The Effects of Voluntary Summer School on Test Scores in the Dominican Republic

Carolina Concha-Arriagada, Guthrie Gray-Lobe, and Michael Kremer

July 2026

We evaluate the impact of a government-run remedial summer school program in the Dominican Republic. Randomizing invitations among students in grades three to five, we find that attending nineteen days of summer camp is associated with a 0.09 SD increase in test scores—a fifth of a year of learning in less than one-tenth of the time. Students in the comparison group did not lose numeracy or literacy skills over the summer, but neither did they make meaningful progress; substantial gains occurred only among students receiving summer-school instruction. Attendance was low in an initial pilot year, moderating aggregate learning gains from the program. Changes including subsidized transportation and expanded meal services improved utilization, reducing the cost per student-day attended.

Keywords: Education, Summer School, Foundational Learning, Remediation, Literacy, Numeracy

JEL Codes: I21, I25

Concha-Arriagada: Columbia University. Gray-Lobe: University of Chicago. Kremer: University of Chicago.
This project received IRB approval from the University of Chicago (IRB24-1522); and was preregistered in the AEA RCT registry with ID 0015257 (https://www.socialscienceregistry.org/trials/15257) (Concha-Arriagada, Gray-Lobe, and Kremer 2025).

1. Introduction

Many students fall behind academically early in their schooling. Public education systems in low- and middle-income countries (LMICs) rely primarily on grade retention to address these learning gaps, requiring students to repeat a grade before advancing. Yet retention is costly and has potentially negative impacts on students. Motivated by concerns of adverse effects, many systems prohibit retention in the early grades. Without effective alternatives for addressing learning gaps, however, these policies allow more students to advance without the foundational skills needed for subsequent learning.

Remedial summer school is another common tool to support students who are falling behind. When regular classes are not in session, students who are falling behind attend special sessions focused on the skills students need to proceed to higher grades. Such programs can be expensive, both in terms of teacher and student time. Teachers typically must be paid to come to school, and students must be persuaded to come to class when many of their friends may be enjoying their holiday. Meanwhile, the costs of running such programs and the limited amount of vacation time itself mean that these programs must achieve a lot in terms of learning in a short amount of time to be worthwhile. Because many of the costs of establishing a summer school are fixed, low attendance attenuates the aggregate benefits of summer school.

We study a voluntary remedial summer school program in the Dominican Republic, where more than a quarter of fifteen-year-olds report having repeated a grade. The camp gives low-performing students in grades 3–5 three to four weeks of full-day instruction in mathematics and Spanish during the summer vacation, using public-school teachers and buildings.

We leverage the random assignment of invitations to eligible students in a sample of over 10,000 students enrolled in 320 schools to estimate the effect of attending one of 52 summer school centers. Because attendance was voluntary and compliance imperfect, we use the invitation as an instrument for days attended, which identifies the average causal response—the per-day return to attendance.

Over half of students do not attend, with attendance especially low among Haitian nationals, older grade students, and those with higher test scores. The effect of invitations on test scores is 0.06 standard deviations. Assuming effects scale linearly with the days of summer school attended, the full nineteen-day program raised aggregate test scores by 0.09 standard deviations, with similar effects in mathematics and Spanish. These impacts represent a 9 percent increase in the fraction who can read a passage with comprehension, a 24 percent increase in the fraction of students who can do subtraction, and a 13 percent increase in the fraction who can do addition.

We discuss how the government modified the program after an earlier experimental pilot showed similar learning gains from camp attendance but found low attendance. The government modified features of the program to improve attendance based on evidence from the pilot. The government provided transportation for students to the summer school centers, reduced the catchment area of individual centers, increased expenditure on meals, and reduced staffing to adjust for lower-than-initially-forecasted attendance. The government also increased the duration of the program from 15 to 19 days. These changes together led to over a 60 percent increase in the number of days students were in attendance at a camp. While some of these changes increased costs, the cost per enrolled day (the assumed mechanism through which summer school affects aggregate learning) fell from 37 to 36 dollars. The example illustrates how efficiencies can be found to improve programs as they scale (Bold et al. 2018; Banerjee et al. 2017; Andersen and Hvidman 2020; Muralidharan and Niehaus 2017; Muralidharan and Singh 2025).

Our study contributes to several literatures related to grade retention policies. Several studies use regression discontinuity designs leveraging rules for assigning students to summer school based on standardized test score results to estimate the impact of being assigned to summer school and grade retention in the United States using standardized test scores. While some of these studies obtain estimates of the effect of summer school itself on academic outcomes, this is typically only possible under strong assumptions because the risk of grade retention typically changes at the same threshold Jacob and Lefgren (2004); Eren, Depew, and Barnes (2017); Matsudaira (2008).

We contribute to a small experimental literature on the impacts of voluntary summer school programs. The National Summer Learning Program evaluation, the largest-scale evaluation of a summer school program before ours, found similar positive effects on learning from voluntary summer school (McCombs et al. 2014), noting gains for mathematics of 0.09 standard deviations, but statistically insignificant effects on reading. Two other smaller randomized evaluations find statistically insignificant effects, except in more assumption-driven extensions (Borman and Dowling 2006; Chaplin and Capizzano 2006). Point estimates for learning gains are positive in all cases. Our study adds to this literature by evaluating at a larger scale than previous evaluations, allowing more precise estimates. We also provide the first experimental estimates of a summer school program in an LMIC.

The study contributes to a growing experimental literature illustrating the potential impact of structured pedagogy as a tool to improve learning outcomes (Lucas et al. 2014; Piper et al. 2018; Cilliers et al. 2020; Kerwin and Thornton 2021; Cilliers et al. 2022; Gray-Lobe et al. 2022). We evaluate the impact of structured pedagogy in delivering summer school. Structured pedagogy may be less effective in remedial summer school because it relies on students receiving a coherent sequence of lessons, while irregular attendance can cause students to miss large portions of a short, cumulative program. Remedial students may also differ widely in their underlying skill levels, making a common scripted sequence poorly matched to some students unless instruction is differentiated. Although we do not observe a counterfactual without structured pedagogy, our results show that it can generate rapid learning gains in this setting.

Our results also speak to summer learning loss (Cooper et al. 1996) and the role of schools in education production (Coleman et al. 1966). Comparison-group students do not appear to lose substantial academic skills over the summer, consistent with the view that summer is not a period of learning loss (Workman, von Hippel, and Merry 2023). However, comparison students also make little academic progress, while students invited to the camps improve substantially on academic benchmarks. This pattern underscores the importance of schools in producing academic skills: nonschool environments may

sustain existing skills, but academic progress appears to depend critically on schools.

We add to a small experimental literature on remedial alternatives to grade retention in middle-income countries. The closest comparators are the Teaching at the Right Level (TaRL) summer-camp evaluations in Bihar, India, where government teachers delivering remedial instruction over the summer produced effects of 0.07–0.09 SD (Banerjee et al. 2016), and high-dosage tutoring in the United States, which produces larger effects but costs far more per student and demands longer programs and more tutor training (Nickow, Oreopoulos, and Quan 2024; Kraft, Schueler, and Falken 2024). Our pooled effects fall within the TaRL range, which shows that public-school teachers in a different middle-income setting, using ministry-developed materials and standard camp orientation, can deliver remediation at scale with comparable results. To our knowledge, this is the first experimental evaluation of a government-designed and government-run remedial summer school program.

The paper proceeds as follows. Section 2 describes the Dominican education context, the camp, and the design changes between the 2024 and 2025 cohorts. Section 3 presents the data and checks baseline balance. Section 4 lays out the estimation framework. Section 5 reports effects on attendance, intent-to-treat and average-causal-response estimates, and heterogeneity by baseline score, grade, demographics, and overage status, and closes with the OLS–ACR comparison that bears on selection. Section 7 concludes.

2. Context and Experiment Design

2.1. Educational Context in the Dominican Republic

Despite sustained investment in primary education, learning outcomes in the Dominican Republic are among the lowest in Latin America: over 80% of sixth-grade students fail to reach minimum reading proficiency, and more than 95% do not meet minimum standards in mathematics (LLECE 2021). This is consistent with World Bank learning poverty estimates, which place 81% of late-primary-age Dominican children below minimum reading proficiency after adjusting for out-of-school children (World Bank 2019). In PISA 2022, the

country ranked among the lowest-performing participants globally, with only 8 percent of 15-year-old students reaching minimum proficiency in mathematics — far below the OECD average of 69 percent.[1]

The primary policy tool educators have to address low learning levels is grade retention. Students can be retained in grade due to low performance starting in third grade. Twenty-six percent of Dominican students reported having repeated at least once. According to MINERD administrative records for the 2023–2024 school year, almost 30 percent of students in grades 3–5 are at least one year behind their age-appropriate grade level, with a failure rate of 7.2 percent and a dropout rate of 1.3 percent in the same grades.

Grade retention is costly. Each additional year of schooling that retention generates costs the public system roughly USD 2,700 per student. Furthermore, the academic benefits of the policy are unclear. While a few regression discontinuity studies find positive impacts of grade retention on test scores of marginal students (Jacob and Lefgren 2004; Schwerdt, West, and Winters 2017), there is also evidence of adverse effects on outcomes like discipline, dropout (Jacob and Lefgren 2009; Manacorda 2012), and violent crime (Eren, Depew, and Barnes 2017).

Without the setting of this study, grade failure rates are highest among those students who are overage for their grade: 5 percent among on-age students, 11 percent among students one year overage, and 14 percent among students two or more years overage. While this gradient may partly reflect selection it is consistent with the view that grade retention has adverse impacts on academic. It also motivates the study's focus (described below) on students who are overage for their grade.

The regions selected by MINERD to participate in the summer camp program have greater academic need and socioeconomic disadvantage. Participating regions have slightly higher shares of overage students in grades 3 to 5 — approximately 31 percent — and higher failure rates approaching 9 percent. These regions also have a larger share of rural schools, reaching 34 percent in the 2024 participating regions compared to 24

[1]These deficits are not merely an artefact of international testing. National standardized assessments at the primary level confirm these deficits — 61 percent of 3rd-grade students scored at the elementary level in Spanish and 59 percent in math, meaning they had not acquired foundational competencies.

percent nationally. On the 2024 national diagnostic assessment (*Pruebas Diagnósticas*), students in participating regions score 5 to 8 points below the national mean in both subjects, with a larger share at the elementary proficiency level. Participating regions are also more socioeconomically disadvantaged: in the 2024 regions, 31 percent of students were in the lowest socioeconomic quintile, compared to 18 percent nationally. Moreover, participating regions exhibit higher failure rates within every overage category, suggesting that the program targets not only regions with more overage students but also regions with worse academic outcomes conditional on overage status.

Finally, participating regions serve a disproportionate share of Haitian migrant students. Haitian students constitute approximately 8 percent of the national student population in grades 3–5, but between 10 and 12 percent in participating regions. These students may face additional barriers to learning, including language differences, as Spanish is the language of instruction in Dominican public schools. We examine whether program effects differ for this population in Section 5.

2.2. Program Design and Implementation

The program established 52 summer school centers in 49 districts (out of 122). The camp ran for four weeks (19 days of school) from 8:30 a.m. to 3:30 p.m. Class sizes were approximately 25 students, with two teachers assigned to each classroom. School bus transportation was provided. Students received three daily meals and an afternoon snack.[2]

2.2.1. Pedagogy

The camp curriculum focused on Spanish and mathematics, with content aligned to the national primary curriculum and designed to reinforce foundational literacy and numeracy skills. In Spanish, instruction covered reading comprehension across different text genres, identification of text structure and purpose, and written production at increasing levels of complexity. In mathematics, instruction progressed from numeration and place value through basic operations (addition, subtraction, multiplication, and division) to

[2]For details of camp staff's vision of on-site implementation, see Appendix **??**.

problem-solving using arithmetic.

Course material alternated between mathematics and Spanish instruction blocks over the course of the program.

Instructional methods combined interactive games and activities and smaller-group instruction for targeted support. Students facing particular difficulties were temporarily separated for more focused reinforcement during Spanish and, in some cases, mathematics sessions. Daily activities were adjusted based on observed learning levels.

Students were grouped into three performance levels (Exploradores, Creadores, and Innovadores), with objectives tailored to each level and teachers expected to adjust the depth and pacing of instruction to students' skill levels.[3][4] In Spanish, students worked with instructive (procedural), narrative, and informational texts, with reading difficulty and writing tasks varying by performance level, from basic sentence construction to passages. In mathematics, instruction covered number sequences and place value, addition and subtraction, and multiplication and division, with differentiation through varying number ranges and levels of operational complexity—for example, the lowest level worked within 1–20 while higher levels extended to 999 and beyond.[5]

The course also included socio-emotional learning, identity development, technology use, arts, and values education. Activities were organized weekly around thematic axes and adapted to students' performance levels, with continued emphasis on interactive, participatory, and game-based learning.

2.3. Experimental Design

This subsection describes the sampling strategy used to identify eligible students in each cohort, followed by the randomization procedure used to assign students to Invited and Comparison groups.

[3] Students could be assigned to any performance level regardless of grade, although different score cutoffs were used for assignment to levels within each grade.

[4] During the first week, students who had not previously completed the baseline diagnostic took it at the camp, and these results were used to finalize classroom assignments.

[5] For details on materials and notebooks, see Appendix B.

2.3.1. Sampling

The Ministry of Education (MINERD) selected participating schools based on indicators of academic need, including the number of overage students, grade repetition rates, and average test scores. Schools up to 5 kilometers from a selected camp were eligible.

Students in grades three to five from schools near the centers were eligible if they were determined to be overage (older than the age of a student who started primary school at age five and progressed one grade each year). This included students who either started late or were held back at some point after starting. Using records from the national student information system (SIGERD), the ministry identified students who were at least one year behind their age-appropriate grade. Selection followed a prioritized rule: all students with three or more years of overage were included first, followed by those with two years of overage, and finally a random sample of students with one year of overage until district-level quotas were met.

In the resulting eligible student sample, 38 percent were one year overage, 32 percent were two years overage, and 30 percent were three or more years overage.

2.3.2. Randomization Procedure

Eligible students were assigned to receive an invitation to attend the summer camp or to a comparison group, with approximately 60 percent receiving an invitation to participate. Randomization was at the student level, stratified to ensure balance on observable characteristics.

The initial stratification was defined by the interaction of district, assigned camp center, school, baseline tests attendance, and nationality. A minimum cell size of 10 students was imposed, and the procedure iteratively collapsed stratification dimensions — first nationality, then baseline tests attendance — in cells that fell below this threshold. The resulting stratification grouped students by district, camp center, and school, with nationality and baseline attendance retained only where cell sizes permitted. This process produced 476 final strata, within which 6,793 students were invited and 4,518 are part of

the comparison group.

3. Data

Our analysis relies on three main sources of data: administrative records, camp's daily attendance, and assessments. We use these records to create the main outcomes of interest and the main variable that represent the dosage that students received during the remedial summer camp. Additionally, from administrative records we obtain student demographics and baseline covariates, including age, overage status, nationality, and gender.

MINERD collected baseline and endline assessments in mathematics and Spanish to measure students' foundational numeracy and literacy progress. Baseline tests were administered in May 2025, before the summer camp. The instruments were developed by MINERD to measure skills aligned with the national primary curriculum and incorporated items from the LAYS Item-Level Design Bank. In mathematics, the assessment covered core numeracy skills as well as place-value decomposition, number ordering, time calculation, and graphical data interpretation. In Spanish, the assessment covered foundational literacy skills and included a listening-comprehension item developed following guidelines from the *Foundational Reading Assessment Tool* (UNICEF, J-PAL, Pratham).

Endline tests for both subjects were administered in November 2025, approximately four months after the conclusion of the camp. This timing means our estimates capture short-term persistence of learning gains rather than immediate post-treatment effects. The endline mirrors the structure and skill domains of the corresponding baseline instrument, with parallel items designed to measure the same underlying constructs. A detailed item-by-item comparison across rounds and cohorts is provided in Appendix C.

