such as Optimal Design (Raudenbush et al., 2011). Consistent with previous research, student-level pretests increase the statistical power of the study considerably, so it may be worth investing in collecting baseline tests for the study sample. Our findings will allow researchers to use more informed data collection strategies for early grade reading studies, in particular, if the widely adopted early grade reading assessment (EGRA) is used. Once pretest data have been included as covariates, adding household data did little to improve statistical power. Researchers need to consider whether collecting household data is worthwhile, considering the cost of data collection. If an outcome such as time spent reading at home is key for the evaluation, household data collection is necessary, but it should not be prioritized if the only reason to collect it is to improve precision of the impact estimates. Specifically, a researcher planning a prospective study can enter different values for R^2 for the different types of data collection planned, using the values in Table 4 as a guide.

Finally, school infrastructure data may be a cost-efficient way of increasing statistical power given that the cost of collecting it may be marginal in cases where researchers plan to collect student test scores. Researchers can perform the same exercise of changing the percentage of variance explained in their power calculations to see if the lower MDES from including school-level covariates is worth the added cost. A potential explanation of why school infrastructure data may increase precision is that, in low- and middle-income countries, school infrastructure may be a proxy for how involved the community is in the school, or the capacity of the principal, both of which may also affect achievement.

The findings we shared about ICCs and the contribution of different types of covariates to the explanatory power of the model can be used to plan efficient experiments. In particular, there is a literature on optimizing sample allocation between levels of a cluster-randomized trial (Bloom, 2005; Hedges & Borenstein, 2014; Liu, 2003; Raudenbush, 1997), where the ICC and the R^2

at each level (student and school) are key inputs. As the ratio of data collection costs increases at a given level (e.g., costs of schools relative to students), the optimal sample allocation shifts more toward the other level (in this case towards fewer schools and more students per school, all things equal). The literature on such optimal design frameworks and an expansion of these frameworks to incorporate more perspectives is covered elsewhere (Shen & Kelcey, 2020).

It remains for future research to identify the factors that determine why outcomes vary more specifically between schools than within schools and why commonly collected data from child assessments, household surveys, school and classroom observations, and administrative data would do a better or worse job of reducing residual variance. This may require a larger sample of studies. In addition, it is unclear whether these results will hold in high-income countries; for example, school infrastructure may have less explanatory power in those contexts and household data may be more informative. As more researchers publish information on the variance components (specifically ICCs and R^2) identified in ongoing and completed RCTs, the field will benefit from better-informed study designs and the pace of scientific knowledge will accelerate.

Appendix

Appendix A. Auxiliary Tables

Table A1. R^2 at School Level, R^2_{cl} , and Student Level, R^2_{st} , and MDDES for Overall Reading, Honduras.

Model	Student covariates	School covariates			Honduras findings		
	Age and gender	School average of 1st-grade pretest in baseline year	School average of 3rd-grade reading test in baseline year	School infrastructure	R^2_{cl}	R^2_{st}	MDDES
2		X			0.17	0.01	0.19
3	X	X			0.21	0.02	0.18
4	X	X		X	0.39	0.02	0.17
5	X	X	X	X	0.41	0.02	0.16
6	X	X	X		0.25	0.02	0.18
7			X		0.20	0.00	0.19
8	X		X		0.23	0.01	0.18
9	X		X	X	0.39	0.01	0.16

MDDES, minimum detectable effect size.

Table A2. School Infrastructure Variables Included, by Study.

School infrastructure variable	Amazon Reads Peru site 1	Amazon Reads Peru site 2	Leer Juntos Guatemala	Leer Juntos Peru	Nicaragua	Honduras
Water supply not piped	X	X	X	X		X
Potable drinking water supply	X	X	X	X		X
Permanent building	X	X	X	X		
Painted exterior walls	X	X	X	X		
Working restroom facilities for children	X	X	X	X	X	X
Functional toilets	X	X	X	X		
Separate facilities for boys and girls	X	X	X	X	X	
Working waste disposal	X	X	X	X		X
Septic (or similar) disposal	X	X	X	X		X
Plumbing waste disposal	X	X	X	X		X
Other waste disposal	X	X	X	X		
Kitchen	X	X	X	X		X
Outdoors recreational space	X	X	X	X		X
Gymnasium or sport facilities	X	X	X	X		
Computers for students	X	X	X	X		X
Internet connectivity	X	X	X	X		X
Library or resource room	X	X	X	X		X

(continued)

Table A2. (continued)

School infrastructure variable	Amazon Reads Peru site 1	Amazon Reads Peru site 2	Leer Juntos Guatemala	Leer Juntos Peru	Nicaragua	Honduras
Number of physical hazards observed in school	X	X	X	X		X
Number of health hazards observed in school	X	X	X	X		X
Classrooms with finished materials (brick)					X	

Table A3. School Household Variables Included, by Study.

