NORC at the University of Chicago

NORC WORKING PAPER SERIES

Attrition in Randomized Control Trials: Evidence from Early Education Interventions in Sub-Saharan Africa

WP-2021.02 | December, 2021

PRESENTED BY:
NORC at the University of Chicago
55 East Monroe Street
30th Floor
Chicago, IL 60603
Ph. (312) 759-4000

AUTHORS:
Alejandro Ome
Cally Ardington
Alicia Menendez

AUTHOR INFORMATION*

Alejandro Ome
NORC University of Chicago
ome-alejandro@norc.org

Cally Ardington
University of Cape Town
cally.ardington@uct.ac.za

Alicia Menendez
University of Chicago and NORC at the University of Chicago
menendez@uchicago.edu

Table of Contents

List of Tables

List of Figures

Abstract

In this study, we analyze the determinants and consequences of survey attrition in the context of three early reading RCT evaluations recently conducted in Zambia, Ethiopia, and South Africa, that used student longitudinal data. To study the determinants of attrition, we analyze whether treatment status and other students' baseline characteristics are correlated with attrition. We find that while treatment status is in most cases not correlated with attrition, students' baseline characteristics are; in particular, baseline reading skills are negatively correlated with attrition. To study the consequences of attrition, we apply inverse probability weighting (IPW), Lee bounds and Manski bounds to each experiment. We show that for both IPW and Lee bounds the results do not change much with respect to the intent to treat parameters. Manski bounds are not informative for any experiment.

1. Introduction

The last few decades have seen an exponential growth in the number of evaluations of early education interventions in low-income countries. To assess the impact of these interventions, evaluations have mostly relied on cross-sectional samples of students. More recently, several studies use longitudinal data, which involves following a sample of students over time. In general, longitudinal data offer advantages over cross-section data such as the possibility of controlling for students' baseline characteristics reducing the risk of omitted variable bias, detecting sequences of events, and measuring duration of events. Importantly, if outcomes of interest are correlated with student performance and characteristics at baseline, longitudinal studies deliver more precise estimates of intervention impacts. However, survey attrition is an important threat to the validity of evaluations that use longitudinal data. Survey attrition has three main consequences. First, a smaller sample reduces the precision of the estimated impacts. Second, unless individuals leave the sample at random, attrition affects the makeup of the final sample, making it no longer equivalent or comparable to the original sample. Therefore, if the original baseline sample was representative of a given population, the endline sample will not necessarily be, and external validity would be compromised. Third, sample deterioration could be such that treatment and control groups are no longer comparable, compromising the internal validity of the results. In this study, we analyze the determinants and consequences of survey attrition in the context of three early reading RCT evaluations recently conducted in Zambia, Ethiopia, and South Africa, that used student longitudinal data.

Attrition in early education studies is mostly driven by students dropping out of school and movement of students between schools. In studies where students' assessments and interviews take place at schools, school absenteeism is probably the main driver of survey attrition. Furthermore, treatment itself can induce individuals to attrit from or remain in the sample.[1] It is also plausible that interventions that improve learning outcomes could also increase attendance. In this study, we apply different econometric techniques to address attrition to analyze how sensitive the program evaluation results are to these corrections.

The literature provides examples of the different ways in which researchers have dealt with attrition in evaluations of early educational interventions in developing countries. A common approach is to report whether attrition is correlated with treatment status. For example, in a work that evaluates the effects of providing laptops vs. traditional textbooks in primary schools in Honduras, Bando et al. (2017) report that

[1] For example, interventions like conditional cash transfers (Fiszbein and Schady 2009; Murnane and Ganimian 2014; Glewwe and Muralidharan 2016) or school feeding programs (Alderman and Bundy, 2012; Drake et al. 2017) tend to affect attendance and therefore the likelihood that students are re-surveyed at endline.