For each subject, cohort, and time of the test (baseline/endline), we estimate a two-parameter logistic (2PL) IRT model and recover student *latent ability scores* (θ_i). These ability measures, standardized to the comparison group mean within each cohort, are our primary academic outcome variables. We report results for mathematics, Spanish, and an aggregate measure constructed as the simple average of the standardized math and Spanish scores. Because we standardize outcomes within cohort, differences in instrument

composition across years do not affect comparability of treatment effect estimates. Details on model estimation, item screening, and reliability analysis are provided in Appendix E.

Finally, daily student attendance was recorded in the classroom throughout the summer camp. Attendance sheets were maintained by camp staff and digitized directly by them.

TABLE 1. Balance Table

	Comparison		Invited		Difference	
	Mean	SE	Mean	SE		
	(1)	(2)	(3)	(4)	(4)-(2)	P-value
Age (in years)	11.242	0.016	11.254	0.011	0.012	0.651
1 year overaged	0.379	0.005	0.373	0.003	-0.006	0.421
2 year overaged	0.317	0.005	0.317	0.004	0.000	0.991
At least 3 years overaged	0.304	0.005	0.310	0.003	0.006	0.464
Underlying ability in math	-0.000	0.009	0.013	0.006	0.013	0.416
Underlying ability in spanish	-0.000	0.010	0.008	0.006	0.009	0.579
Haitian	0.205	0.004	0.208	0.003	0.003	0.595
Missing nationality	0.011	0.001	0.010	0.001	-0.001	0.516
Female	0.366	0.005	0.365	0.004	-0.000	0.961
Number students	4,518		6,793			
p-value from joint significance test						0.983

Note: This table shows baseline characteristics by treatment status. We report means, standard errors, and pairwise differences between treatment and comparison groups. Estimates are from an OLS regression of each baseline variable on the treatment indicator, including stratum fixed effects. The sample includes all students selected by the Ministry of Education to participate in the study in 2025. For students with missing baseline assessments, we impute the within-cohort mean (zero by construction since baseline scores are standardized within cohort). Standard errors are clustered at the school level. We also present p-values from a joint significance test of all variables included in the balance test. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 1 presents baseline characteristics by treatment status. Differences between invited and comparison groups are small in magnitude and statistically insignificant across all covariates, including age, overage status, baseline test scores in math and Spanish, nationality, and gender. A joint F-test of the null hypothesis that all covariates simultaneously predict treatment assignment fails to reject.

Follow-up rates were 70 percent at endline, with no difference in the rate of attrition between comparison and invited students. We do find that when adjusting for controls, invited students are slightly more likely be assessed, so we conduct sensitivity analysis

using Lee bounds.

4. Estimation Framework

The primary empirical question of this paper is whether attending the summer improves test scores. We assume that the impact of the camps scales approximately linearly with attendance. A naive approach to estimating the impact of the camp would be to use the variation in the number of days that the student attended the camp in

$$Y_i = \alpha + \beta D_i + \varepsilon_i \,, \tag{1}$$

where y_i is the test score of student i, $D_i \in \{0, 1, ..., \bar{D}\}$, is the number of days that the student attends the camp (with $\bar{D}$ in our case being either 15 in 2024 or 19 in 2025), and ε_i represents the influence of unobserved factors. Estimation of β in 1 by ordinary least squares can lead to biased estimates if D_i is correlated with ε_i. Lower-performing students may disproportionately attend the camps if they can benefit more from them. Meanwhile, the skills and supportive structures at home that allow students to attend the camp regularly may be more prevalent among high-performing students, generating positive selection among attendees.

The analysis presented in this study is centered around the impact of the invitation to attend the camp. As we will show, invitations to the camp increased camp attendance D_i and are uncorrelated with the error term ε because they were randomly assigned. Specifically, we estimate the following model of test scores as a function of invitations:

$$Y_i = \delta Z_i + X_i'\Gamma + \eta_i \tag{2}$$

where X_i contains randomization strata fixed effects, as well as controls for grade and baseline test scores. The estimand δ in Equation 2 is referred to as the intent-to-treat (ITT), insofar as the offers to the camp were meant to ration access to attend the camp (D_i).

Using Z_i as an instrument for D_i allows us to estimate the average causal response of D_i on test scores. The ITT effect and the IV estimate are connected by the first-stage effect of invitations on attendance

$$D_i = \pi Z_i + X_i'\Omega + \upsilon_i \,. \tag{3}$$

The estimand π reflects the increase in attendance induced by offering an invitation.

Dropping the controls in X_i for expositional simplicity, the Wald estimand identifies a weighted average effect of increasing D_i by one unit (Theorem 4.5.3 Angrist and Pischke

2009). Specifically,

$$\frac{E[Y_i|Z=1]-E[Y_i|Z=0]}{E[D_i|Z=1]-E[D_i|Z=0]} = \sum_{d=1}^{\bar{D}} \omega_d E[Y_i(d) = Y_i(d-1)|D_i(1) \geq d > D_i(0)] \,, \tag{4}$$

where $Y_i(d)$ represents the potential outcome when attending d days of the camp, $D_i(z)$ represents the potential attendance level for individual i with invitation status $z \in \{0,1\}$, and $\omega_d = \frac{P[D_i(0)\geq d > D_i(0)]}{\sum_{j=1}^{\bar{D}} P[D_i(1)\geq j > D_i(0)]}$. We report estimates of the ACR in Equation 4 scaling D by 15, so that the estimate reflects the impact of attending 15 days, the full length of the 2024 program.

Identification of the ACR requires four assumptions. The first three are the standard assumptions for binary endogenous variable: relevance, independence and the exclusion restriction. The fourth assumption is monotonicity, which in the variable-intensity case is more demanding than its binary counterpart. In our setting, the first two assumptions are satisfied by construction. The exclusion restriction and monotonicity assumptions are discussed below.

The exclusion restriction requires that the random invitation affects student learning outcomes only through its effect on camp attendance. This assumption would be violated if, for example, the invitation itself conveyed information to parents — such as a signal that their child was academically at risk — that prompted changes in home learning behaviors independent of camp participation. We view these concerns as limited because invitation letters did not disclose individual performance information, and the camp took place during the school vacation period when no regular instruction was scheduled.

Monotonicity in the variable-treatment-intensity setting is more demanding than in the binary case. Rather than simply requiring that the instrument shifts all individuals in the same direction (into or out of treatment), monotonicity here requires that the invitation weakly increases the attendance dosage for every student: $D_i(1) \geq D_i(0)$ for all i. This rules out the existence of "defiers at any margin" — students who would attend fewer days if invited than if not invited. We view this assumption as plausible in our setting. The invitation provides access to a free program with no associated costs or penalties; there is no obvious mechanism through which receiving an invitation would reduce a student's attendance at the camp. The main threat would arise if invited students were displaced from alternative learning opportunities by the invitation, but since the camp operated during the school vacation period when no regular instruction was scheduled, this concern is minimal.

Under the further assumptions that the effect scales linearly with each unit and that the average marginal effects over the full population, the ACR from Equation 4 and Equation 1 reflect the same structural estimand. We compare OLS estimates of Equation 1 to the identified estimates of the ACR in section **??**. The better identification of the estimates

in Equation 4 come at the cost of less precision because the ACR is identified only from variation in attendance that arises from the invitation. OLS estimates of Equation 1 are more precise, using all variation in attendance, at the cost of potentially biased estimates due to selection effects. Comparison of OLS and ACR estimates yields a test of the selection effect under the assumption of linearity.

5. Results

5.1. Effects on Attendance

We estimate the effect of receiving an invitation to participate in the remedial summer school on the days of attendance D_i from Equation 3. Invited students attended 7.9 days more than comparison students (Table 2). There was some non-compliance among comparison students, with treated students attending on average 0.4 days. A substantial share of invited students did not attend at all, with only 49 percent attending at least one day. However, most students who attended once attended up to 15 days: nearly 42 percent of invited students attended 15 days. Twenty-nine percent of invited students attended all nineteen days of the camp.

The effect of invitations on attendance varies across sub-populations. Students whose nationality is Haitian attend 7.3 percentage points fewer summer school days, potentially reflecting concerns about deportation risk (Table A.1).[6] Students in older grades (Table A.2) and higher-performing students within grades are also less likely to attend the camps. We find no evidence that attendance varies systematically by gender. We also find no evidence that attendance varies systematically with the distance of the student (measured by the distance between the school where the student attends during the regular school session and the location of the summer school center) which may be suggestive of the effectiveness of transportation subsidies.

5.2. Effect of Summer School on Learning

Students who attended the camps for a larger share of the nineteen-day program have test scores 0.11 standard deviations above those who attend zero days, controlling for their baseline test scores (Column 1 Table 3). This difference is slightly larger for math (0.11

[6] In October 2024, the Dominican government ordered the deportation of up to 10,000 Haitian migrants per week. Over 276,000 people were deported in 2024, and more than 310,000 in the first ten months of 2025 alone, raising concerns among human rights organizations. The enforcement campaign included street sweeps and raids on workplaces, public transportation, and homes. See Inter-American Commission on Human Rights (IACHR), "IACHR concerned about rights of individuals in human mobility contexts in the Dominican Republic," Press Release No. 023/2026, February 6, 2026; Loiseau, M. (2026, February 26), "How Far Will the Dominican Republic Go in Deporting Haitians?," *Inkstick Media*.

TABLE 2. Effect of Being Invited to the Remedial Camp on Attendance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Share att.	Total Days	>= 1 days	>= 5 days	>= 10 days	>= 15 days	= 19 days
Treatment	0.526***	7.894***	0.463***	0.447***	0.432***	0.402***	0.275***
	(0.02)	(0.23)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	11,311	11,311	11,311	11,311	11,311	11,311	11,311
Control mean	0.025	0.373	0.023	0.023	0.021	0.017	0.012

Note: This table shows the effect of being invited to the remedial camp on the share of days attended out of 15 core days (Column 1), total days attended (Column 2), and the likelihood of attending at least 1, 5, 10, 15, or a total of 19 days (Columns 3–7). Estimates are from an OLS regression including stratum and baseline controls pre-specified in our pre-analysis plan, including baseline test scores and grade. We also include cohort fixed effects in the pooled analysis. Missing baseline variables are imputed at the mean, with indicator variables for missingness included. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

standard deviations) than Spanish (0.92 standard deviations). Whether these estimates identify the effect of attending summer camp rests on a selection on observables assumption. In this case, Equation 1 can be estimated directly using OLS and the coefficient β can be estimated. The formulation is similar to value-added models for teachers (Rockoff 2004; Rivkin, Hanushek, and Kain 2005) and schools (Deming 2014; Angrist et al. 2017). However, these results may be subject to selection bias if, holding baseline test scores fixed, students who attend more summer school differ in unobserved ways from those who attend less.

TABLE 3. Estimations of Remedial Camp Effects

	(1)	(2)	(3)
	OLS	ITT	ACR
Average test scores	0.110***	0.056***	0.090***
	(0.02)	(0.02)	(0.03)
Math	0.110***	0.061***	0.098***
	(0.02)	(0.02)	(0.03)
Spanish	0.092***	0.040**	0.066**
	(0.02)	(0.02)	(0.03)
Observations	7,999	7,999	7,999
Control mean	0.000	0.000	0.000

Note: This table shows the different estimators for the remedial camp effects. Column (1) reports OLS estimates of endline test scores on camp attendance. Column (2) reports intent-to-treat (ITT) estimates of the effect of being randomly assigned to the remedial camp invitation. Column (3) reports the average causal response (ACR) using treatment assignment as an instrument for camp attendance. All specifications include stratification fixed effects, year fixed effects, and baseline test score controls interacted with grade and year. Baseline test scores set to zero when missing, with missing indicators included as controls. Standard errors are clustered at the school level. * p < 0.10, ** p < 0.05, *** p < 0.01.

Students randomly invited to summer school score 0.056 standard deviations higher than comparison students on an aggregate measure of math and Spanish skills (Column 2 Table 3). The effect in math is 0.061 standard deviations, compared to 0.04 standard deviations for Spanish. We report impacts on test scores standardized to the comparison group mean within each cohort following specification 2. The coefficient of interest, δ, captures the average effect of being invited to the program on endline test scores, regardless of whether the student actually attended the camp.

Table A.3 examines heterogeneous effects of invitations by nationality, gender, and distance to the camp site. None of these interaction terms are statistically significant, indicating that the summer school invitation effect does not vary meaningfully along these dimensions. Haitian students show a negative but insignificant interaction with treatment across all three outcomes (−0.031 SD for aggregate test scores, −0.045 for math, −0.018 for Spanish), consistent with the hypothesis that the camp was similarly effective for Haitian and non-Haitian students despite the differential attendance documented in Section 5.1.

Students with 3+ years overage have lower effects in Spanish relative to students with one year overage, -0.09 SD (Table A.4).

Having established that the camp invitation significantly improved learning outcomes, we now turn to the average causal response (ACR) estimates, which recover the impact of camp attendance for compliers.

Effects of attending the camp on learning follow a similar pattern to the ITT estimates: the most overage students benefit less from Spanish instruction, with a significant negative

effect for the 3+ years category in 2025 (Table A.6). The grade 5 advantage in math also persists in the 2025 cohort (Table A.7). No significant heterogeneity appears by nationality, gender, distance to camp, or baseline test score (Tables A.5).

A Wu–Hausman test of the equality between the OLS and 2SLS estimators fails to reject exogeneity of attendance in the pooled sample for average test scores, math, and Spanish.[7]

5.3. Skill benchmarks and summer learning loss/gain

In this section, we examine the impact of summer camps on learning benchmarks related to numeracy and literacy. We then compare these impacts to how numeracy and literacy skills evolve over the summer in the comparison group.

We estimate positive gains on all measured learning levels. The impacts on learning benchmarks are similar across different measures, and in most cases the magnitude of the impacts is between 30 and 90 percent of an equivalent year of schooling (the coefficient from a univariate regression of the measured skill on the students' grade within the comparison group) (Table 4). The fraction of students who are assessed to be able to perform subtraction increased by 8.4 percentage points . In the comparison 36 percent of students are able to perform subtraction, and the magnitude of the effect is comparable to the differences observed between grades. Grade four students are about 9.5 percentage points more likely to be able to do subtraction than grade three students, and grade five students are 5.5 percentage points more likely than grade four students.

[7] Two caveats temper a strong selection interpretation of the OLS–ACR comparison. First, the equivalence of OLS and ACR rests on homogeneity of per-day returns across student types: if compliers differ from always-takers in their per-day returns, the OLS–ACR gap conflates selection on attendance with treatment-effect heterogeneity across the complier and always-taker populations, and a small gap is consistent with selectio n bias and complier-population differences offsetting one another. Second, the OLS estimand under non-linear marginal returns averages a misspecified model over the full attendance distribution, while the ACR remains a properly weighted average of margin-specific effects; in that case, the OLS–ACR gap reflects functional-form misspecification in addition to any selection on unobservables.

TABLE 4. Effect of attending summer camp on learning levels

	Math level			Reading level		
	Recognize numbers	Addition	Subtraction	Recognize letters	Read Syllables	Reading comprehension
	(1)	(2)	(3)	(4)	(5)	(6)
Full attendance effect	0.064***	0.067***	0.085***	0.038***	0.063***	0.048**
	(0.020)	(0.021)	(0.022)	(0.012)	(0.017)	(0.022)
Observations		6,302			6,523	
Comparison mean	0.694	0.531	0.357	0.886	0.776	0.547
EYS gradient	0.140	0.116	0.092	0.089	0.143	0.159

Figure A.1 reports the share of students at or above each performance level on the baseline and endline assessments, separately for the comparison and invited students. For math, we examine the share who recognize numbers, perform addition problems, and perform subtraction problems. For Spanish, we examine the share who can identify letters, decode syllables, and read a passage with comprehension. It should be noted that baseline was measured two months into the school year and so may reflect advancement in these months.