School infrastructure variable	Leer Juntos Guatemala	Leer Juntos Peru
Child attended preschool or kindergarten before starting first grade	X	X
Average number of people living in the household	X	X
Number of rooms in the house	X	X
Highest grade attained by mother	X	X
Parent reading skills (knows how to read)	X	X
Mother language K'iche' only	X	X
Family monthly income	X	X
Land owned by a household member	X	X
House with finished floor	X	X
House with finished roof	X	X
House with electricity	X	X
House with phone	X	X
House with radio	X	X
House with television	X	X
House with refrigerator	X	X
House with computer	X	X
House with bicycle	X	X
House with motorcycle	X	X
House with car	X	X

Appendix B. LAC Reads Study Descriptions

Amazon Reads

Context. In Peru, Amazon Reads (*Amazonía Lee* in Spanish) was implemented in two sites by the Department of Regional Education Offices (DREs) of Ucayali and San Martín with funding from USAID starting in 2015. Site 1, San Martín lies in the northern part of the Peruvian Amazon forest. Site 2, Ucayali is located south of San Martín, bordering Brazil. These departments are primarily Spanish speaking, with a small percentage of the population speaking other indigenous languages. San Martín is more densely populated than Ucayali. San Martín has approximately 1302 active primary public education institutions, whereas Ucayali has approximately 796. Ucayali and, to a lesser extent, San Martín lag behind the nation on student reading achievement indicators.

Sample. The DRE in San Martín prepared a list of schools eligible to participate in the intervention. Eligible schools were public schools that had no other support programs and a minimum of 16 second graders in 2012. In San Martín, the final study sample included 200 eligible schools. In Ucayali, the DRE identified schools to participate in the study. The final study sample included 70 schools. In each study school, one first-grade classroom was randomly selected for data collection. Based on the classroom roster for the selected classroom, on average 8 first graders in San Martín and 10 in Ucayali were randomly selected (evenly divided by gender when possible). Students who were identified as having a physical or severe cognitive disability or as no longer attending the school were excluded. At endline, we tracked the study students from the 2015 first-grade cohort to evaluate their reading skills at the end of the second year of the intervention, when they were expected to be in second grade.

Data. The main data sources for the study are student reading assessments and a school infrastructure survey. The student assessments were collected at baseline (when students were attending first grade) and endline (when students were attending second grade). The endline assessment was a version of the Early Grade Reading Assessment (EGRA) that tested oral comprehension, familiar word reading, decoding, reading fluency, and reading comprehension. We designed and piloted an endline EGRA, in collaboration with a local data collection partner following the guidelines of the Early Grade Reading Assessment (EGRA) Toolkit ([RTI International, 2016](#)). The reading assessments were administered individually by a trained enumerator. Most students were assessed at school during regular school hours, either in their baseline school or in another school if they had transferred. The final student reading outcomes were as

follows: decoding accuracy (correct pseudo-words per minute), reading fluency (correct words per minute), familiar word reading (number of familiar words read correctly), and reading comprehension (number of correct answers).

Leer Juntos, Aprender Juntos

Context. *Leer Juntos, Aprender Juntos* was implemented in two sites, Peru and Guatemala. In Peru and Guatemala, the official language is Spanish, but some children speak other languages and have no knowledge of Spanish. *Leer Juntos, Aprender Juntos* is a program funded by USAID that aims to improve early grade reading instruction in communities with linguistically diverse populations. It was developed by Save the Children based on its Literacy Boost model, which includes teacher training and community involvement, and it was implemented in the K'iche'-speaking region of Guatemala and the Quechua-speaking region of Apurímac in Peru. In Guatemala, the study schools were located in five municipalities in the Department of El Quiché. The Department of El Quiché was selected for the study because its population is ethnically and linguistically diverse. In Peru, the study schools were located in the Apurímac region, in the Andes of southern-central Peru. Apurímac was selected because its population is linguistically diverse. The schools were selected from two provinces, Andahuaylas and Chincheros. The implementation of *Leer Juntos, Aprender Juntos* began in May 2013 in Peru and Guatemala, and continued through December 2015 in Peru and March 2016 in Guatemala.