the attrition rate after one year is 7 percent, and that attrition is not correlated with treatment. An evaluation of different approaches to teachers' coaching delivery in South Africa, (Kotze et al. 2019) indicates that there is no correlation between treatment and the 8.6 percent attrition rate that they find during a shorter-than-one-year period. Other research goes a step further into establishing the relationship between attrition and treatment status. Aurino et al. (2020) study whether an intervention providing child feeding in Ghanaian schools is associated with some child characteristics in predicting likelihood of remaining in the sample by interacting treatment assignment with the background characteristics. They find that children from poor households tend to be older in the treatment group and employ age-standardized outcome variables in an attempt to address the problem. Cilliers et al. (2018), compare centralized teacher training and teachers' in-classroom coaching in 230 primary schools in the Northwest Province in South Africa. Using baseline data, they show that the 16.8 percent rate of attrition is not correlated with treatment status and estimate the relative differences between those that attrit and those that do not in the control and treatment schools by regressing students' characteristics on treatment, attrition, and the interaction between treatment and attrition. They do not find any significant differences between groups. A very similar approach is used by He et al. (2009) in their evaluation of the Pratham Shishuvachan preschool program. The attrition rate is 24.6 percent in a period shorter than a year and the authors find some differences between groups but deem them small enough to disregard them. In contrast, evaluations of conditional cash transfers like Progresa have found differential attrition rates between treatment and control groups. Parker, Rubalcava, and Teruel (2006) indicate that the first Progresa studies did not take into account attrition when estimating the program impacts, but later studies did. For example, Behrman, Parker, and Todd (2005) used a difference-in-differences approach combined with a density reweighting method to take into account attrition.

The focus on differential attrition rates between treatment and control groups in much of the literature is potentially misleading. Firstly, if attrition is completely random, then there are no internal or external validity concerns. The only implication is a reduction in precision. This is true, even with differential attrition between treatment and control, as long as within each group attrition is completely random.

Conversely, a lack of observed differential attrition between treatment and control does not imply that there is no threat to internal and external validity. If the drivers of attrition are different in each group, then estimates of treatment effects could still be biased. If particular individuals are more likely to attrit from both groups, then inferences can no longer be made about the full population from which the sample was drawn.

The most common approaches found in the evaluation literature to address attrition are inverse probability weighting (IPW), Lee bounds and Manski bounds. IPW uses the propensity to attrit to construct weights to overweight students who were more likely to leave the sample. This method corrects for attrition assuming selection on observable characteristics only. The Lee bounds approach trims the experimental group with the lower attrition rate, so the evaluation uses only those who are always observed (Lee 2002). This method assumes that treatment affects attrition in one direction only (monotonicity). Manski bounds impute missing outcome data for treatment and control groups using the maximum and minimum possible values of the outcome distribution (Horowitz and Manski 2000).

In this study we analyze whether treatment status and other students' baseline characteristics are correlated with attrition using data from external impact evaluations we conducted in the past few years. We collected data and evaluated the impact of one experiment in Zambia, four experiments in Ethiopia, and two experiments in South Africa, for a total of seven experiments. We find that while treatment status is in most cases not correlated with attrition, students' baseline characteristics are; in particular, baseline reading skills are negatively correlated with attrition. Next, we apply IPW and Lee and Manski bounds to each experiment. We show that for both IPW and Lee bounds the results do not change much with respect to the "raw" intent to treat (ITT) parameters, meaning that for the experiments for which the ITT are (positive) and statistically significance we find significant IPW and informative Lee bounds, and for the cases where the ITT are not significant, the IPW are also not significant and Lee bounds are not informative. Manski bounds are not informative for any experiment.

This study makes two main contributions to the evaluation literature on early reading interventions. First, it analyzes what student and household characteristics correlate with attrition. Policymakers are often concerned about program impacts for different subpopulations (e.g., low-skill readers, socio-economically disadvantaged students, boys/girls, etc.). The fact that certain types of students (e.g., low-skill readers) are more likely to attrit from the sample, casts doubts on the validity of program impacts on these types of readers. Second, it provides evidence that, in the context of early education interventions, under certain conditions, survey attrition probably has minor consequences on *average* treatment effects.