Across each skill, the share who met each performance level was fairly stable from baseline until endline in the comparison group. The share who can recognize numbers was equal across periods. The share who could perform addition fell slightly, while the share who could perform subtraction rose slightly. One notable exception is that letter identification skill rose by 6 percentage points from 83 to 90 percent. Syllable reading rose by 3 percentage points. Reading comprehension fell by 4 percentage points.

5.4. Effect of invitations on Non-academic Outcomes

We find no statistically significant effects of the summer camp invitation on any socioemotional outcome measure (see Table 5).[8] Point estimates are close to zero across all dimensions, though the signs are suggestive: emotional regulation shows the largest positive (non-significant) point estimate (0.023 SD), consistent with qualitative evidence from teacher focus groups highlighting improved self-regulation among participants. Bullying victimization carries a small negative sign (−0.015), pointing in the desired direction of reduced victimization.

[8] see Appendix D for details on the questionnaire and variables' constructions.

TABLE 5. Effects on Non-Academic Outcomes of Being Invited to the Remedial Camp

	(1)	(2)	(3)	(4)	(5)	(6)
	PCA Index	Empathy	Emotional Regulation	Academic Self-Concept	Growth Mindset	Bullying Victimization
Treatment	-0.003	-0.003	0.023	-0.019	0.001	-0.015
	(0.03)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)
Observations	5,423	7,781	7,581	7,545	7,511	5,665
Control mean	0.000	-0.005	-0.011	-0.003	-0.002	0.018

Note: This table reports intent-to-treat (ITT) estimates of the effect of being invited to the remedial camp on socioemotional outcomes for the 2025 cohort. Column (1) reports effects on the first principal component index of five socioemotional indices (with bullying victimization reverse-coded so that higher values indicate better outcomes for all indices). Columns (2)–(6) report effects on each index separately. Each index is the mean of its standardized component items: Empathy (items 1, 2, 4, 5, 15), Emotional Regulation (items 3, 6), Academic Self-Concept (items 8, 9), Growth Mindset (items 11, 12), and Bullying Victimization (items 13, 14). Items 1 and 12 are reverse-coded. All regressions include randomization stratum fixed effects, grade fixed effects, and baseline math and Spanish test scores interacted with grade (with missing value imputation and flags) as controls. Standard errors clustered at the school level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6. Costs, learning, and scaling

The costs of summer camps are substantial, requiring time from teachers and students. We estimate that the cost per student hinge significantly on the attendance rate. Training and staffing camps and other camp inputs are fixed and must be made based on forecasts of the number of students who will attend. When attendance is low (capacity is underutilized) the aggregate effect of the camp is reduced.

In 2024, the government ran a similar experiment to one evaluated here, finding similar (but less precise) estimates of the impact on learning.

It began providing transportation subsidies after observing that students living farther from the center were much less likely to attend. It increased expenditure on meals to make the camps more attractive.

Subsidizing transportation and improving meal provision were costly, but some of this cost was offset by reducing staffing requirements at the school after revising forecasts of the number of students who would attend. Similarly, the government was able to invite more students. Having seen that attendance rates were low, the government could improve utilization rates even more by increasing the number of students invited.

Table 6 summarizes the key design differences between the 2024 and 2025 implementations. While core features of the program — including daily schedule, class size, and treatment share — remained largely unchanged, the 2025 cohort introduced several important modifications, including a 30-min longer instructional period, explicit curriculum differentiation by performance level, transportation provision, and the use of administrative records for student identification.

The higher compliance in 2025 likely reflects both the reduced contamination and the design improvements summarized in Table 6, including the provision of transportation, shorter distances to camp sites, reorganization of classrooms by student level, and the addition of meals. Overall, the utilization rate — defined as the share of total student-days of capacity actually attended — rose from 52 percent in 2024 to 61 percent in 2025. Among invited students only, the increase was sharper, from 42 to 59 percent, suggesting the program improved from one year to another.

TABLE 6. Program Design Comparison: 2024 vs. 2025

Feature	**2024 Cohort**	**2025 Cohort**
Districts (camp sites)	29 (29)	49 (52)
Instructional days	15	19
Daily schedule	8:30 a.m.–3:30 p.m.	8:00 a.m.–3:30 p.m.
Student grouping	By grade	By performance level (3 tiers)
Curriculum differentiation	None	Explicit (3 leveled workbooks)
Student identification	Teacher nomination	Administrative records (SIGERD)
Transportation provided	No	Yes
Max distance to camp	25 km	5 km
Meals per day	3	3 + afternoon snack
Socio-emotional component	No	Yes
Counselor/psychology support	No	Yes (select sites)
Class size	≈25 (2 teachers)	≈25 (2 teachers)
Study sample	5,482 students	11,311 students
Participating schools	302 schools	320 schools
Treatment share	61%	60%

7. Conclusion

We show that attending the full course of a government-designed and government-led voluntary remedial summer school raised test scores by 0.09 standard deviations low-performing primary school students. Dominican public school teachers selected for their high results in FLN during the previous academic year, working with MINERD-provided materials and training, generated learning gains after the one-month program.

The magnitude of impacts is fairly small, but should be considered within the context of the amount of time that it took to produce. The effect estimated for the impact of attending the full summer school program places it at the median for education RCTs in low- and middle-income countries (Evans and Yuan 2022) and fall in the range that Kraft (2020) characterizes as educationally meaningful for interventions delivered at scale. Yet, they were produced with only a month of teacher and student time.

These results suggest that short, intensive remedial camps offer a viable complement

to the regular school year for addressing foundational learning gaps. The program operates during the summer vacation without displacing regular instructional time, uses existing public school teachers and infrastructure, and can be scaled through the existing administrative apparatus.

The reliance of the program on teachers identified based on strong performance raises some questions about scalability. To the extent that the effect is driven by the quality of teachers, expansion of the program may require use of teachers who do not meet the current criteria. While teachers were recruited locally, most head teachers felt that the program still did not serve all students who needed it, so that local expansion was needed to meet the demands within the system.

That over fifty percent of students do not attend summer school even when invited raises questions about whether voluntary programs can reach all students.

The one substantive exception to the otherwise uniform pattern of effects — smaller gains in Spanish among students with three or more years of overage — points to a subgroup whose distinct barriers may warrant curriculum adaptation or complementary interventions in future iterations. While we find that never-takers are somewhat positively selected on baseline test scores, it is still possible that some students who would benefit from the program are not attending.

References

Andersen, Simon Calmar and Ulrik Hvidman. 2020. "Implementing Educational Interventions at Scale." HCEO Working Paper 2020-039, Human Capital and Economic Opportunity Working Group.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. "Leveraging Lotteries for School Value-Added: Testing and Estimation." *The Quarterly Journal of Economics* 132 (2): 871–919. 10.1093/qje/qjx001.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India." *NBER Working Paper No. 22746* .

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives* 31 (4): 73–102.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2018. "Experimental Evidence on Scaling Up Education Reforms in Kenya." *Journal of Public Economics* 168: 1–20.

Borman, Geoffrey D. and N. Maritza Dowling. 2006. "Longitudinal Achievement Effects of Multiyear

Summer School: Evidence from the Teach Baltimore Randomized Field Trial." *Educational Evaluation and Policy Analysis* 28 (1): 25–48. 10.3102/01623737028001025.

Chaplin, Duncan and Jeffrey Capizzano. 2006. "Impacts of a Summer Learning Program: A Random Assignment Study of Building Educated Leaders for Life (BELL)." Tech. rep., Urban Institute, Washington, DC.

Cilliers, Jacobus, Brahm Fleisch, Janeli Kotze, Nompumelelo Mohohlwane, Stephen Taylor, and Tsegofatso Thulare. 2022. "Can Virtual Replace In-Person Coaching? Experimental Evidence on Teacher Professional Development and Student Learning." *Journal of Development Economics* 155: 102815. 10.1016/j.jdeveco.2021.102815.

Cilliers, Jacobus, Brahm Fleisch, Cas Prinsloo, and Stephen Taylor. 2020. "How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and In-Classroom Coaching." *Journal of Human Resources* 55 (3): 926–962. 10.3368/jhr.55.3.0618-9538R1.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. 1966. "Equality of Educational Opportunity." Tech. rep., U.S. Department of Health, Education, and Welfare, Office of Education, Washington, DC.

Cooper, Harris, Barbara Nye, Kelly Charlton, James Lindsay, and Scott Greathouse. 1996. "The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review." *Review of Educational Research* 66 (3): 227–268. 10.3102/00346543066003227.

Deming, David J. 2014. "Using School Choice Lotteries to Test Measures of School Effectiveness." *American Economic Review* 104 (5): 406–411. 10.1257/aer.104.5.406.

Eren, Ozkan, Briggs Depew, and Stephen Barnes. 2017. "Test-based promotion policies, dropping out, and juvenile crime." *Journal of Public Economics* 153 (C): 9–31. https://EconPapers.repec.org/RePEc:eee:pubeco:v:153:y:2017:i:c:p:9-31.

Evans, David K. and Fei Yuan. 2022. "How Big Are Effect Sizes in International Education Studies?" *Educational Evaluation and Policy Analysis* 44 (3): 532–540.

Gray-Lobe, Guthrie, Anthony Keats, Michael Kremer, Isaac Mbiti, and Owen W. Ozier. 2022. "Can Education Be Standardized? Evidence from Kenya." Working Paper 2022-68, Becker Friedman Institute for Economics at the University of Chicago. 10.2139/ssrn.4129184.

Jacob, Brian A. and Lars Lefgren. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics* 86 (1): 226–244.

Jacob, Brian A. and Lars Lefgren. 2009. "The Effect of Grade Retention on High School Completion." *American Economic Journal: Applied Economics* 1 (3): 33–58. 10.1257/app.1.3.33.

Kerwin, Jason T. and Rebecca L. Thornton. 2021. "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." *Review of Economics and Statistics* 103 (2): 251–264. 10.1162/rest_a_00911.

Kraft, Matthew A. 2020. "Interpreting effect sizes of education interventions." *Educational researcher* 49 (4): 241–253.

Kraft, Matthew A., Beth E. Schueler, and Grace T. Falken. 2024. "What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability." EdWorkingPaper No. 24-1031, Annenberg Institute at Brown University.

LLECE. 2021. "Estudio Regional Comparativo y Explicativo (ERCE 2019): Reporte Nacional de Resultados – República Dominicana." Tech. rep., UNESCO Office Santiago, Regional Bureau for Education in Latin America and the Caribbean, Santiago. https://unesdoc.unesco.

org/ark:/48223/pf0000380257.

Lucas, Adrienne M., Patrick J. McEwan, Moses Ngware, and Moses Oketch. 2014. "Improving Early-Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda." *Journal of Policy Analysis and Management* 33 (4): 950–976. 10.1002/pam.21782.

Manacorda, Marco. 2012. "The Cost of Grade Retention." *Review of Economics and Statistics* 94 (2): 596–606.

Matsudaira, Jordan D. 2008. "Mandatory summer school and student achievement." *Journal of Econometrics* 142 (2): 829–850.

McCombs, Jennifer Sloan, John F. Pane, Catherine H. Augustine, Heather L. Schwartz, Paco Martorell, and Laura Zakaras. 2014. "Ready for Fall? Near-Term Effects of Voluntary Summer Learning Programs on Low-Income Students' Learning Opportunities and Outcomes." Tech. Rep. RR-815-WF, RAND Corporation, Santa Monica, CA. 10.7249/RR815.

Muralidharan, Karthik and Paul Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives* 31 (4): 103–124.

Muralidharan, Karthik and Abhijeet Singh. 2025. "Adapting for Scale: Experimental Evidence on Technology-Aided Instruction in India." NBER Working Paper 34205, National Bureau of Economic Research, Cambridge, MA. 10.3386/w34205. Revised November 2025.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan. 2024. "The Promise of Tutoring for PreK–12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence." *American Educational Research Journal* 61 (1): 74–107.

Piper, Benjamin, Stephanie Simmons Zuilkowski, Margaret Dubeck, Evelyn Jepkemei, and Simon J. King. 2018. "Identifying the Essential Ingredients to Literacy and Numeracy Improvement: Teacher Professional Development and Coaching, Student Textbooks, and Structured Teachers' Guides." *World Development* 106: 324–336. 10.1016/j.worlddev.2018.01.018.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–458. 10.1111/j.1468-0262.2005.00584.x.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94 (2): 247–252. 10.1257/0002828041302244.

Schwerdt, Guido, Martin R. West, and Marcus A. Winters. 2017. "The Effects of Test-Based Retention on Student Outcomes over Time: Regression Discontinuity Evidence from Florida." *Journal of Public Economics* 152: 154–169. 10.1016/j.jpubeco.2017.06.004.

Workman, Joseph, Paul T. von Hippel, and Joseph Merry. 2023. "Findings on Summer Learning Loss Often Fail to Replicate, Even in Recent Data." *Sociological Science* 10: 251–285. 10.15195/v10.a8.

World Bank. 2019. "Dominican Republic Learning Poverty Brief." Tech. rep., World Bank Group, Washington, DC. https://documents1.worldbank.org/curated/en/994181637665163043/pdf/Dominican-Republic-Learning-Poverty-Brief-2019.pdf.

Appendix A. Appendix Tables and Figures

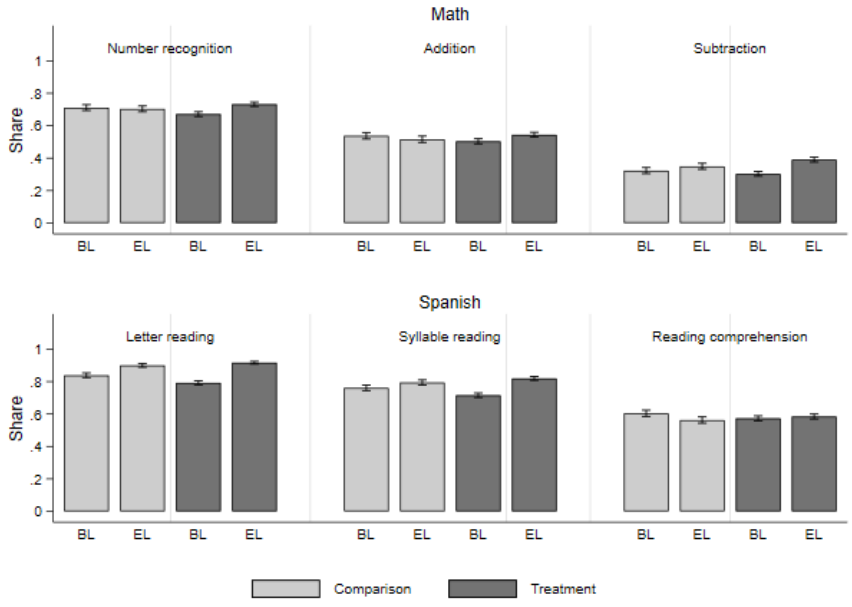


FIGURE A.1. Performance on baseline (BL) and endline (EL) learning benchmarks.

