USAID
FROM THE AMERICAN PEOPLE

Methodological Insights for Impact Evaluations

This brief is part of a series uncovering lessons learned from four impact evaluations of promising reading interventions funded by USAID as part of the Latin America and the Caribbean Reads (LAC Reads) project. The evaluations were conducted by Mathematica.

Background

Randomized controlled trials of education interventions present evaluators with many choices in the form of both challenges and opportunities. Impact evaluations of programs designed to improve early grade reading in Guatemala, Honduras, Nicaragua, and Peru generated methodological insights that could help researchers, implementers, and donors make decisions that improve evidence. These include decisions about baseline and interim data collection, the unit of randomization, and the contrasts that provide meaningful learning.

Baseline data are important, but not all baseline data are equally important

It is well known that baseline data can be used to adjust impact estimates to increase a study's statistical power. To assess the statistical power, researchers calculate the minimum detectable effect (MDE), which is the smallest true effect that the study would be able to detect. The smaller the MDE, the higher the statistical power. A study's sample size, such as the number of schools or students, is the main driver of statistical power. But other factors can also help researchers increase the statistical power of a study. Researchers often collect baseline data and use them in statistical models to increase the statistical power of a study. Here we discuss strategies for increasing statistical power assuming a given sample size.

The types of baseline data and the costs of obtaining them should be weighed against any resulting gains in precision. The LAC Reads evaluations, which measured programs' impacts on student reading skills, used student-level pre-intervention data on key reading skills to adjust impact estimates and increase precision. The studies gained additional precision by adjusting for (1) school infrastructure characteristics such as access to potable water, working restroom facilities, library, and internet connectivity; and (2) household-level data such as number of rooms in the house, highest grade attained by mother, access to key services, and assets. Adjusting for baseline student reading test scores provided the largest gains in statistical power. We calculated MDEs with and without baseline test scores in the studies' regression models, assessing how much the MDEs decreased when we included pre-test scores in the models. This revealed gains in statistical power as large as 44 percent and as small as 20 percent. The evidence for other early reading interventions shows effect sizes between 0.20 and 0.30 standard deviations for early reading outcomes (Young-Suk, Kim; Hansol Lee; Stephanie Zuilkowski, 2020). For the LAC Reads evaluations, using baseline data allowed the evaluation team to detect impacts close to the lowest effect size in that range without expanding the sample size. Detecting impacts within comparable effect sizes allows for comparison with the impacts found in other reading interventions.

Baseline test example

For example, in one of the *Amazonía Lee* sites, the MDE without baseline test information was 9 correct words read per minute (0.34 standard deviations). When we included baseline test scores, the MDE decreased to 5 words per minute (0.19 standard deviations), a reduction of 44 percent. In Guatemala, without baseline test scores the MDE was 10 correct words per minute and decreased to 7 words per minute when baseline tests were included (30 percent reduction). In other sites, the MDE decreased by 1 word per minute (20 percent) in the Andean region of Peru and by 3 words per minute (30 percent) in Nicaragua.

May 2021

Adjusting impacts for school-level characteristics proved to be a low-cost approach to increase statistical power for several of the LAC Reads evaluations.

For the impact evaluations of interventions targeting learning outcomes in LAC, including information on school infrastructure characteristics increased statistical power substantially with only a marginal increase in overall data collection costs. For example, in Guatemala, the MDE for a model that adjusted only for student pre-test scores was 7 correct words read per minute (0.25 standard deviations). When school-level infrastructure characteristics were included in the regression, the MDE decreased by 12 percent to 6 words per minute (0.23 standard deviations). In other sites, including school characteristics in the estimation similarly decreased the MDEs. We hypothesize that these school-level variables may actually reflect differences in principal effectiveness related to school maintenance and resources, as well as perhaps community engagement. Collecting these data added little cost, since each LAC Reads evaluation included visits to schools to obtain student test data or interview teachers. This is often the case with education evaluations. Education researchers should consider collecting additional information about school characteristics during school visits, as such data collection typically represents only a small cost increase.

Adding household data provided minimal gains in power after including baseline test scores and did not provide sufficient power gains to be cost-effective. In the case of the *Leer Juntos, Aprender Juntos* evaluation, household data provided minimal gains in statistical power after baseline test scores were included. For example, given the sample size for Guatemala, the MDE was 7 words per minute after adjusting for student pre-tests. When household characteristics were also included in the regression, the MDE reduction was less than 1 word. In the Andean region of Peru, the MDE reduction was also less than 1 word when we included household characteristics. In the evaluation of *Espacios para Crecer* in Nicaragua, the other LAC Reads evaluation with household data collection, the data also did not increase statistical power. Collecting data on family and household characteristics is often costly. It typically requires resource-intensive visits to each student's household, which involves locating, scheduling, and transportation costs. Unless the visits are otherwise required (such as to test out-of-school children, or to assess other key outcomes, such as time spent reading at home), collecting data on household characteristics may not be worth the cost, particularly if there is not much variation in the socio-demographic characteristics of the households in the evaluation sample.

Individual-level random assignment increases statistical power, but group randomization can reduce contamination and facilitate program participation

Individual-level randomization improves statistical power. It may appeal to program implementers to use individual-level (or within-school) assignment because it enables them to serve as many children as possible while also generating information about impact. This is a good option in places where demand is expected to exceed supply. For example, in Nicaragua, we were able to conduct individual random assignment in large communities because the demand for services was expected to exceed the number of slots available for after-school services, and the intervention was limited to certain classrooms with trained facilitators.