This paper has four additional sections. In section 2, we review the implications of survey attrition and the methods used in the literature to control for it. We discuss the data sets we use in section 3. In section 4 we describe the results. Section 5 concludes.

2. Methods to address sample attrition

To summarize the implications of survey attrition, we consider a simple two-period sample selection model, as proposed by Fitzgerald, Gottschalk, and Moffitt (1998), where the outcome of interest at endline, y_1, is a function of the (randomized) treatment status, denoted by *D*, student and households' characteristics at baseline x_0, and an error term ε_1.

$$y_1 = \alpha + \beta D + \gamma x_0 + \varepsilon_1 \quad (1)$$

Whether y_1 is observed, is determined by the attrition model described by:

$$A^* = \delta + \theta D + \vartheta x_0 + \pi z_0 + \mu_1$$

and,

$$A = \{0 \; if \; A^* < 0 \; 1 \; if \; A^* \geq 0$$

where A^* is a latent attrition variable that is a function of all the covariates in equation (1) and variables that affect sample attrition but not the outcome of interest, z_0, and *A* is an indicator variable for whether y_1 is observed or not. If the error terms ε_1 and μ_1 are correlated, then a regression of (1) using only the observed data leads to a biased estimate of β, as:

$$E[y_1|D, x_0, z_0, A = 0] = \alpha + \beta D + \gamma x_0 + E[\varepsilon_1|D, x_0, z_0, \mu_1 < -(\delta + \theta D + \vartheta x_0 + \pi z_0)]$$

We could identify β if it were possible to estimate:

$$E[y_1|D = 1, x_0, z_0, \mu_1 < -(\delta + \vartheta x_0 + \pi z_0)] = \alpha + \beta D + \gamma x_0 + E[\varepsilon_1|D = 1, x_0, z_0, \mu_1 < -(\delta + \vartheta x_0 + \pi z_0)]$$

which is the expected value of the outcome of interest for a subset of the treatment group, specifically those that would have not attrited even if they had not been treated. Note that all the data to estimate this is observed, the problem is that we cannot differentiate between the individuals that would not have attrited even if they had not been treated ($\mu_1 < -(\delta + \vartheta x_0 + \pi z_0)$), and those that did not attrite because were treated ($-(\delta + \vartheta x_0 + \pi z_0) \leq \mu_1 < -(\delta + \theta D + \vartheta x_0 + \pi z_0)$).

Several approaches to deal with this issue have been proposed in the literature. Lee (2009) proposes bounding the treatment effect by assuming worst/best case scenarios for those that are observed because they were treated and trimming those observations from the treatment group. This reduces to trimming the treatment group from above and below by the fraction:

$$p = \frac{Pr\,[D = 1]\, - Pr\,[D = 0]}{Pr\,[D = 1]}$$

Trimming the treatment group in this way provides bounds for the treatment effect. Other than random assignment, the underlying assumption is monotonicity on the effect of treatment assignment on sample selection. This means that treatment should either increase or decrease the likelihood that some students are surveyed, but not increase it for some cases and decrease it for others. In early reading interventions, this means that if the attrition rate is lower for the treatment group than for the control group, all the observed students in the control group would have been observed even if they had been treated, hence there are no control students who were observed only because they were assigned to the control group. In this context, monotonicity guarantees that the comparison is made with control students who were going to be observed regardless of their treatment status, which is the equivalent of what the proposed bounds for the trimmed treatment sample encompass.