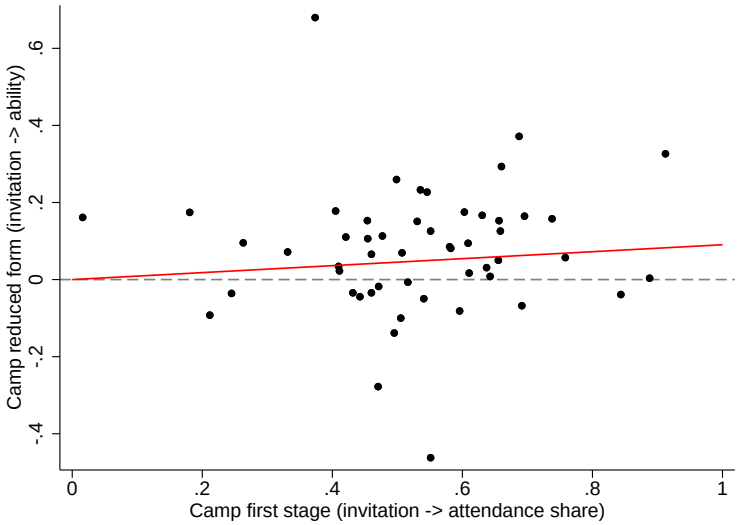


FIGURE A.2. Camp-specific reduced forms against first stages.

Notes: Each point is one of the 52 camps in the 2025 cohort. The horizontal axis is the camp's first stage — the within-camp coefficient of attendance (share of the program attended) on the invitation; the vertical axis is the camp's reduced form — the within-camp coefficient of endline aggregate ability (in standard deviations) on the invitation. Both come from separate per-camp regressions with stratum fixed effects and robust standard errors. The solid line is the pooled average causal response drawn through the origin, the proportional relationship that would hold for every camp if effects were homogeneous. Scatter about the line beyond sampling error reflects cross-camp heterogeneity in effectiveness.

TABLE A.1. Heterogeneous Effects of Being Invited to the Remedial Camp on Attendance (2025)

	(1) Share att.	(2) Total Days	(3) >= 1 days	(4) >= 5 days	(5) >= 10 days	(6) >= 15 days	(7) = 19 days
***Panel A**: Heterogeneity by Gender*							
Female	0.004	0.054	-0.000	0.002	0.004	0.004	0.003
	(0.01)	(0.12)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Treatment	0.525***	7.873***	0.462***	0.447***	0.433***	0.399***	0.272***
	(0.02)	(0.25)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Treatment x Female	0.004	0.058	0.002	-0.001	-0.002	0.009	0.008
	(0.01)	(0.22)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.170	2.546	0.145	0.148	0.142	0.130	0.085
Observations	11,311	11,311	11,311	11,311	11,311	11,311	11,311
***Panel B**: Heterogeneity by Nationality*							
Haitian	-0.005	-0.072	-0.007	-0.005	-0.004	-0.005	0.001
	(0.02)	(0.23)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Treatment	0.548***	8.215***	0.482***	0.466***	0.450***	0.417***	0.284***
	(0.02)	(0.25)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Treatment x Haitian	-0.088***	-1.317***	-0.081***	-0.082***	-0.074***	-0.059***	-0.031
	(0.03)	(0.38)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.173	2.588	0.147	0.151	0.145	0.133	0.085
Observations	11,191	11,191	11,191	11,191	11,191	11,191	11,191
***Panel C**: Heterogeneity by Camp Being Host at Student's Regular School*							
Treatment	0.522***	7.835***	0.457***	0.443***	0.429***	0.399***	0.276***
	(0.02)	(0.26)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Treatment x Zero Distance	0.020	0.304	0.030	0.022	0.017	0.014	-0.008
	(0.04)	(0.57)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Constant	0.172	2.573	0.145	0.149	0.144	0.132	0.086
Observations	11,311	11,311	11,311	11,311	11,311	11,311	11,311
***Panel D**: Heterogeneity by Distance between Camp and Student's Regular School*							
Treatment	0.513***	7.698***	0.455***	0.439***	0.424***	0.393***	0.249***
	(0.02)	(0.34)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Treatment x Distance	0.009	0.134	0.006	0.005	0.005	0.006	0.018*
	(0.01)	(0.18)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.171	2.571	0.145	0.149	0.144	0.132	0.086
Observations	11,311	11,311	11,311	11,311	11,311	11,311	11,311

Note: This table shows heterogeneous effects of being invited to the remedial camp on attendance for the 2025 cohort. Each panel reports estimates from a separate OLS regression in which the treatment indicator is interacted with a student characteristic: nationality (Panel A), gender (Panel B), whether the camp is located in the student's school from the previous academic year (Panel C), and distance in kilometers between the student's school and the camp site (Panel D). Columns report results for the share of days attended out of 15 (Column 1), total days attended (Column 2), and the likelihood of attending at least 1, 5, 10, 15, or a total of 19 days (Columns 3–7). All specifications include stratum fixed effects and baseline controls pre-specified in our pre-analysis plan, including baseline test scores and grade. The main effects of distance and camp at student's school are absorbed by the stratum fixed effects as these variables do not vary within strata. In 2025, the reduction in maximum distance to camp sites to less than 5 kilometers substantially reduced variation in the distance variable. The sample includes students selected by the Ministry of Education to participate in the 2025 remedial camp, including students for whom baseline test scores are unavailable. Missing baseline variables are imputed at the mean, with indicator variables for missingness included. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.2. Heterogeneous Effects of Being Invited to Summer School on Attendance by Grade

	(1) Share att.	(2) Total Days	(3) >= 1 days	(4) >= 5 days	(5) >= 10 days	(6) >= 15 days	(7) = 19 days
Treatment (β_T)	0.573***	8.600***	0.499***	0.484***	0.469***	0.446***	0.309***
	(0.02)	(0.32)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Treatment × Grade 4 ($\beta_{T,4}$)	-0.058***	-0.867***	-0.037**	-0.043**	-0.047**	-0.056***	-0.047***
	(0.02)	(0.33)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Treatment × Grade 5 ($\beta_{T,5}$)	-0.069***	-1.035***	-0.058***	-0.056***	-0.052***	-0.062***	-0.046***
	(0.02)	(0.30)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Joint F-test	6.21	6.21	5.54	5.47	5.09	7.69	5.08
	0.002	0.002	0.004	0.005	0.007	0.001	0.007
$\beta_T + \beta_{T,4}$	0.516	7.733	0.462	0.440	0.422	0.389	0.262
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\beta_T + \beta_{T,5}$	0.504	7.565	0.441	0.428	0.417	0.384	0.264
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Observations	11,311	11,311	11,311	11,311	11,311	11,311	11,311

Note: This table shows heterogeneous effects of being invited to the remedial camp on attendance by grade. Estimates are from an OLS regression in which the treatment indicator is interacted with grade indicators. Columns report results for total days attended (Column 1) and the likelihood of attending at least 1, 5, 10, 15, or a total of 19 days (Columns 2–6). All specifications include cohort and stratum fixed effects and baseline controls pre-specified in our pre-analysis plan, including baseline test scores interacted with grade. The sample includes all students selected by the Ministry of Education to participate in the study, including students for whom baseline test scores are unavailable. Missing baseline variables are imputed at the mean, with indicator variables for missingness included. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.3. Heterogeneous Effects on Learning of Being Invited to the Remedial Camp (2025)

	(1) Aggregate Ability	(2) Math	(3) Spanish
***Panel A**: Heterogeneity by Gender*			
Female	0.037	-0.052*	0.119***
	(0.03)	(0.03)	(0.03)
Treatment	0.041*	0.051**	0.024
	(0.02)	(0.02)	(0.02)
Treatment x Female	0.040	0.025	0.048
	(0.03)	(0.04)	(0.03)
Constant	-0.126	-0.132	-0.109
Observations	8,002	7,906	7,944
***Panel B**: Heterogeneity by Nationality*			
Haitian	0.110***	0.167***	0.041
	(0.04)	(0.04)	(0.04)
Treatment	0.062***	0.070***	0.045**
	(0.02)	(0.02)	(0.02)
Treatment x Haitian	-0.031	-0.045	-0.018
	(0.04)	(0.05)	(0.04)
Constant	-0.137	-0.194	-0.069
Observations	7,962	7,866	7,908
***Panel C**: Heterogeneity by Camp Being Host at Student's Regular School*			
Treatment	0.054***	0.057***	0.043**
	(0.02)	(0.02)	(0.02)
Treatment x Zero Distance	0.008	0.020	-0.011
	(0.05)	(0.05)	(0.05)
Constant	-0.109	-0.153	-0.057
Observations	8,002	7,906	7,944
***Panel D**: Heterogeneity by Distance between Camp and Student's Regular School*			
Treatment	0.038	0.050*	0.012
	(0.03)	(0.03)	(0.03)
Treatment x Distance	0.012	0.007	0.020
	(0.01)	(0.01)	(0.01)
Constant	-0.109	-0.154	-0.057
Observations	8,002	7,906	7,944

Note: This table shows heterogeneous intent-to-treat effects of being invited to the remedial camp on endline test scores for the 2025 cohort. Each panel reports estimates from a separate OLS regression in which the treatment indicator is interacted with a student characteristic: nationality (Panel A), gender (Panel B), whether the camp is located in the student's school from the previous academic year (Panel C), and distance in kilometers between the student's school and the camp site (Panel D). Columns report results for aggregate test scores (Column 1), mathematics (Column 2), and Spanish (Column 3). All specifications include stratum fixed effects and baseline controls pre-specified in our pre-analysis plan, including baseline test scores interacted with grade. The main effects of distance and camp at student's school are absorbed by the stratum fixed effects as these variables do not vary within strata. In 2025, the reduction in maximum distance to camp sites to less than 5 kilometers substantially reduced variation in the distance variables. The sample includes students selected by the Ministry of Education to participate in the 2025 remedial camp for whom endline test scores are available. Missing baseline variables are imputed at the mean, with indicator variables for missingness included. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.4. Heterogeneous Effects on Learning of Being Invited to the Remedial Camp by Students' Overage Status (2025)

	(1) Aggregate	(2) Math	(3) Spanish
Treatment	0.076***	0.074***	0.065**
	(0.03)	(0.03)	(0.03)
Treatment × 2 years overage	-0.005	-0.011	0.004
	(0.04)	(0.04)	(0.04)
Treatment × 3+ years overage	-0.067	-0.039	-0.092**
	(0.04)	(0.05)	(0.04)
Joint F-test	1.51	0.40	2.99
	0.223	0.670	0.052
Observations	7,999	7,903	7,941

Note: This table shows heterogeneous intent-to-treat effects of being invited to the remedial camp on endline test scores by overage status for the 2025 cohort. Estimates are from an OLS regression in which the treatment indicator is interacted with indicators for overage categories, with one year of overage as the omitted category. Columns report results for aggregate test scores (Column 1), mathematics (Column 2), and Spanish (Column 3). The specification includes stratum fixed effects and baseline controls pre-specified in our pre-analysis plan, including baseline test scores interacted with grade. The sample is restricted to overage students, as no on-age students were included in the 2025 cohort. The sample includes students selected by the Ministry of Education to participate in the 2025 remedial camp for whom endline test scores are available. Missing baseline variables are imputed at the mean, with indicator variables for missingness included. Standard errors are clustered at the school level. * p < 0.10, ** p < 0.05, *** p < 0.01.

TABLE A.5. Heterogeneous Effects on Learning of Attending the Remedial Camp (2025)

	(1) Aggregate Ability	(2) Math	(3) Spanish
***Panel A**: Heterogeneity by Gender*			
Female	0.034	-0.054*	0.116***
	(0.03)	(0.03)	(0.03)
Share of days attended	0.069**	0.084**	0.040
	(0.03)	(0.04)	(0.04)
Share of days attended x Female	0.061	0.037	0.074
	(0.05)	(0.05)	(0.05)
Observations	7,999	7,903	7,941
***Panel B**: Heterogeneity by Nationality*			
Haitian	0.113***	0.170***	0.043
	(0.04)	(0.04)	(0.04)
Share of days attended	0.099***	0.110***	0.072**
	(0.03)	(0.03)	(0.03)
Share of days attended x Haitian	-0.043	-0.066	-0.023
	(0.07)	(0.07)	(0.07)
Observations	7,959	7,863	7,905
***Panel C**: Heterogeneity by Camp Being Host at Student's Regular School*			
Share of days attended	0.088***	0.092***	0.069**
	(0.03)	(0.03)	(0.03)
Share of days attended x Zero Distance	0.012	0.030	-0.019
	(0.08)	(0.07)	(0.07)
Observations	7,999	7,903	7,941
***Panel D**: Heterogeneity by Distance between Camp and Student's Regular School*			
Share of days attended	0.065	0.084*	0.023
	(0.04)	(0.04)	(0.04)
Share of days attended x Distance	0.017	0.009	0.029
	(0.02)	(0.02)	(0.02)
Observations	7,999	7,903	7,941

Note: This table presents average causal response estimates of attending the remedial camp on endline test scores for the 2025 cohort. Each panel reports estimates from a separate two-stage least squares regression in which the share of camp days attended (instrumented by the randomly assigned invitation) is interacted with a student characteristic: nationality (Panel A), gender (Panel B), whether the camp is located in the student's school from the previous academic year (Panel C), and distance in kilometers between the student's school and the camp site (Panel D). Columns report results for aggregate test scores (Column 1), mathematics (Column 2), and Spanish (Column 3). All specifications include stratum fixed effects and baseline controls pre-specified in our pre-analysis plan, including baseline test scores interacted with grade. The main effects of distance and camp at student's school are absorbed by the stratum fixed effects as these variables do not vary within strata. In 2025, the reduction in maximum distance to camp sites to less than 5 kilometers substantially reduced variation in the distance variables. The sample includes students selected by the Ministry of Education to participate in the 2025 remedial camp for whom endline test scores are available. Missing baseline variables are imputed at the mean, with indicator variables for missingness included. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.6. Effects on Learning of Attending the Remedial Camp by Overage Status (2025)

	(1) Aggregate	(2) Math	(3) Spanish
Share of days attended	0.122***	0.121***	0.104**
	(0.04)	(0.04)	(0.05)
Share of days attended × 2 years overage	-0.013	-0.022	0.002
	(0.06)	(0.07)	(0.07)
Share of days attended × 3+ years overage	-0.106	-0.060	-0.148**
	(0.07)	(0.07)	(0.07)
Joint F-test (heterogeneity)	1.33	0.34	2.77
p-value	0.267	0.712	0.064
Observations	7,999	7,903	7,941

Note: This table shows heterogeneous average causal response estimates of attending the remedial camp on endline test scores by overage status for the 2025 cohort. Estimates are from a two-stage least squares regression in which the share of camp days attended (instrumented by the randomly assigned invitation) is interacted with indicators for overage categories, with one year of overage as the omitted category. Columns report results for aggregate test scores (Column 1), mathematics (Column 2), and Spanish (Column 3). The specification includes stratum fixed effects and baseline controls pre-specified in our pre-analysis plan, including baseline test scores interacted with grade. The sample is restricted to overage students, as no on-age students were included in the 2025 cohort. The sample includes students selected by the Ministry of Education to participate in the 2025 remedial camp for whom endline test scores are available. Missing baseline variables are imputed at the mean, with indicator variables for missingness included. Standard errors are clustered at the school level. * p < 0.10, ** p < 0.05, *** p < 0.01.