Sample. The study recruited 147 schools in Peru and 150 schools in Guatemala. In each country, two-thirds of the schools were randomly assigned to an intervention group and one-third was assigned to a control group. Within each school, a group of children were followed from first grade through the end of third grade. The evaluation team randomly selected one first-grade classroom, one teacher, and 10 first-grade students from each school to serve as the analysis sample for the evaluation. The endline follow-up data collection took place three years after the baseline, when most of the children in the study should have progressed to third grade and have had two or more years of potential exposure to the intervention.

Data. During the baseline data collection, when children were in the first grade, the study assessed children's oral language proficiency and emergent literacy skills (letter identification, emergent writing, emergent reading, phonemic awareness, pseudo-word decoding, and passage comprehension skills) in Spanish and Quechua in Peru and Spanish and K'iche' in Guatemala. At endline, the individual assessment was administered when most of the children in the evaluation attended the third

grade and it focused on reading skills in Spanish. The endline assessments were conducted in Spanish because Spanish is the language of instruction in the third-grade classrooms. Children's literacy skills were assessed in three tasks: (1) pseudo-word reading (decoding), (2) reading fluency, and (3) reading comprehension. These three tasks were adapted for Peru and Guatemala following the guidelines of the Early Grade Reading Assessment (EGRA) Toolkit (RTI International, 2016). A household survey was conducted at midline when most students were in second grade. This survey was administered in person to the main caregivers (usually the mother or father) of the children in the evaluation at their homes. The purpose of this survey was to learn about the household composition, family socio-economic status, household assets, children's schooling background and routines at home, and children's and families' participation in reading activities offered in their communities.

Nicaragua

Context. USAID Nicaragua funded the Community Action for Reading and Security (CARS) program in five municipalities in the Southern Caribbean Atlantic Autonomous Region (Región Autónoma de la Costa Caribe Sur, RACCS) from 2011 through 2017. The RACCS is rural, hard to access, and half the population lives in extreme poverty. The intervention, Spaces to Learn, known by its Spanish name Espacios para Crecer (EpC), is an afterschool program implemented as part of CARS. Eligible communities had to be accessible, have schools that offered first to third grades, and have at least 20 children who were eligible to participate in the intervention. To be eligible to participate in the afterschool program, children had to be 6–16 years old; and they had to be out of school with the equivalent of a third-grade education or less (or enrolled in first, second, or third grade), or enrolled in school and have one risk factor such as having failed a grade, high absenteeism, or a mother tongue different from the language of instruction. Spanish was the primary language of instruction in most communities.

Sample. The communities included in the study were small, with at least 20 and fewer than 50 eligible children. The evaluation followed two cohorts of students for approximately one-and-a-half years of exposure to the afterschool program. We collected base-year data to measure children's literacy skills but could do so in only one of the cohorts. We collected follow-up data for each cohort (in 2016 for Cohort 1 and in 2017 for Cohort 2). The sample of students was representative of students eligible for the afterschool program: it included children in school with at-risk factors, as well as out-of-school children, within the ages of 6 to 13.

Data. Literacy skills were measured in the language of instruction in schools in the community. To measure children's literacy skills, we created an assessment based on the EGRA and the Dynamic Indicators of Reading Success (IDEL in Spanish). The four skills measured were (1) decoding pseudo-words or unfamiliar words, (2) oral comprehension (administered only to children unable to decode pseudo-words), (3) oral reading fluency, and (4) reading comprehension. Local stakeholders reviewed all assessment materials to ensure that the questions were appropriate for the cultural and linguistic context of the RACCS.

Honduras

Context. EducAcción was a USAID/Honduras-funded project carried out by AIR, who worked with municipalities, districts, and schools in Honduras since 2011 to promote improved school management techniques, community involvement in schools, and teacher training on Spanish and math instruction. The EducAcción project's school supports included assistance in using formative assessments and end-of-grade (EOG) results to improve instruction. The project was implemented in the two predominantly rural departments of Lempira and Santa Barbara and the two urban areas of Tegucigalpa and La Ceiba.