Rather than trimming the sample, another option is to impute the missing data. Horowitz and Manski (2000) propose a worst/best case scenario approach where missing data are imputed using extreme values of the outcome of interest. We can write the expectation of the outcome of interest as:

$$E[y_1|D, A] = E[y_1|D, A = 1]Pr\,[A = 1]\, + E[y_1|D, A = 0]Pr\,[A = 0]$$

where the only construct in the right hand side that is not observed is $E[y_1|D, A = 0]$. However, for each censored case we know that the outcome value falls between the minimum and maximum values that the outcome of interest can take. In the case of a reading assessment, it is natural to use zero as the lowest possible value, and the total number of words in the reading piece as the maximum value. Therefore, the worst possible scenario, that is, the most conservative imputation we can make of the missing data, is that all censored outcome data in the treatment group equals zero, and all censored outcome data in the control group equals the total number of words. Conversely, the best-case scenario is to impute all missing data in the treatment group as the total number of words, and all missing data in the control group as zero. Kling and Liebman (2004) propose tighter bounds by not using extreme values for the imputation, but rather the mean plus/minus some fraction of the standard deviation of the observed distributions. Obviously, this produces much tighter bounds, but it also relies on untestable (and strong) assumptions on the distribution of the missing data.

An extension to Lee bounds consists of imputing the missing data for the compliers in the control group using a worst/best case scenario approach. Under monotonicity, this method bounds the treatment effect for the always observed and the compliers. Huber and Mellace (2015) propose a similar approach except they bound the treatment effect for all selected, which includes the always observed both in the treatment

and control groups, and the compliers. The method we propose bounds the treatment effect for the always observed in the treatment group only, and the compliers.

Rather than bounding the treatment effect, researchers often try to "correct" for attrition, by modeling the sample selection process. Selection driven by unobservables can be corrected following a Heckman approach (Heckman 1979), as long as there are valid instruments for selection. If researchers have access to variables that explain attrition but not the outcome of interest, then consistent estimates of β can be derived using a two-step estimator. The limitation of this approach is that finding valid instruments for attrition can be challenging. In the context of early reading interventions, for example, where student assessments and interviews usually take place at school, attrition is likely associated with attendance and dropping out of school, and therefore any variable that affects attrition will also have an impact on academic performance. Given this practical limitation, sample selection bias can only be corrected if the process is driven by observable characteristics.

A more feasible approach involves assuming that ε_1 and μ_1 are independent, i.e. selection on observables. Fitzgerald et al. (1998) propose estimating a weighted regression of the outcome of interest, where the weights are based on the estimation of two attrition models, one where attrition is modeled as a function of the x_0, the variables in the structural model of y_1, and another where auxiliary variables are also included. These auxiliary variables may be correlated with y_1 but do not enter the structural or theoretical model of y_1. In the context of estimating an earnings regression, Fitzgerald and coauthors argue that while education and experience are part of the structural model, other variables like occupation and industry, that might be correlated with y_1, can be used as auxiliary variables because they do not enter the structural or theoretical model. This method involves making decisions about which variables enter the structural model of the outcome of interest, and which can be used as auxiliary variables. Wooldridge (2002) proposes an alternative approach where only one attrition model is estimated, and the weights are constructed using the estimated probabilities of such model. Intuitively it is easy to see how weighting can correct for attrition. If individuals that were more likely to attrite from the sample but were actually observed are overweighted in the regression, then the parameters for the original sample are more likely to be recovered.

3. Three early reading evaluations in Sub-Saharan countries

We use longitudinal data collected for external impact evaluations recently conducted in three Sub-Saharan countries to explore determinants and consequences of attrition, and to analyze the sensitivity of the evaluation results to alternative correction approaches. The focus of these three projects is to improve

early reading skills, so all respondents are primary school children, and the outcomes of interest are reading skills, measured using different versions of the Early Grade Reading Assessment – EGRA (RTI International 2015). We focus on oral reading fluency (number of correct words read per minute) as the outcome of interest. All data collection activities were conducted at the schools, so attrition could be the consequence of children moving to another school, dropping out or simply not attending the day of data collection. Across all interventions treatment was assigned randomly.