TABLE A.7. Heterogeneous Effects on Learning of Attending the Remedial Camp by Grade

	(1)	(2)	(3)
	Aggregate	Math	Spanish
***Panel A**: Pooled years*			
Share of days attended	0.052	0.006	0.050
	(0.05)	(0.05)	(0.05)
Share of days attended × Grade 4	0.077	0.060	0.109
	(0.08)	(0.08)	(0.08)
Share of days attended × Grade 5	0.093	0.156**	0.050
	(0.08)	(0.08)	(0.08)
Joint F-test	0.77	1.29	0.54
	0.544	0.273	0.706
Observations	10,632	10,185	10,309
***Panel B**: 2024*			
Share of days attended	0.157	-0.035	0.165
	(0.12)	(0.14)	(0.15)
Share of days attended × Grade 4	-0.007	-0.036	0.142
	(0.24)	(0.27)	(0.26)
Share of days attended × Grade 5	0.037	0.200	0.071
	(0.22)	(0.24)	(0.23)
Joint F-test	0.02	0.44	0.16
	0.981	0.645	0.855
Observations	2,633	2,282	2,368
***Panel C**: 2025*			
Share of days attended	0.008	0.005	0.015
	(0.05)	(0.05)	(0.06)
Share of days attended × Grade 4	0.113*	0.104	0.101
	(0.07)	(0.07)	(0.07)
Share of days attended × Grade 5	0.118	0.157**	0.042
	(0.07)	(0.07)	(0.07)
Joint F-test	1.69	2.39	0.98
	0.186	0.094	0.376
Observations	7,999	7,903	7,941

Note: This table shows heterogeneous average causal response estimates of attending the remedial camp on endline test scores by grade. Estimates are from a two-stage least squares regression in which the share of camp days attended (instrumented by the randomly assigned invitation) is interacted with grade indicators, with grade 3 as the omitted category. Columns report results for aggregate test scores (Column 1), mathematics (Column 2), and Spanish (Column 3). Panel A reports pooled results across both cohorts, Panel B reports results for the 2024 cohort, and Panel C reports results for the 2025 cohort. All specifications include stratum fixed effects and baseline controls pre-specified in our pre-analysis plan, including baseline test scores interacted with grade. The pooled specification in Panel A additionally includes cohort fixed effects. Grade is not included as a separate control as it enters the specification through the interaction terms. The sample includes all students selected by the Ministry of Education to participate in the study for whom endline test scores are available. Missing baseline variables are imputed at the mean, with indicator variables for missingness included. Standard errors are clustered at the school level. * p < 0.10, ** p < 0.05, *** p < 0.01.

To document implementation processes during both years of the summer camp, we conducted focus groups with camp staff in 2024 and 2025. In 2024, focus groups were conducted with teachers only. In 2025, we conducted separate focus groups with teachers, camp coordinators, and district-level technicians.

A.1. 2024 Focus Groups

Focus groups were conducted with 28 teachers. The discussion guide covered preparation and logistics, the use of instructional materials, classroom organization, and student participation.

Teachers described receiving the teacher guides shortly before the start of the camp, in some cases with content for the initial days still incomplete. This limited their ability to plan ahead and required additional work outside scheduled hours to organize and adapt the material. Several also noted the limited availability of physical classroom materials such as manipulatives and art supplies, which constrained the variety of activities that could be implemented.

When asked about the alignment between the instructional guides and students' observed learning levels, teachers reported classrooms in which students' reading abilities were below the level assumed by the daily activities. In response, they described modifying activities, reducing the number of exercises completed per day, omitting components, or reorganizing tasks. Several reported that the volume of planned activities exceeded what could be covered within the camp schedule given students' pace.

Teachers also described variation in attendance patterns across the three weeks, including students joining after the first week and others discontinuing attendance. Transportation barriers were frequently mentioned, particularly for students living farther from camp locations, and were described as affecting consistent attendance at some sites. Teachers reported that late arrivals disrupted classroom dynamics, requiring repetition of prior material or individual catch-up instruction.

Finally, teachers described classroom management challenges, particularly during the initial weeks and in classrooms with mixed age groups. Engagement varied across students, and teachers reported making instructional adjustments for students who were not yet literate, including providing individualized support within the classroom.

A.2. 2025 Focus Groups

Focus groups were conducted with 34 teachers, 39 camp coordinators, and 37 district-level technicians. Separate discussion guides were developed for each group to reflect differences in responsibility across classroom instruction, camp-level organization, and district oversight.

Teachers and coordinators highlighted the formal organization of students into three performance levels as a central change relative to 2024. Classrooms were structured accordingly, and both instructional guides and student workbooks were aligned to these levels. They described this structure as reducing within-class heterogeneity compared to the prior year, and the availability of level-specific materials as facilitating daily instructional

planning and implementation.

Logistical arrangements were also described as improved. In several locations, counseling and psychology support staff were assigned to assist with implementation, providing additional help with student discipline and well-being. Teachers noted that transportation was more consistently available and that meal and snack provision had improved relative to the prior year. These changes were described as supporting attendance and student participation.

Teachers reported that monitoring visits were frequent at some sites, noting that repeated classroom observations could at times interrupt instructional flow.

Classroom management remained a recurring theme. Teachers described discipline challenges and lower initial engagement during the first week, particularly among older students, but reported that engagement improved as socio-emotional and recreational activities were introduced. Several teachers described students who were often characterized negatively during the regular school year as responding positively to the camp's structured and affirming environment. Participants also noted limited strategies for supporting students with different or special learning needs within the camp structure.

Camp coordinators discussed operational aspects of implementation that were less visible at the classroom level. In particular, they described the process of organizing students into classrooms and coordinating teachers across levels during the first days of the camp. Because the distribution of students across learning levels was finalized only after the camp had begun, the initial organization of classrooms required adjustments during the first days. Several coordinators noted that this process was particularly challenging in camps without counseling or psychology support, as they were often expected to assist teachers in managing behavioral or learning-related issues.

Coordinators also described a substantial operational workload, including coordinating materials, conducting pedagogical classroom visits and providing feedback to teachers, communicating with families, and addressing day-to-day logistical issues. Those who had also participated in the 2024 camp reported that the workload was greater than expected under the single-coordinator model and suggested that having two coordinators — one focused on logistics and another on pedagogical support, as in the previous year — would facilitate operations. This concern was particularly salient for coordinators responsible for more than one camp site, as occurred in three districts where a single coordinator oversaw two camps.

At the same time, coordinators highlighted several positive aspects of the program, including improved transportation arrangements, the provision of snacks for students, and the structure of the instructional materials by level. Some also noted that end-of-day debrief sessions with teachers provided a useful space to discuss challenges and coordinate adjustments for subsequent sessions.

District-level technicians emphasized delays in the delivery of student level-assignment lists and instructional materials in some locations, which complicated the initial planning and organization of classrooms. Technicians also discussed their role in monitoring implementation and providing feedback to camp coordinators and teachers during classroom visits. While monitoring was generally viewed as important for maintaining implementation quality, its frequency at some sites was noted as potentially disruptive to instructional flow — a concern echoed by teachers and coordinators. At the same time, technicians highlighted that regular monitoring helped identify implementation challenges early and facilitated communication between schools, districts, and the central team, allowing logistical issues to be addressed more quickly during the camp.

Appendix B. Summer Camp Material

This appendix describes the instructional materials used in the 2024 and 2025 summer camps, including teacher guides, student notebooks, and supplementary classroom materials.

B.1. Teacher Guides

2024. A single teacher guide was used across all classrooms, covering both Spanish and mathematics. The guide provided daily instructional sequences organized by week, with each week structured around a theme, a specific text type in Spanish, and a defined mathematics focus.

In the first week (cooperative games and technology), Spanish instruction focused on instructive texts — understanding structure, identifying purpose, and following written instructions — while mathematics covered natural number sequences, place value, and number recognition and ordering. In the second week (arts and their different expressions), Spanish shifted to narrative texts, including reading comprehension, narrative structure, and basic written production, and mathematics addressed addition, subtraction, and problem-solving using basic operations. In the third week (environmental care), Spanish covered informational and news texts through reading and discussion, and mathematics introduced multiplication and division.

The daily schedule also included a one-hour ludic activities block at the end of the day, devoted to recreational and complementary activities for the whole group. Suggested activities included cooperative games, group dynamics, artistic and creative activities such as drawing, dramatization, or music, and outdoor or movement-based games. Teachers could also use this time to work with smaller groups of students, allowing for differentiated support or additional practice.

2025. The teacher guide was redesigned in two main ways: it was extended to cover a four-week camp, and it incorporated explicit differentiation by performance level. Students were organized into three groups — Exploradores, Creadores, and Innovadores — and the guide included a table describing expected competencies by level in both subjects. It also integrated socio-emotional components, a bank of recreational activities, structured daily time allocation, and a daily 15-minute reading block.

The weekly progression followed a similar thematic structure. The first week (socio-emotional education and technology) focused on identity documents and personal identity in Spanish, and numeration, recognition, writing, and ordering of numbers in mathematics. The second week (cooperative games) covered shopping lists and functional texts in Spanish, and place value and addition in mathematics. The third week (arts and imagination) addressed short stories and narrative comprehension and production in Spanish, and addition and subtraction in mathematics. The fourth week (environmental care) introduced children's songs and lyrical texts in Spanish, and multiplication and division in mathematics.

Content differed across levels in terms of complexity. In mathematics, differentiation operated primarily through number range and operational complexity. For example, in the first week, Exploradores worked with numbers 1–20 on ordering, identifying previous and next numbers, and writing numbers; Creadores covered the same skills but extended the range to 1–999; and Innovadores worked within the same number range but focused on solving everyday problems using mathematical operations. In Spanish, content varied in reading and writing complexity across levels.

The 2025 guide also included a daily ludic activities block devoted to structured recreational activities designed to promote participation, collaboration, and well-being. Activities included cooperative games, movement-based activities, music and rhythm exercises, and short group dynamics aimed at strengthening social interaction and classroom climate. In some cases, these activities were also used to introduce socio-emotional themes, such as recognizing emotions, building self-esteem, and fostering respectful relationships among students.

B.2. Student Notebooks and Other Materials

In 2024, no separate student notebooks were provided. All instruction was delivered through the teacher guide, with activities implemented based on the daily instructional sequences described therein. In 2025, three separate student notebooks were developed, one for each performance level: Exploradores, Creadores, and Innovadores. Each notebook was organized by week and included explicit weekly learning goals (*Metas de aprendizaje*), separate sections for Spanish and mathematics, a student identification page, socio-emotional elements, and weekly self-assessment sections.

In both years, the activities described in the guides relied on simple classroom materials to implement instructional and recreational activities. These included materials for literacy activities (e.g., cards, short printed texts, or posters to support reading and writing exercises), manipulatives for mathematics (e.g., objects or visual representations for number recognition, ordering, and basic operations), and basic materials for the artistic and movement-based activities in the ludic workshops.

Appendix C. Tests Comparison (2024-2025)

The Summer Camp evaluation administered four rounds of student assessments. Within each cohort, the Baseline and Endline follow a *mirror design*, same structure and skill domains. The tables below compare instruments *across cohorts*: Baseline 2024 against Baseline 2025 (left pair), and Endline 2024 against Endline 2025 (right pair).

TABLE C.1. Spanish Items — Change Summary (2024 vs. 2025)

Question Type	Task	Change type	What changed
L1	Listening: identify dictated word among 4 options (3 real + 2 nonsense words)	*New in 2025*	Item added in 2025; not present in 2024 (BL or EL)
L2	Word decoding: complete word with correct missing letter (image-supported, 4 items)	*Minor*	One target word replaced: *montaña* → *paloma* (BL); *elefante* → *mesa* (EL)
L3	Syllable assembly: rearrange syllables into the correct word (4 items)	*None*	Identical across years (BL and EL)
L4	Reading comprehension: read passage and answer 3 MC questions	*Modified*	BL: same story; Q(c) replaced with inferential question. EL: new story (*Tomás el gato*); Q(b) and Q(c) changed
L5	Open-ended justification: write justification for previous answer	*Dropped in 2025*	Item removed in 2025 (BL and EL)
L6	Sentence reconstruction: rewrite the scrambled sentence correctly	*Dropped in 2025*	Item removed in 2025 (BL and EL)
L7	Fill in the blanks: complete 3 sentences using words from a box	*Minor*	BL word box identical. EL word box changed: *amigos/humilde/gigante* → *dormido/curioso/ratón*
L8	Spelling (dictation): write four words dictated by the teacher	*Dropped in 2025*	Item removed in 2025 (BL and EL)

*Note: B*L = Baseline; EL = Endline. *None* = identical across years; *Minor* = single word or numeral changed; *Modified* = structure or content substantively changed.

TABLE C.2. Mathematics Items — Change Summary (2024 vs. 2025): Part I

Question Type	Task	Change type	What changed
M1	Arithmetic operations: vertical addition and subtraction	*Modified*	Expanded from 4 to 6 operations; two simpler 2-digit problems added
M2	Number recognition: identify correct written form of a numeral	*Minor*	BL numeral identical (708). EL numeral changed: 678 → 748
M3	Number conversion: words to digits	*Minor*	BL identical (1,844). EL number and context changed: 1,949 → 1,963
M4	Comparing numbers	*Modified*	Format changed: open-ended with justification (2024) → MC only (2025)
M5	Number ordering on a number line: order values from least to greatest	*Minor*	BL identical. EL price range adjusted: RD$ 130–170 → RD$ 120–160
M6	Addition word problem	*None*	Identical across years (BL and EL)
M7	Subtraction word problem	*Minor*	Numbers changed (BL: 55/17 → 63/25; EL: 65/13 → 74/22); structure unchanged
M8	Division word problem	*Dropped in 2025*	Item removed in 2025 (BL and EL)

*Note: B*L = Baseline; EL = Endline. *None* = identical across years; *Minor* = single word or numeral changed; *Modified* = structure or content substantively changed.

TABLE C.3. Mathematics Items — Change Summary (2024 vs. 2025): Part II

Question Type	Task	Change type	What changed
M9	Multiplication word problem — single-digit multiplier	*None*	Identical across years (BL and EL)
M10	Multiplication word problem — two-digit multiplier	*Dropped in 2025*	Item removed in 2025 (BL and EL)
M11	Place value decomposition — identify hundreds, tens, and units digits (open-ended)	*New in 2025*	Item added in 2025. Number differs by round: 679 (BL) vs. 842 (EL)
M12	Time calculation — add one hour to a given time	*New in 2025*	Item added in 2025. Start time differs by round: 7:00 (BL) vs. 8:00 (EL)
M13	Bar chart interpretation — identify two categories with equal values	*New in 2025*	Item added in 2025. Context differs by round: erasers sold (BL) vs. bottles sold (EL)

Note: None = identical across years; *Minor* = single word or numeral changed; *Modified* = structure or content substantively changed.

The tables below compare instruments *across cohorts*: Baseline 2024 against Baseline 2025 (left pair), and Endline 2024 against Endline 2025 (right pair).

TABLE C.4. Spanish Items by Round (Part 1)

	Baseline		Endline	
	2024	**2025**	**2024**	**2025**
L1	*Listening — Identify the dictated word among 4 options (3 real words + 2 nonsense words)*			
L1	—	**3 real: *moto, paloma, cocodrilo*. Nonsense: *lopame, trusal***	—	**3 real: *luna, camisa, elefante*. Nonsense: *misola, frusal***
L2	*Word decoding — Complete the word with the correct letter (image-supported, 4 items)*			
L2	*casa, tigre, escuela, montaña*	*casa, tigre, escuela,* ***paloma***	*zapato, canasta, pluma, elefante*	*zapato, canasta, pluma,* ***mesa***
L3	*Syllable Assembly — Rearrange syllables into the correct word (4 items)*			
L3	*bri-lla-som* (sombrilla), *tra-es-ma* (maestra), *co-bar* (barco), *po-sa-ri-ma* (mariposa)	Same as BL 2024	*na-za-ria-ho* (zanahoria), *za-pi-rra* (pizarra), *bro-li* (libro), *chi-la-mo* (mochila)	Same as EL 2024
L4	*Reading Comprehension — Read the passage and answer 3 MC questions*			
L4	Story: *El gato y los ratones*. MC: (a) cat's name; (b) who was scared; (c) did they resolve the problem?	Same story; same (a) and (b). **(c) replaced: inferential question on why mice applauded.**	Story: *Nito and Leo*. MC: (a) lion's name; (b) who was scared; (c) did they resolve the problem?	**Story: *Tomás el gato*.** MC: (a) cat's name; **(b) what mice were doing; (c) what mouse proposed to the cat.**

Note: BL = Baseline; EL = Endline. Bold text indicates content that changed relative to 2024. '—' = item not included in that round.