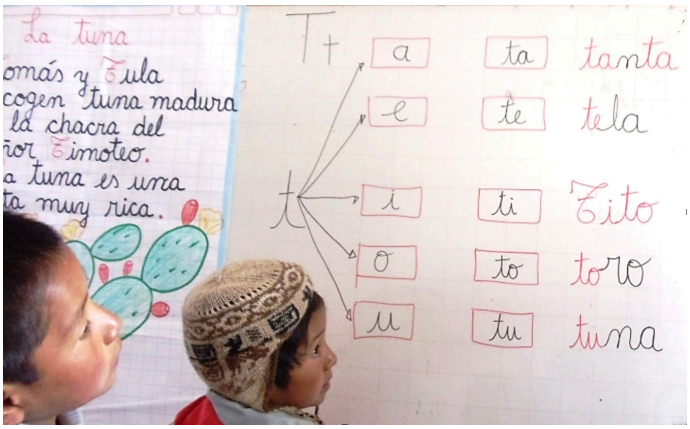
However, individual-level random assignment can be challenging for interventions rolled out at the school level. Also, it may not be advisable in other settings where it is difficult to prevent a subset of potential beneficiaries from accessing the intervention, which occurred when we used individual-level assignment in Nicaragua. Most of the interventions we evaluated required implementation at the school level. For instance, teacher training had to be offered to all teachers in the school (or in the targeted grades). In these instances, we implemented cluster randomized controlled trials, which used groups of individuals (such as schools or communities) as the unit of assignment to treatment or control groups.

Despite the lower statistical power of a design that assigned groups to receive the intervention, we found that the results could be more accurate. For example, when a program was offered at the community level, contamination was low because children in control communities were unlikely to attend program activities. Offering the program to an entire community also facilitated program take-up and consistent participation. In contrast, when random assignment was conducted at the child level, withholding services from children in the control group required program staff to monitor and report when children from the control group participated. At the same time, in order to help reach program targets for beneficiaries reached, these same staff sometimes felt conflicting pressures to provide services to any child willing to participate.

The counterfactual matters

Impact evaluations estimate the benefit to the treatment group compared to a contrast or counterfactual. What the intervention is compared to matters. In certain situations, the LAC Reads evaluations estimated the impact of a program compared to prevailing practice or to the existing education system. In others, the evaluations compared one program to another, or to no program at all, as in the case of the *Espacios para Crecer* program in Nicaragua.

For example, the evaluations of *Leer Juntos, Aprender Juntos* in Peru and Guatemala showed that the teacher training and coaching component of the program had positive effects in Peru but almost no effects in Guatemala. Although the study used similar design and data collection protocols in both countries, in Peru the intervention was markedly distinct from the communicative textual pedagogical approach to reading instruction followed by Peru's Ministry of Education. In Guatemala, the *Leer Juntos, Aprender Juntos* approach was more in line with the Ministry of Education's approach to reading instruction, which emphasizes foundational reading skills to promote fluency and comprehension. The evaluation results suggest that the *Leer Juntos, Aprender Juntos* program was effective when it was distinct from other teaching approaches already in use—as in Peru—and that other local prevailing practices may be as effective as the program when they share core components of foundational reading skills instruction—as in Guatemala.



Students' language background may have contributed to differences in program effectiveness between the two countries, by making the Quechua language-speaking skills of teachers in Peru less critical to the success of the program than K'iche' language-speaking skills of Guatemalan teachers. In Peru, despite the high prevalence of Quechua spoken at home, most students in the sample (92 percent) demonstrated proficiency in Spanish at baseline, when they were starting first grade. In contrast, relatively few first-grade students in Guatemala—only 32 percent of the students in the sample—demonstrated proficiency in oral Spanish language skills at baseline. Despite the differences in proficiency in the language of instruction for first-grade students between the two countries, in both Peru and Guatemala teachers primarily used Spanish for reading instruction, even those who were proficient in local languages.

Parental literacy levels and engagement in school—as well as the pre-primary education opportunities for children—were higher in Peru than in Guatemala; these factors, too, may have affected student achievement.

The evaluation of *Amazonía Lee* in Peru also found different effects on reading outcomes in the two regions where the evaluation was conducted, which likely related to differences in the type of instruction offered to the control group. In the San Martín region, the control group received services from *Soporte Pedagógico*—the Peruvian Ministry of Education's flagship education quality improvement program. These services included workshops to strengthen first- through sixth-grade teachers' reading, math, and social studies instruction; remedial academic support for underperforming students; and other related services to improve teachers' performance and students' academic outcomes. In the Ucayali region, the control group did not receive additional support beyond what the Ministry of Education typically provides. The evaluation showed positive impacts when *Amazonía Lee* was compared to the prevailing practice on reading instruction in Ucayali, but not when it was compared to *Soporte Pedagógico*.

Finally, the intervention in Honduras aimed to improve teachers' use of formative and summative assessments in urban and rural primary schools. Providing support to teachers and principals in the use of formative and summative assessments to improve instruction led to improvements in math and reading test scores. But the support of summative assessments was more effective in urban schools, while the support of formative assessments was more effective in rural schools.