The Makhalidwe Athu project (MA) – Eastern Province, Zambia

The MA project was a nine-month pilot intervention aimed at improving the reading skills of ~1,200 students in grades 2 and 3 in two districts of Zambia's Eastern province. The program was designed and implemented by Chemonics International. Every week, MA sent three SMS messages comprising a short story for the students to read with their families. In addition, participants could call in for a prepaid recorded voice message (IVR), which included comprehension questions, and a recording of the story itself.

To evaluate MA, we randomly assigned 80 schools to treatment or comparison groups. Baseline data from 15 2nd graders and 15 3rd graders were collected in January 2016. One comparison school could not be surveyed due to weather conditions, so only 79 schools are included in the analysis. In total, 2,263 students answered a version of the EGRA that included five subtasks in ChiNyanja language, namely letter sound identification, non-word decoding, oral reading fluency, reading comprehension, and listening comprehension.

Endline data were collected in January 2017. Out of the 2,263 students surveyed at baseline, 1,973 were surveyed at endline, for an attrition rate of 12.8 percent. In parallel, one caregiver for each child was surveyed both at baseline and endline, to record basic sociodemographic characteristics, and children's reading habits at home. Ome and Menendez (2020) find that the program had a positive and significant impact on oral reading fluency of 3.4 words per minute.

Reading for Ethiopia's Achievement Developed Community Outreach (READ CO)

The Ethiopian program READ CO--designed and implemented by Save the Children--provided supplementary reading materials, supported school and community reading activities, and trained teachers and parent-teacher-student associations on how to manage and promote these activities effectively.

To evaluate this program, we randomized 150 schools into one of two treatment groups and a control group. One treatment group fielded school-based reading activities only, while the other implemented school-based and community-based reading activities. School-based activities included support to set up reading corners and reading/writing clubs in schools, among other activities. Community-based activities included setting up book banks and reading camps. The impact evaluation of READ CO was conducted in the regions of Amhara and Oromia.

Baseline data were collected in February 2016. The target was to survey 30 2nd graders in each school, for a total sample of 4,500 students. At baseline 4,275 students were surveyed. Reading skills were measured using a version of EGRA in Amharic and Afaan Oromo languages.

Endline data were collected in May 2018. Out of the 4,273 students surveyed at baseline, 2,495 were surveyed at endline, for an attrition rate of 41.6 percent. This attrition rate varies substantially by region, which is why we conducted separate analyses by region, as well as by treatment arm. Attrition in this study is high and different factors explain his phenomenon. Oromia and Amhara regions are the main origin areas of internal migrants in Ethiopia (Bundervoet 2018). In addition, school absenteeism and drop-out rates are both high[2].

While we focus on the student surveys for this paper, it is worth mentioning that this study also surveys, both at baseline and endline, samples of students' parents, teachers and head teachers from the 150 schools.

The results show that the program had positive impacts only for the school-based treatment group, and only in Amhara (NORC 2018). At endline, students in the school-based treatment group in Amhara read 7.5 correct words per minute more than their counterparts in control schools. By contrast, we found no significant effects in schools with School + Community interventions in Amhara. This result may be explained by the fact that books and other resources in the School + Community arm were split between school-based and (out-of-school) community-based activities. In other words, schools in the two treatment arms received the same number of books, but schools in the School + Community group needed to split these resources between, for example, reading corners in the schools and book banks located outside the schools.

[2] Rates of absenteeism based on rosters of schools included in our sample tended to range from 20%-35%. In addition, data from EMIS suggest a drop-out rate of approximately 20% in two years (NORC, 2015).

In Oromia, READ CO implementation faced major challenges and delays and we did not find evidence that the program affected reading skills in the area.

Nal'ibali Story Powered Schools project (SPS)

SPS aims to develop and sustain a culture of reading for enjoyment in 720 primary schools in the Eastern Cape (EC) and KwaZulu-Natal (KZN) provinces of South Africa. The program--designed and implemented by Nal'ibali--focuses on developing and nurturing reading habits in mother tongue and English. Several activities take place at each SPS school. These include training teachers and community volunteers in reading for enjoyment, providing reading materials, visits to schools by reading-for-enjoyment mentors, supporting schools to register reading clubs, organizing holiday programs that include reading activities, and events, campaigns and reading competitions.