TABLE C.5. Spanish Items by Round (Part 2)

	Baseline		Endline	
	2024	**2025**	**2024**	**2025**
L5	*Open-ended Justification — Write justification for the previous answer*			
L5	Write justification for Q3	—	Write justification for Q3	—
L6	*Sentence Reconstruction — Rewrite the scrambled sentence correctly*			
L6	"*Los ratones y ratas se reunieron para buscar una solución*"	—	"*valiente era un Nito muy ratoncito*"	—
L7	*Fill in the Blanks — Complete sentences using words from the box (3 items)*			
L7	*cascabel / astuto / madrigueras*	Same as BL 2024	***amigos / humilde / gigante***	***dormido / curioso / ratón***
L8	*Spelling (dictation) — Write four words dictated by the teacher*			
L8	Four dictated words; party supply list context	—	Four dictated words; party supply list context	—

Note: BL = Baseline; EL = Endline. Bold text indicates content that changed relative to 2024. '—' = item not included in that round.

TABLE C.6. Mathematics Items by Round (Part 1)

	Baseline		Endline	
	2024	**2025**	**2024**	**2025**
M1	*Arithmetic Operations — Vertical addition and subtraction*			
M1	4 ops: 234+45, 347+367, 465−54, 546−537	**6 ops:** 34+12**,** 234+45**,** 347+367**,** 79−56**,** 465−54**,** 546−537	4 ops: 345+56, 458+478, 576−65, 657−648	**6 ops:** 42+14**,** 345+56**,** 458+478**,** 89−45**,** 576−65**,** 657−648
M2	*Number recognition — Identify the correct written form of a number (MC, 4 options)*			
M2	Written form of 708	Same as BL 2024	Written form of 678	**Written form of 748**
M3	*Number conversion (words to digits) — Identify the correct numeral (MC, 4 options)*			
M3	*Mil ochocientos cuarenta y cuatro* → 1,844 (Independence year)	Same as BL 2024	*Mil novecientos cuarenta y nueve* → 1,949 (grandmother's birth year)	***Mil novecientos sesenta y tres*** → **1,963 (school inauguration)**
M4	*Comparing Numbers*			
M4	Compare 8493 vs. 9834; justify; construct smallest 4-digit number from {8,3,4,9}	**MC only: greatest among {2735, 2537, 2573, 2753}**	Compare 7561 vs. 7651; justify; construct smallest 4-digit number from {5,7,1,6}	**MC only: greatest among {3648, 3468, 3684, 3486}**
M5	*Number Line — Order prices from least to greatest (MC, 4 options)*			
M5	Five prices, RD$ 140–160	Same as BL 2024	Five prices, RD$ 130–170	**Five prices, RD$ 120–160**
M6	*Addition Word Problem (MC)*			
M6	Wilson has RD$575; dad gives RD$360. Total?	Same as BL 2024	Juan has RD$352; dad gives RD$465. Total?	Same as EL 2024

Note: BL = Baseline; EL = Endline. Bold text indicates content that changed relative to 2024. '—' = item not included in that round.

TABLE C.7. Mathematics Items by Round (Part 2)

	Baseline		Endline	
	2024	**2025**	**2024**	**2025**
M7 Subtraction Word Problem (MC)				
M7	Manuel has 55 lives; gives sister 17. Remaining?	**Manuel has 63 lives; gives sister 25. Remaining?**	Juan has 65 bottles; gives cousin 13. Remaining?	**Juan has 74 bottles; gives cousin 22. Remaining?**
M8 Division Word Problem (MC)				
M8	Masks factory: 1,740 units in boxes of 6. How many boxes?	—	Ice cream factory: 1,456 units in boxes of 4. How many boxes?	—
M9–M10 Multiplication Word Problems (MC)				
M9	Restaurant: 54 tables × 6 plates. Total?	Same as BL 2024	Birthday: 42 tables × 5 plates. Total?	Same as EL 2024
M10	Juan: 35 candy bags × 24 candies. Total?	—	Juan: 45 candy bags × 6 candies. Total?	—
M11 Place value decomposition — Identify digits by hundreds, tens, and units (open-ended)				
M11	—	**Hundreds, tens, units digits in 679**	—	**Hundreds, tens, units digits in 842**
M12 Time calculation — Add one hour to a given time (MC, 4 options)				
M12	—	**School starts at 7:00; arrives 1 hour late**	—	**School starts at 8:00; arrives 1 hour late**
M13 Bar chart interpretation — Identify equal values across categories (MC, 4 options)				
M13	—	**Bar chart: two days with same number of erasers sold**	—	**Bar chart: two days with same number of bottles sold**

Note: BL = Baseline; EL = Endline. Bold text indicates content that changed relative to 2024. '—' = item not included in that round.

Appendix D. Socio-Emotional Survey (2025)

The 2025 evaluation included a socio-emotional survey administered at both Baseline and Endline. This instrument was not part of the 2024 evaluation. The Baseline survey measured empathy and emotional awareness, while the Endline survey covered a broader set of constructs including emotion regulation, school belonging, academic self-concept, and growth mindset. Tables D.1 and D.2 list all items for each round.

D.1. Index Construction

For each Endline item, responses are first reverse-coded where applicable (items 1 and 12) so that higher values consistently reflect higher levels of the measured construct. Each item is then standardized to have mean zero and standard deviation one. Domain indices are constructed as the mean of the standardized items within each domain, following the groupings below:

Index	Items	Reverse-coded items
Emotional Awareness & Empathy	EL 1, 2, 4, 5, 15	Item 1
Emotion Regulation & Self-Management	EL 3, 6	—
School Belonging	EL 7	—
Academic Self-Concept & Effort	EL 8, 9	—
Classroom Norms & Discipline	EL 10	—
Growth Mindset	EL 11, 12	Item 12
Bullying Victimization	EL 13, 14	—

TABLE D.1. Socio-Emotional Survey Items — Baseline 2025

#	Question	Response type	Reverse-coded
1	Children who cry because they are happy are dumb.	Binary (Yes/No)	Yes
2	I really like watching people when they open gifts, even when I don't receive any.	Binary (Yes/No)	No
3	Seeing a child crying makes me want to cry.	Binary (Yes/No)	No
4	Even when I don't know why someone is laughing, I laugh too.	Binary (Yes/No)	No
5	Sometimes I cry when I watch television.	Binary (Yes/No)	No
6	Girls who cry because they are happy are dumb.	Binary (Yes/No)	Yes
7	I find it difficult to understand why another person gets angry.	Binary (Yes/No)	Yes
8	I feel sad when I see a child who can't find anyone to play with.	Binary (Yes/No)	No
9	Some songs make me so sad that they make me want to cry.	Binary (Yes/No)	No
10	Adults sometimes cry even when they have no reason to feel bad.	Binary (Yes/No)	No
11	It is silly to treat dogs and cats as if they had feelings just like people.	Binary (Yes/No)	Yes
12	I get angry when I see a classmate who pretends to need the teacher's help all the time.	Binary (Yes/No)	Yes
13	Children who don't have friends probably don't want to have them.	Binary (Yes/No)	Yes
14	Seeing a girl crying makes me want to cry.	Binary (Yes/No)	No
15	I think it is ridiculous that some people cry during a sad movie or while reading a sad book.	Binary (Yes/No)	Yes
16	I am able to eat all my candy even when I see someone looking at me and wanting one.	Binary (Yes/No)	Yes
17	I do not get upset when I see a classmate punished for not obeying school rules.	Binary (Yes/No)	Yes

TABLE D.2. Socio-Emotional Survey Items — Endline 2025

Construct / Domain	**#**	**Question**	**Response type**	**Reverse-coded**
Emotional Awareness & Empathy	1	Children who cry because they are happy are silly.	Binary (Yes/No)	Yes
	2	Seeing a child cry makes me feel like crying.	Binary (Yes/No)	No
	4	Sometimes, when I get angry, it is actually because I feel sad or hurt.	Binary (Yes/No)	No
	5	Sometimes people get angry because they feel left out.	Binary (Yes/No)	No
	15	When someone gets angry with me, I try to imagine what they are thinking or feeling.	Never / Sometimes / Several times / Always	No
Emotion Regulation & Self-Management	3	I find it easy to calm myself down when I feel upset.	Binary (Yes/No)	No
	6	When I get angry with someone, I try to calm myself down even if that person does not apologize.	Binary (Yes/No)	No
School Belonging & Institutional Support	7	My school cares about my well-being and about me doing well.	Binary (Yes/No)	No
Academic Self-Concept & Effort	8	I am the kind of student who makes an effort and works hard.	Binary (Yes/No)	No
	9	When I learn something new, I feel proud of myself.	Binary (Yes/No)	No
Classroom Norms & Discipline	10	It is important to stay quiet while the teacher is speaking.	Binary (Yes/No)	No
Growth Mindset	11	If I study a lot, I can improve in reading and math.	Binary (Yes/No)	No
	12	Some children are simply good at school and others are not.	Binary (Yes/No)	Yes
Bullying Exposure	13	During the last month: classmates laughed at me or insulted me.	Never / Sometimes / Many times / Every or almost every day	No
	14	During the last month: they left me out of games or activities.	Never / Sometimes / Many times / Every or almost every day	No

Appendix E. IRT and Cronbach's alpha Analysis

We use Item Response Theory (IRT) to construct latent ability measures for each subject (Mathematics and Spanish) and cohort (2024 and 2025). IRT has several advantages over simpler scoring approaches (see, Muralidharan, Singh, and Ganimian (2019)). Unlike percentage-correct or normalized raw scores, IRT allows items to contribute differently to the latent score depending on their estimated discrimination and difficulty, placing greater weight on items that are more informative for distinguishing students along the ability distribution. It also summarizes performance on the retained item set using a common latent scale within each subject and cohort. Finally, it provides item-level diagnostics (discrimination, difficulty, and ICC fit) that allow us to identify and remove poorly performing items before constructing the final scores.

E.1. 2PL Model

Under the two-parameter logistic (2PL) specification, the probability that student i answers item j correctly is:

$$P(Y_{ij} = 1 \mid \theta_i) = \frac{1}{1 + \exp\left[-a_j\left(\theta_i - b_j\right)\right]} \tag{A1}$$

where θ_i is the latent ability of student i, a_j is the *discrimination* parameter, and b_j is the *difficulty* parameter of item j. The difficulty parameter gives the ability level at which the probability of a correct response equals 50%; items with higher b_j require greater ability to answer correctly. The discrimination parameter a_j governs how steeply the probability of success increases with ability around b_j; higher values indicate that the item better differentiates between students of different ability levels. The logistic functional form in equation (A1) implies a monotonic S-shaped Item Characteristic Curve (ICC), mapping latent ability to the probability of a correct response.

For each subject and cohort, we estimate a two-parameter logistic (2PL) model using the retained item set and predict student-specific latent ability scores $\hat{\theta}_i$ from the fitted model. The resulting scores are standardized to mean zero and standard deviation one within cohort. Students who did not attend the endline assessment receive a missing score rather than an imputed value.

E.2. Empirical Distribution of IRT Scores

Figures E.1 and E.2 present the distribution of IRT scores in Mathematics and Spanish at baseline and endline for each cohort. The distributions are unimodal and well spread

across the ability range, with little mass at the floor or ceiling of the distribution. This indicates that the tests provide informative measures of student achievement across a wide range of ability levels in this setting.

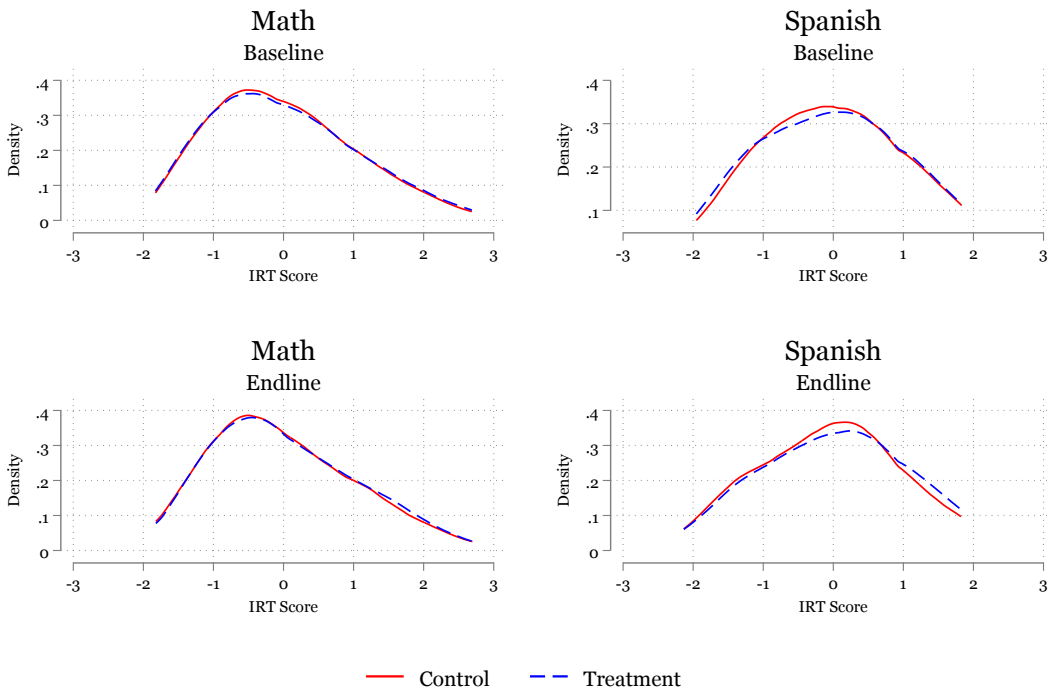


FIGURE E.1. Distribution of IRT Scores by Treatment Status, 2024 Cohort

Each panel shows the kernel density of IRT scores separately for treatment and comparison students. The top panels correspond to the baseline assessment (Mathematics and Spanish) and the bottom panels to the endline assessment. IRT scores are estimated using a two-parameter logistic model on the retained item sets and standardized to have mean zero and standard deviation one within cohort

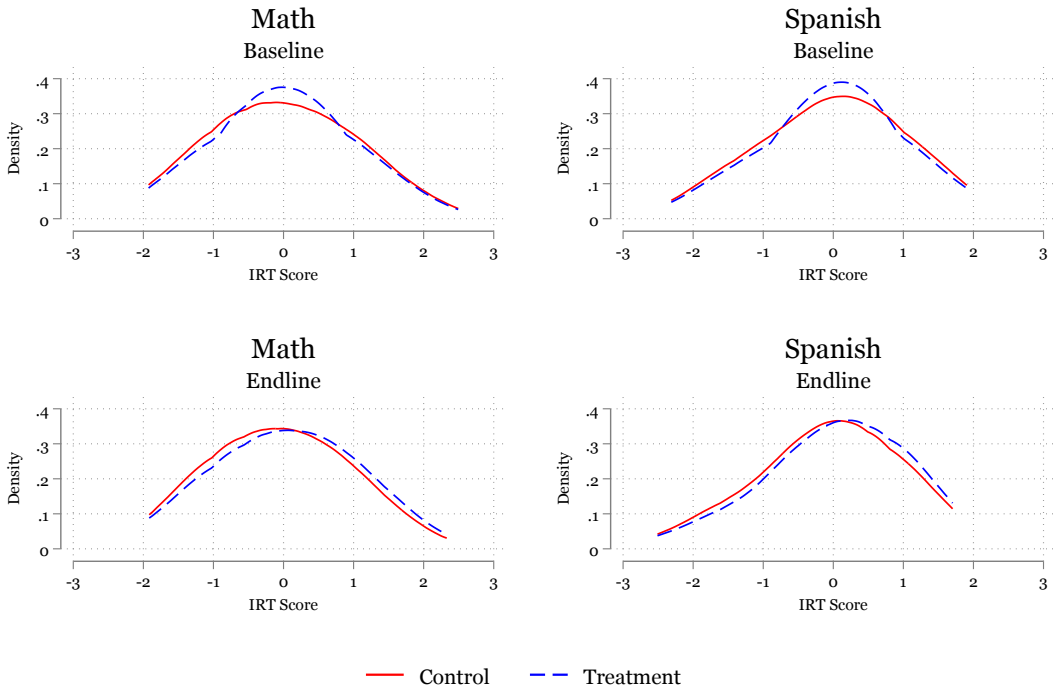


FIGURE E.2. Distribution of IRT Scores by Treatment Status, 2025 Cohort

Each panel shows the kernel density of IRT scores separately for treatment and comparison students. The top panels correspond to the baseline assessment (Mathematics and Spanish) and the bottom panels to the endline assessment. IRT scores are estimated using a two-parameter logistic model on the retained item sets and standardized to have mean zero and standard deviation one within cohort.