Sample. We selected an initial sample of 240 schools that were low-performing, had at least 10 first-grade students in 2013, and EOG test scores in the bottom three quintiles of performance. Schools that were infeasible to visit because they were in areas that were highly prone to violence were excluded. The study followed the cohort of students who were enrolled in second grade as of the last day before the intervention began—May 31, 2015. We measured learning after half a year of support in second grade (2015) and a full year of support in third grade (2016).

Data. In most years since 2007, the Ministry of Education, with international donor funding from several sources, has supported administration of the EOG tests. Mejorando el Impacto al Desempeño Estudiantil de Honduras (MIDEH) developed and has administered the tests. For this evaluation, we used EOG test data from four academic years: 2013 through 2016. The items in the tests varied from year to year; however, all items were drawn from the same item bank that MIDEH developed and considered to be equally difficult. The study's main outcome measures are third grade EOG test scores in reading for the study cohort in 2016.

Acknowledgments

The authors acknowledge substantive input from colleagues Camila Fernandez, Sarah Liuzzi, Julieta Lugo-Gil, Emilie Bagby, and Virginia Poggio. Ivonne Padilla, Irina Cheban, and Galina Lapadatova provided expert programming assistance. Errors are sole responsibility of the authors.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partly supported by the United States Agency for International Development under Task Order AID-OAA-M12-00020.

Data Availability Statement

The data that support the findings of this study are available from United States Agency for International Development (USAID) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

ORCID iD

Larissa Campuzano  <https://orcid.org/0009-0003-0386-9959>

Note

1. The results for Honduras presented in the appendix are based on a general reading test that includes comprehension, but does not explicitly report results by literacy skill, such as reading fluency.

References

- Bagby, E., Torrente, C., Glazerman, S., Murray, N., & Padilla, I. (2019). Evaluation of Espacios para Crecer (EpC), an afterschool program. In *Nicaragua: Final report. Mathematica*. Report prepared for USAID. https://pdf.usaid.gov/pdf_docs/PA00XDR4.pdf
- Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). Russell Sage Foundation.
- Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions.

- Educational Evaluation and Policy Analysis*, 29(1), 30–59. <https://doi.org/10.3102/0162373707299550>
- Brunner, M., Keller, U., Wenger, M., Fischbach, A., & Lüdtke, O. (2018). Between-school variation in students' achievement, motivation, affect, and learning strategies: Results from 81 countries for planning group-randomized trials in education. *Journal of Research on Educational Effectiveness*, 11(3), 452–478. <https://doi.org/10.1080/19345747.2017.1375584>
- Campuzano, L., Fernández, C., Lugo-Gil, J., Glazerman, S., & Murray, N. (2018). *Evaluation of Amazonia Lee reading intervention in Peru: Final report*. Mathematics. Report prepared for USAID. https://pdf.usaid.gov/pdf_docs/PA00TCQ1.pdf
- Duflo, E., Glennerster, R., & Kremer, M. (2008). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, 4(5), 3895–3962. <https://econpapers.repec.org/bookchap/eedevchp/5-61.htm>
- Hedberg, E., & Hedges, L. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics: Results from a meta-analysis of district-specific values. *Evaluation Review*, 38(6), 546–582. <https://doi.org/10.1177/0193841X14554212>
- Hedges, L., & Hedberg, E. (2013). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review*, 37(6), 445–489. <https://doi.org/10.1177/0193841X14529126>
- Hedges, L. V., & Borenstein, M. (2014). Conditional optimal design in three- and four-level experiments. *Journal of Educational and Behavioral Statistics*, 39(4), 257–281. <https://doi.org/10.3102/1076998614534897>
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60–87. <https://doi.org/10.3102/0162373707299706>
- Jacob, R., Zhu, P., & Bloom, H. (2010). New empirical evidence for the design of group randomized trials in education. *Journal of Research on Educational Effectiveness*, 3(2), 157–198. <https://doi.org/10.1080/19345741003592428>
- Kelcey, B., & Phelps, G. (2013). Strategies for improving power in school-randomized studies of professional development. *Evaluation Review*, 37(6), 520–554. <https://doi.org/10.1177/0193841X14528906>
- Kelcey, B., Shen, Z., & Spybrook, J. (2016). Intraclass correlation coefficients for designing cluster-randomized trials in sub-Saharan Africa education. *Evaluation Review*, 40(6), 500–525. <https://doi.org/10.1177/0193841x16660246>
- Konstantopoulos, S. (2012). The impact of covariates on statistical power in cluster randomized designs: Which level matters more? *Multivariate Behavioral Research*, 47(3), 392–420. <https://doi.org/10.1080/00273171.2012.673898>
- Konstantopoulos, S. (2013). Optimal design in three-level block randomized designs with two levels of nesting: An ANOVA framework with random effects.