To evaluate the effects of the SPS project we collected quantitative data in two rounds: a baseline prior to program implementation, and an endline two (academic) years later. To align with the program implementation, the baseline was conducted in two years. Half of the baseline school visits took place in February/March 2017 and the rest around the same time the following year. The endline was also conducted over two years; the cohort assessed in 2017 was re-interviewed in October/November 2018, while the second cohort was assessed in October/November of 2019. We analyzed the data pooling the cohorts and also separately. The results are very similar and therefore, we present here the pooled data.

In each of the 360 schools randomly selected for data collection, we sampled 30 students from grades 2, 3, and 4 at random, creating a total sample of 10,404 students. At baseline, we interviewed 5,002 students in the EC and 5,372 in KZN. We followed these students for two years and at endline we re-interviewed 4,284 in the EC and 4,490 in KZN, for an attrition rate of 14 and 16 percent respectively. The students in the panel were interviewed and assessed using subtasks of the EGRA and several other reading and writing exercises in isiZulu and isiXhosa.

SPS does not show an impact on learners' reading and writing skills in mother tongue or in English. We tested a wide array of reading and writing skills and the estimated effects are all very small and not statistically different from zero for any of the outcomes of interest in either province. Around 15 percent of learners in SPSs attend reading clubs. Compared to learners in control schools with similar baseline characteristics, these learners appear to read on their own at home more frequently and for fun in the holidays. In the Eastern Cape, reading club participants also have higher fluency and do better on the oral reading and productive listening comprehension tasks than the matched comparison learners in control schools. The effects range from 0.15 to 0.26 standard deviations.

On average, the take-up and adherence to SPS is low. Program activities did not take place in the schools as expected. A large number of schools did not fulfill the minimum expectations of the program and very few met the target for all components.