E.3. Item Fit and Differential Item Functioning

A standard assumption in IRT is that item parameters are invariant across subgroups of respondents, what the psychometrics literature calls the absence of *differential item functioning* (DIF). In a randomized experiment, this assumption is particularly consequential: if the same item behaved differently for treatment and comparison students at the same ability level, the latent scores would not be on a common scale across arms, and treatment effect estimates could be confounded by the measurement model rather than reflecting genuine learning differences.

We assess DIF following the approach of Muralidharan, Singh, and Ganimian (2019): for each item we plot the estimated ICC together with the empirical proportion correct computed within quintiles of the estimated ability distribution, calculated separately for treatment and comparison students. If the model fits well and DIF is absent, the empirical points for both groups should lie close to the same curve. Figures E.3–E.6 show these plots for all four test specifications. Across items and cohorts, the empirical proportions for

treatment and comparison students broadly track the same estimated curve, with no clear visual evidence of systematic differential item functioning across experimental arms. This suggests that the estimated treatment effects are unlikely to be driven by differential behavior of the measurement model between treatment and comparison students.

2024 Cohort

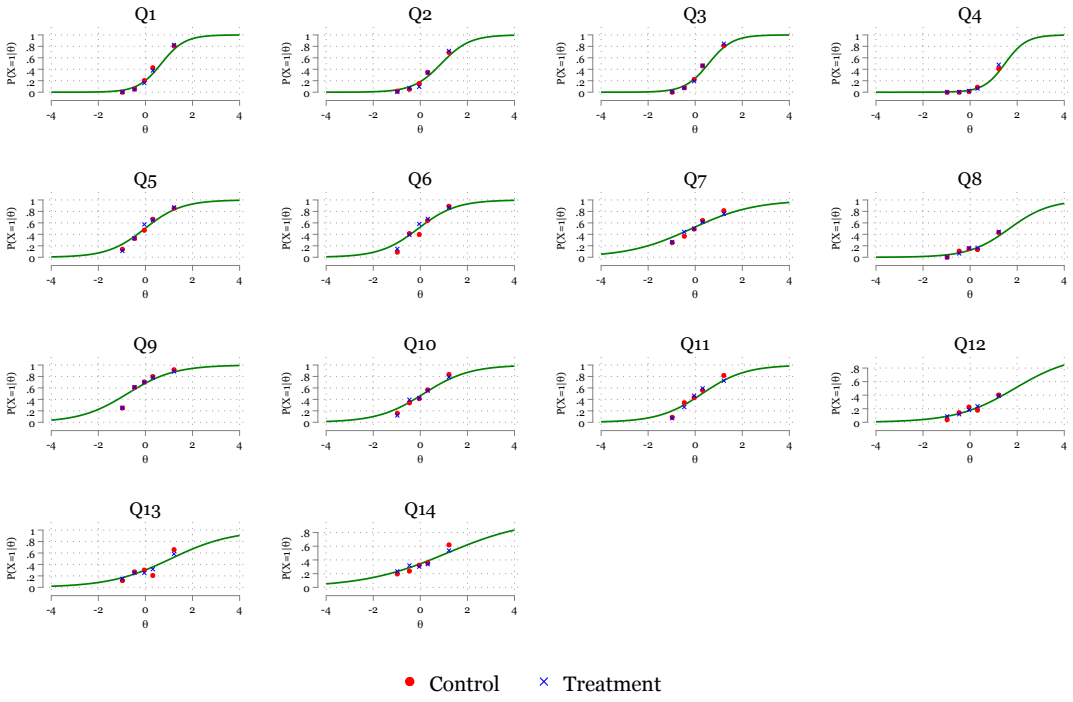


FIGURE E.3. Item Characteristic Curves, Math Endline 2024

Each subplot shows the estimated Item Characteristic Curve (ICC) for one item from the 2024 Math endline assessment together with the empirical proportion of correct responses computed within quintiles of the estimated ability distribution, shown separately for treatment (crosses) and comparison (circles) students.

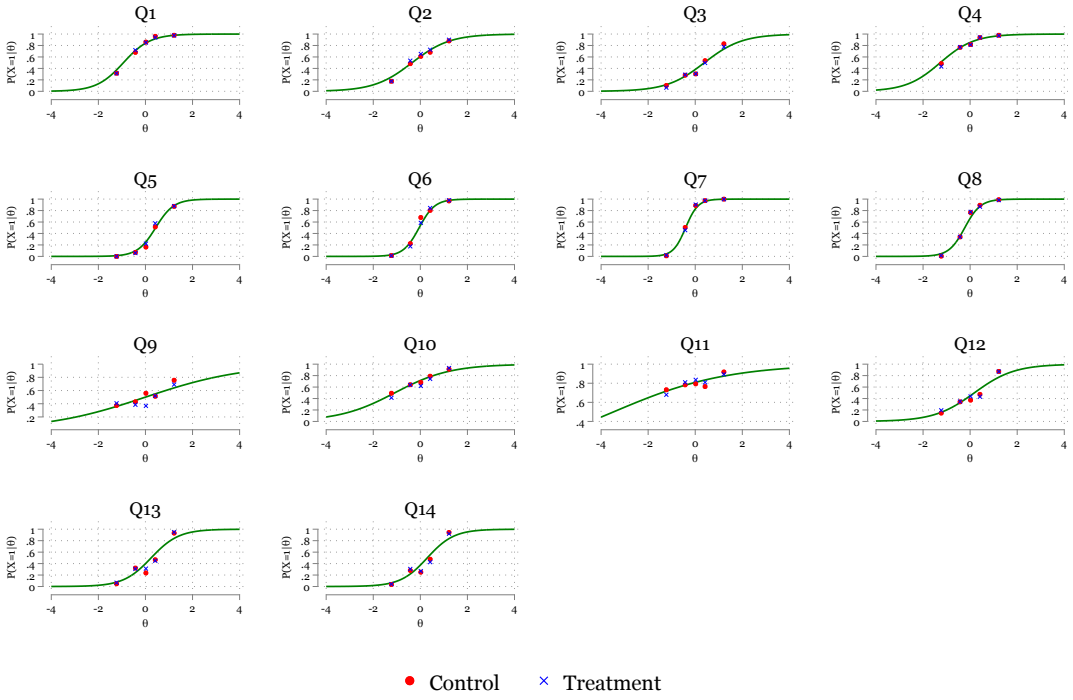


FIGURE E.4. Item Characteristic Curves, Spanish Endline 2024

Each subplot shows the estimated Item Characteristic Curve (ICC) for one item from the 2024 Spanish endline assessment together with the empirical proportion of correct responses computed within quintiles of the estimated ability distribution, shown separately for treatment (crosses) and comparison (circles) students.

2025 Cohort

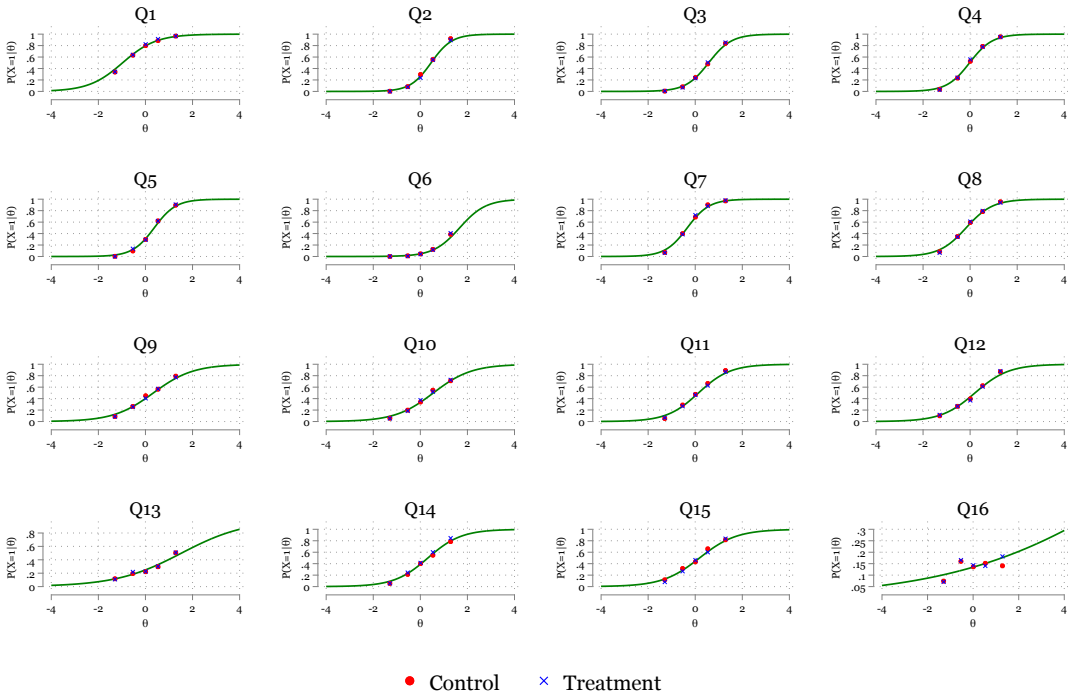


FIGURE E.5. Item Characteristic Curves, Math Endline 2025

Each subplot shows the estimated Item Characteristic Curve (ICC) for one item from the 2025 Mathematics endline assessment together with the empirical proportion of correct responses computed within quintiles of the estimated ability distribution, shown separately for treatment (crosses) and comparison (circles) students.

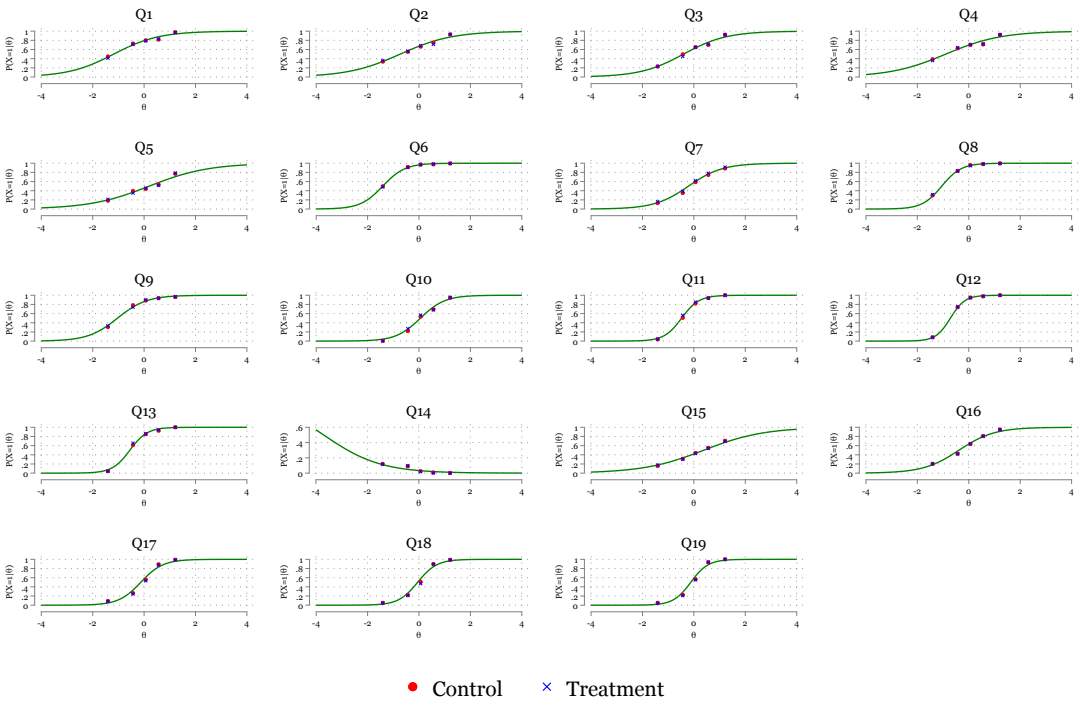


FIGURE E.6. Item Characteristic Curves, Spanish Endline 2025

Each subplot shows the estimated Item Characteristic Curve (ICC) for one item from the 2025 Spanish endline assessment together with the empirical proportion of correct responses computed within quintiles of the estimated ability distribution, shown separately for treatment (crosses) and comparison (circles) students.

E.4. Item Screening and Exclusions

For each test specification, we examined each item's estimated discrimination parameter, difficulty parameter, ICC shape, item-rest correlation, and contribution to overall scale reliability. Items exhibiting flat or irregular ICC patterns, weak discrimination, or near-zero item-rest correlations were flagged and reviewed. After this diagnostic review, one item was excluded from each test:

- **2024 Mathematics:** Item Q7 was excluded.
- **2024 Spanish:** Item Q11 was excluded.
- **2025 Mathematics:** Item Q16 was excluded.
- **2025 Spanish:** Item Q14 was excluded.

The final ability scores $\hat{\theta}_i$ used as outcome variables were predicted from the 2PL model estimated on the retained item sets. The excluded items are shown in the reliability tables below (marked with an asterisk) for transparency.

E.5. Scale Reliability

Tables E.1–E.2 report Cronbach's alpha and item-level diagnostics for each test specification. Cronbach's alpha is defined as:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{j=1}^{K} \sigma_j^2}{\sigma_T^2}\right) \tag{A2}$$

where K is the number of items, σ_j^2 is the variance of item j, and σ_T^2 is the variance of the total score. Values above 0.70 are generally considered acceptable for research purposes. For each item we report: (i) difficulty, measured as the proportion of students answering correctly; (ii) the item-test correlation, i.e. the correlation of the item with the full scale score; (iii) the item-rest correlation, i.e. the correlation of the item with the sum of all other items; and (iv) the alpha that would result from deleting that item. Items marked with * were excluded from the final IRT estimation.

2024 Cohort

TABLE E.1. Scale Reliability: Endline 2024

Panel A. Mathematics

Item	Difficulty	Item-Test Corr.	Item-Rest Corr.	α if Deleted
Q1	0.304	0.596	0.498	0.763
Q2	0.269	0.556	0.455	0.769
Q3	0.341	0.603	0.503	0.762
Q4	0.131	0.527	0.451	0.766
Q5	0.531	0.538	0.421	0.770
Q6	0.542	0.537	0.419	0.770
Q7*	0.548	0.437	0.309	0.781
Q9	0.204	0.475	0.378	0.774
Q10	0.678	0.448	0.327	0.778
Q11	0.487	0.495	0.369	0.774
Q12	0.452	0.525	0.404	0.772
Q13	0.211	0.397	0.286	0.782
Q14	0.343	0.454	0.331	0.778
Q15	0.364	0.388	0.257	0.783
Overall α			0.786	

Panel B. Spanish

Item	Difficulty	Item-Test Corr.	Item-Rest Corr.	α if Deleted
Q1	0.755	0.549	0.467	0.816
Q2	0.592	0.532	0.434	0.817
Q3	0.417	0.545	0.448	0.817
Q4	0.794	0.468	0.382	0.822
Q5	0.367	0.653	0.575	0.810
Q6	0.533	0.678	0.600	0.807
Q7	0.662	0.685	0.613	0.805
Q8	0.595	0.675	0.598	0.807
Q9	0.515	0.350	0.228	0.833
Q10	0.713	0.450	0.348	0.823
Q11*	0.811	0.247	0.145	0.837
Q14	0.496	0.530	0.426	0.819
Q15	0.479	0.648	0.562	0.811
Q16	0.471	0.651	0.566	0.811
Overall α			0.828	

Each panel reports item-level reliability diagnostics for the 2024 endline assessment. Difficulty is the proportion of students answering the item correctly. Item-test correlation is the correlation between the item score and the total scale score. Item-rest correlation is the correlation between the item score and the sum of all other items. Alpha if deleted reports the Cronbach's alpha of the scale after removing that item. Overall alpha is computed using the full set of items shown in the table. Items marked with * were excluded from the final IRT estimation due to poor diagnostic performance and are shown here for transparency.