- Educational and Psychological Measurement*, 73(5), 784–802. <https://doi.org/10.1177/0013164413485752>
- Liu, X. (2003). Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. *Journal of Educational and Behavioral Statistics*, 28(3), 231–248. <https://doi.org/10.3102/10769986028003231>
- Liuzzi, S., Murray, N., Glazerman, S., & Cheban, I. (2019). *Data-driven instruction in Honduras: An impact evaluation of the EducAcción-PRI promising reading intervention*. Mathematica. Final report. Report prepared for USAID. https://pdf.usaid.gov/pdf_docs/PA00WDW1.pdf
- Lugo-Gil, J., Murray, N., Glazerman, S., Fernández, C., & Campuzano, L. (2019a). *Evaluation of Leer Juntos, Aprender Juntos early grade intervention in Guatemala: Final report*. Mathematica. Report prepared for USAID. <https://www.edu-links.org/sites/default/files/media/file/Evaluation/of/Leer/Juntos%2C/Aprender/Juntos/Early/Grade/Reading/Intervention/in/Guatemala/Final/Report.pdf>
- Lugo-Gil, J., Murray, N., Glazerman, S., Fernández, C., Campuzano, L., & Padilla, I. (2019b). *Evaluation of Leer Juntos, Aprender Juntos early grade intervention in Peru: Final report*. Mathematica. Report prepared for USAID. https://pdf.usaid.gov/pdf_docs/PA00XJKM.pdf
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2), 173–185. <https://doi.org/10.1037/1082-989X.2.2.173>
- Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, 29(1), 5–29. <https://doi.org/10.3102/0162373707299460>
- Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., & Hill, C. (2011). *Optimal design plus empirical evidence (Version 3.0)*. <https://www.wtgrantfoundation.org/resources/optimal-design>
- RTI International (2016). *Early grade reading assessment (EGRA) toolkit* (2nd ed.). United States Agency for International Development.
- Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, 33(1), 62–87. <https://doi.org/10.3102/1076998607302714>
- Seidenfeld, D., Handa, S., de Hoop, T., & Morey, M. (2023). Intraclass correlations values in international development: Evidence across commonly studied domains in sub-Saharan Africa. *Evaluation Review*, 47(5), 786–819. <https://doi.org/10.1177/0193841X231154714>
- Shen, Z., Curran, F. C., You, Y., Splett, J. W., & Zhang, H. (2023). Intraclass correlations for evaluating the effects of teacher empowerment programs on student educational outcomes. *Educational Evaluation and Policy Analysis*, 45(1), 134–156. <https://doi.org/10.3102/01623737221111400>

- Shen, Z., & Kelcey, B. (2020). Optimal sample allocation under unequal costs in cluster-randomized trials. *Journal of Educational and Behavioral Statistics*, 45(4), 446–474. <https://doi.org/10.3102/1076998620912418>
- Spybrook, J., & Kelcey, B. (2016). Introduction to three special issues on design parameter values for planning cluster randomized trials in the social sciences. *Evaluation Review*, 40(6), 491–499. <https://doi.org/10.1177/0193841X16685646>
- Stallach, S. E., Lüdtke, O., Artelt, C., & Brunner, M. (2021). Multilevel design parameters to plan cluster-randomized intervention studies on student achievement in elementary and secondary school. *Journal of Research on Educational Effectiveness*, 14(1), 172–206. <https://doi.org/10.1080/19345747.2020.1823539>
- Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, 37(6), 490–519. <https://doi.org/10.1177/0193841X14531584>