2025 Cohort

TABLE E.2. Scale Reliability: Endline 2025

Panel A. Mathematics

Item	Difficulty	Item-Test Corr.	Item-Rest Corr.	α if Deleted
Q1	0.732	0.494	0.406	0.834
Q2	0.360	0.648	0.572	0.824
Q3	0.332	0.620	0.542	0.826
Q4	0.509	0.639	0.558	0.825
Q5	0.388	0.641	0.562	0.825
Q6	0.115	0.437	0.371	0.835
Q7	0.606	0.655	0.579	0.824
Q8	0.551	0.625	0.542	0.826
Q9	0.423	0.524	0.429	0.832
Q10	0.372	0.532	0.440	0.832
Q11	0.465	0.597	0.510	0.828
Q12	0.449	0.576	0.486	0.829
Q13	0.270	0.377	0.279	0.840
Q14	0.420	0.571	0.482	0.829
Q15	0.457	0.561	0.469	0.830
Q16*	0.137	0.161	0.078	0.851
Overall α			0.840	

Panel B. Spanish

Item	Difficulty	Item-Test Corr.	Item-Rest Corr.	α if Deleted
Q1	0.747	0.514	0.444	0.882
Q2	0.647	0.493	0.414	0.883
Q3	0.595	0.542	0.465	0.881
Q4	0.670	0.475	0.395	0.884
Q5	0.464	0.453	0.366	0.885
Q6	0.869	0.575	0.526	0.879
Q7	0.560	0.571	0.496	0.880
Q8	0.813	0.651	0.601	0.877
Q9	0.774	0.583	0.522	0.879
Q10	0.488	0.620	0.551	0.878
Q11	0.671	0.702	0.648	0.875
Q12	0.750	0.719	0.671	0.874
Q13	0.694	0.704	0.651	0.875
Q14*	0.048	-0.160	-0.203	0.890
Q15	0.431	0.421	0.333	0.886
Q16	0.603	0.572	0.499	0.880
Q17	0.549	0.646	0.580	0.878
Q18	0.526	0.667	0.604	0.877
Q19	0.552	0.693	0.634	0.876
Overall α			0.886	

Each panel reports item-level reliability diagnostics for the 2025 endline assessment. Difficulty is the proportion of students answering the item correctly. Item-test correlation is the correlation between the item score and the total scale score. Item-rest correlation is the correlation between the item score and the sum of all other items. Alpha if deleted reports the Cronbach's alpha of the scale after removing that item. Overall alpha is computed using the full set of items shown in the table. Items marked with * were excluded from the final IRT estimation due to poor diagnostic performance and are shown here for transparency.

Appendix F. Lee bound sensitivity analysis

TABLE F.1. Lee Bounds on the Effects on Learning of Being Invited to the Remedial Camp (Residualized Ranking)

	(1)	(2)	(3)
	Aggregate Ability	Math	Spanish
Panel A: Pooled years			
Invited	0.057***	0.050***	0.048***
	(0.015)	(0.016)	(0.015)
Lee Lower bound	0.046***	0.033**	0.045***
	(0.015)	(0.016)	(0.015)
Lee Upper bound	0.103***	0.103***	0.093***
	(0.014)	(0.016)	(0.015)
Imbens-Manski 95% CI	[0.022, 0.126]	[0.007, 0.129]	[0.020, 0.117]
Trim share (q)	0.021	0.027	0.019
Group trimmed	Invited	Invited	Invited
Observations	10,632	10,185	10,309
Panel B: 2024			
Invited	0.054*	0.005	0.072**
	(0.029)	(0.034)	(0.034)
Lee Lower bound	0.024	-0.033	0.072**
	(0.029)	(0.034)	(0.034)
Lee Upper bound	0.187***	0.182***	0.205***
	(0.029)	(0.032)	(0.034)
Imbens-Manski 95% CI	[-0.023, 0.235]	[-0.088, 0.234]	[0.016, 0.262]
Trim share (q)	0.074	0.097	0.067
Group trimmed	Invited	Invited	Invited
Observations	2,633	2,282	2,368
Panel C: 2025			
Invited	0.056***	0.061***	0.040**
	(0.016)	(0.018)	(0.017)
Lee Lower bound	0.055***	0.057***	0.039**
	(0.016)	(0.018)	(0.017)
Lee Upper bound	0.056***	0.065***	0.042**
	(0.016)	(0.018)	(0.017)
Imbens-Manski 95% CI	[0.023, 0.087]	[0.025, 0.096]	[0.007, 0.074]
Trim share (q)	0.000	0.002	0.001
Group trimmed	Comparison	Invited	Invited
Observations	7,999	7,903	7,941

Note: This table reports Lee (2009) sharp bounds on the intent-to-treat (ITT) effect of being invited to the remedial camp on standardized endline test scores, addressing differential attrition between invited and comparison groups. Columns (1)–(3) report results for aggregate test scores, mathematics, and Spanish, respectively. Panel A reports pooled estimates across both cohorts, Panel B reports the 2024 cohort, and Panel C reports the 2025 cohort. The row labeled *Invited* reports the conventional ITT estimate using the full survivor sample. The *Lee Lower bound* and *Lee Upper bound* rows report estimates from re-running the ITT regression after trimming a fraction q of the higher-survival group's observed outcomes, where q equals the difference in survival rates divided by the higher survival rate. The row *Trim share (q)* reports the trimming fraction applied in each specification. Trimming is based on the rank of the residualized endline score, where residuals come from a regression of the outcome on baseline controls and randomization-strata fixed effects estimated on the full survivor sample. Imbens-Manski (2004) 95% confidence intervals are reported in brackets. Because Lee bounds partially identify the treatment effect under monotonicity, the intervals in brackets constitute the relevant inference for the identified set. Standard errors clustered at the school level are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE F.2. Lee Bounds on the Effects on Learning of Being Invited to the Remedial Camp (Raw Outcome Ranking)

	(1)	(2)	(3)
	Aggregate Ability	Math	Spanish
Panel A: Pooled years			
Invited	0.057***	0.050***	0.048***
	(0.015)	(0.016)	(0.015)
Lee Lower bound	0.037**	0.020	0.029*
	(0.015)	(0.016)	(0.015)
Lee Upper bound	0.092***	0.090***	0.086***
	(0.014)	(0.016)	(0.015)
Imbens-Manski 95% CI	[0.013, 0.116]	[-0.006, 0.115]	[0.003, 0.110]
Trim share (q)	0.021	0.027	0.019
Group trimmed	Invited	Invited	Invited
Observations	10,632	10,185	10,309
Panel B: 2024			
Invited	0.054*	0.005	0.072**
	(0.029)	(0.034)	(0.034)
Lee Lower bound	-0.021	-0.092***	-0.000
	(0.029)	(0.035)	(0.035)
Lee Upper bound	0.136***	0.132***	0.171***
	(0.030)	(0.033)	(0.034)
Imbens-Manski 95% CI	[-0.069, 0.186]	[-0.150, 0.187]	[-0.057, 0.228]
Trim share (q)	0.074	0.097	0.067
Group trimmed	Invited	Invited	Invited
Observations	2,633	2,282	2,368
Panel C: 2025			
Invited	0.056***	0.061***	0.040**
	(0.016)	(0.018)	(0.017)
Lee Lower bound	0.055***	0.058***	0.039**
	(0.016)	(0.018)	(0.017)
Lee Upper bound	0.056***	0.062***	0.041**
	(0.016)	(0.018)	(0.017)
Imbens-Manski 95% CI	[0.023, 0.088]	[0.025, 0.096]	[0.007, 0.074]
Trim share (q)	0.000	0.002	0.001
Group trimmed	Comparison	Invited	Invited
Observations	7,999	7,903	7,941

Note: This table reports Lee (2009) sharp bounds on the intent-to-treat (ITT) effect of being invited to the remedial camp on standardized endline test scores, addressing differential attrition between invited and comparison groups. Columns (1)–(3) report results for aggregate test scores, mathematics, and Spanish, respectively. Panel A reports pooled estimates across both cohorts, Panel B reports the 2024 cohort, and Panel C reports the 2025 cohort. The row labeled *Invited* reports the conventional ITT estimate using the full survivor sample. The *Lee Lower bound* and *Lee Upper bound* rows report estimates from re-running the ITT regression after trimming a fraction q of the higher-survival group's observed outcomes, where q equals the difference in survival rates divided by the higher survival rate. The row *Trim share (q)* reports the trimming fraction applied in each specification. In this robustness specification, observations are ranked directly using the raw endline outcome in the trimming step, rather than the residualized outcome. Imbens-Manski (2004) 95% confidence intervals are reported in brackets. Because Lee bounds partially identify the treatment effect under monotonicity, the intervals in brackets constitute the relevant inference for the identified set. Standard errors clustered at the school level are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE F.3. Lee Bounds on the Effects on Learning of Attending the Remedial Camp (Residualized Ranking)

	(1)	(2)	(3)
	Aggregate Ability	Math	Spanish
Panel A: Pooled years			
Share of days attended	0.103***	0.090***	0.086***
	(0.026)	(0.029)	(0.027)
Lee Lower bound	0.085***	0.060**	0.081***
	(0.027)	(0.029)	(0.027)
Lee Upper bound	0.188***	0.186***	0.167***
	(0.025)	(0.028)	(0.027)
Imbens-Manski 95% CI	[0.041, 0.229]	[0.013, 0.232]	[0.036, 0.211]
Trim share (q)	0.021	0.027	0.019
Group trimmed	Invited	Invited	Invited
Observations	10,632	10,185	10,309
Panel B: 2024			
Share of days attended	0.166*	0.016	0.222**
	(0.092)	(0.105)	(0.106)
Lee Lower bound	0.075	-0.100	0.222**
	(0.089)	(0.103)	(0.106)
Lee Upper bound	0.576***	0.561***	0.625***
	(0.099)	(0.108)	(0.115)
Imbens-Manski 95% CI	[-0.072, 0.738]	[-0.269, 0.739]	[0.047, 0.814]
Trim share (q)	0.074	0.097	0.067
Group trimmed	Invited	Invited	Invited
Observations	2,633	2,282	2,368
Panel C: 2025			
Share of days attended	0.090***	0.098***	0.066**
	(0.026)	(0.029)	(0.027)
Lee Lower bound	0.089***	0.092***	0.064**
	(0.026)	(0.028)	(0.027)
Lee Upper bound	0.090***	0.105***	0.068**
	(0.026)	(0.028)	(0.027)
Imbens-Manski 95% CI	[0.038, 0.141]	[0.041, 0.155]	[0.012, 0.120]
Trim share (q)	0.000	0.002	0.001
Group trimmed	Comparison	Invited	Invited
Observations	7,999	7,903	7,941

Note: This table reports Lee (2009) sharp bounds on the average causal response (ACR) of attending the remedial camp on standardized endline test scores, addressing differential attrition between invited and comparison groups. Columns (1)–(3) report results for aggregate test scores, mathematics, and Spanish, respectively. Panel A reports pooled estimates across both cohorts, Panel B reports the 2024 cohort, and Panel C reports the 2025 cohort. The endogenous variable is the share of camp days attended, defined as total days attended divided by 15. The instrument is the randomly assigned invitation to attend the camp. The row labeled *Share of days attended* reports the conventional 2SLS estimate using the full survivor sample. The *Lee Lower bound* and *Lee Upper bound* rows report estimates from re-running the 2SLS specification after trimming a fraction q of the higher-survival group's observed outcomes, where q equals the difference in survival rates divided by the higher survival rate. The row *Trim share (q)* reports the trimming fraction applied in each specification. Trimming is based on the rank of the residualized endline score, where residuals come from a regression of the outcome on baseline controls and randomization-strata fixed effects estimated on the full survivor sample. Imbens-Manski (2004) 95% confidence intervals are reported in brackets. Because Lee bounds partially identify the treatment effect under monotonicity, the intervals in brackets constitute the relevant inference for the identified set. Standard errors clustered at the school level are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE F.4. Lee Bounds on the Effects on Learning of Attending the Remedial Camp (Raw Outcome Ranking)

	(1)	(2)	(3)
	Aggregate Ability	Math	Spanish
Panel A: Pooled years			
Share of days attended	0.103***	0.090***	0.086***
	(0.026)	(0.029)	(0.027)
Lee Lower bound	0.067**	0.035	0.052*
	(0.027)	(0.028)	(0.028)
Lee Upper bound	0.167***	0.161***	0.155***
	(0.025)	(0.027)	(0.026)
Imbens-Manski 95% CI	[0.023, 0.209]	[-0.011, 0.206]	[0.007, 0.197]
Trim share (q)	0.021	0.027	0.019
Group trimmed	Invited	Invited	Invited
Observations	10,632	10,185	10,309
Panel B: 2024			
Share of days attended	0.166*	0.016	0.222**
	(0.092)	(0.105)	(0.106)
Lee Lower bound	-0.065	-0.286**	-0.000
	(0.093)	(0.110)	(0.110)
Lee Upper bound	0.423***	0.409***	0.519***
	(0.098)	(0.109)	(0.111)
Imbens-Manski 95% CI	[-0.219, 0.585]	[-0.467, 0.589]	[-0.180, 0.701]
Trim share (q)	0.074	0.097	0.067
Group trimmed	Invited	Invited	Invited
Observations	2,633	2,282	2,368
Panel C: 2025			
Share of days attended	0.090***	0.098***	0.066**
	(0.026)	(0.029)	(0.027)
Lee Lower bound	0.089***	0.095***	0.064**
	(0.026)	(0.028)	(0.027)
Lee Upper bound	0.091***	0.101***	0.067**
	(0.026)	(0.029)	(0.027)
Imbens-Manski 95% CI	[0.038, 0.142]	[0.042, 0.154]	[0.012, 0.119]
Trim share (q)	0.000	0.002	0.001
Group trimmed	Comparison	Invited	Invited
Observations	7,999	7,903	7,941

Note: This table reports Lee (2009) sharp bounds on the average causal response (ACR) of attending the remedial camp on standardized endline test scores, addressing differential attrition between invited and comparison groups. Columns (1)–(3) report results for aggregate test scores, mathematics, and Spanish, respectively. Panel A reports pooled estimates across both cohorts, Panel B reports the 2024 cohort, and Panel C reports the 2025 cohort. The endogenous variable is the share of camp days attended, defined as total days attended divided by 15. The instrument is the randomly assigned invitation to attend the camp. The row labeled *Share of days attended* reports the conventional 2SLS estimate using the full survivor sample. The *Lee Lower bound* and *Lee Upper bound* rows report estimates from re-running the 2SLS specification after trimming a fraction q of the higher-survival group's observed outcomes, where q equals the difference in survival rates divided by the higher survival rate. The row *Trim share (q)* reports the trimming fraction applied in each specification. In this robustness specification, observations are ranked directly using the raw endline outcome in the trimming step, rather than the residualized outcome. Imbens-Manski (2004) 95% confidence intervals are reported in brackets. Because Lee bounds partially identify the treatment effect under monotonicity, the intervals in brackets constitute the relevant inference for the identified set. Standard errors clustered at the school level are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.