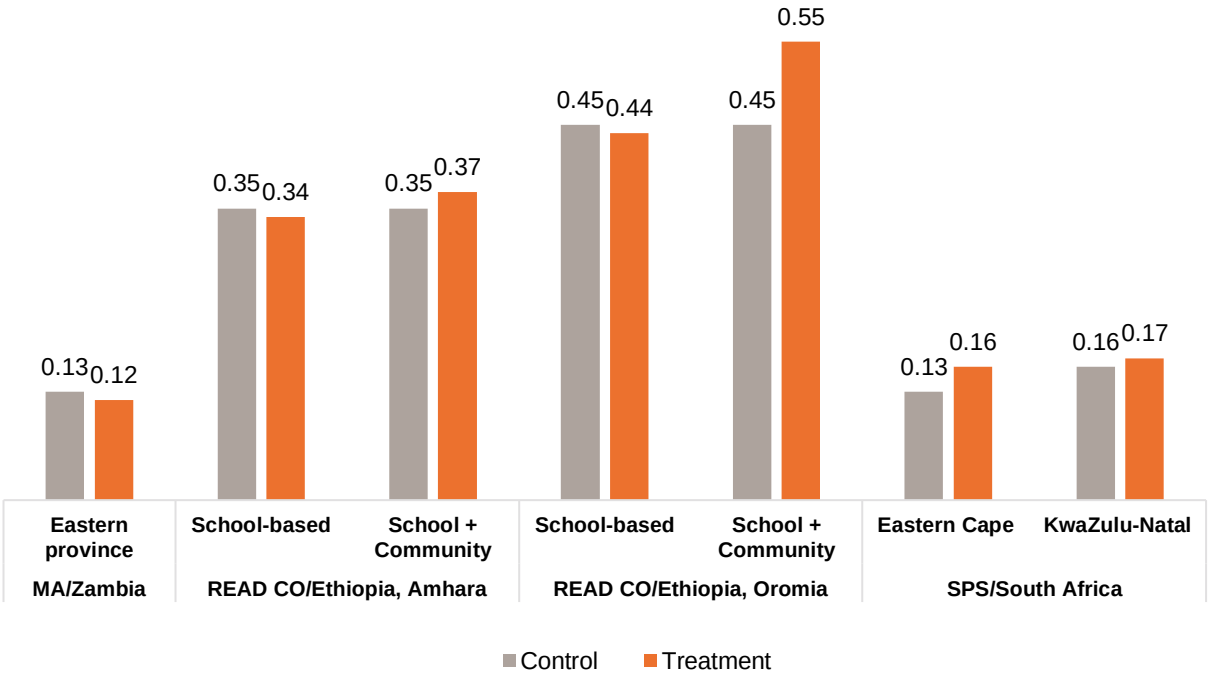
4. Addressing attrition

4.1 Determinants of attrition

Attrition rates by treatment status for each experiment are shown in Figure 1. Attrition rates vary substantially between experiments. While for MA/Zambia and SPS/South Africa attrition rates are between 12-17 percent depending on the experiment and treatment status, in Ethiopia the range is 34-55 percent. As discussed in the previous section, the high attrition observed for Ethiopia is explained mainly by high rates of student absenteeism and dropout, in addition to migration. However, there are no major differences between treatment status within each experiment. In most cases the difference is 1-2 percentage points, with the notable exception of School + Community intervention in Oromia (READ CO/Ethiopia), where the treatment group has an attrition rate of 55 percent, 10 percentage points larger than the corresponding control group.

Not only are the differences in attrition rates by treatment group small, the sign of the difference is not even consistent across experiments. In fact, while attrition is higher for the control than for the treatment group for MA/Zambia and the two School-based experiments in READ CO/Ethiopia, the opposite pattern is observed for the two School + Community experiments in READ Co/Ethiopia, and both experiments in South Africa. This suggests that treatment status is not correlated with attrition, which is consistent with previous studies. In fact, in a review of 91 recent field experiments, Ghanem et al. (2020) show that only 12 percent of the surveyed studies have differential rates greater than 5 percentage points, and for 66 percent the differences are less than 2 percentage points.

Figure 1. Attrition rates by treatment status



To analyze the determinants of attrition, Table 1 shows logistic regression results where the dependent variable is a dummy for attrition and the covariates are treatment status and student characteristics, namely students' baseline reading skills, sex, age and an asset index. The objective of this exercise is to identify patterns in terms of what seem to be determinants of attrition across all experiments, which is why we only include variables that are available across all three studies. In the following section, we consider many more variables as determinants of attrition, with the purpose of producing more saturated models. The coefficients shown are marginal impacts evaluated at the mean of each independent variable, so they can be interpreted as percentage points of the outcome of interest. Reading skills are measured as the first principal component of the EGRA subtasks fielded in each country. The asset index corresponds to the first principal component of indicator variables for whether households own a range of goods including radios, televisions, refrigerators and cars.[3] Confirming what was mentioned before, there is no correlation between treatment status and attrition, as the corresponding coefficients are not significant, with the sole exception of the School + Community experiment in Oromia (READ CO/Ethiopia).

[3] In Zambia and Ethiopia, the items were a chair, a bed, a clock, a radio, a television, a computer, a bicycle, a motorcycle, a car, a refrigerator, and a stove. In South Africa, the items were a radio, a television, a computer, a refrigerator, a flush toilet, a mobile phone, a bicycle, and a car.

There is a negative and statistically significant correlation between baseline reading skills and attrition across all seven experiments; this could reflect the fact that low-skill students also tend to miss more days of school and/or are more likely to drop out of school or change schools. There is no clear pattern for the female dummy. Within the Ethiopian sample the coefficient is always negative but significant only in two cases. The coefficient on age is positive across countries--indicating the older children are more likely to attrit--but only significant in the Ethiopian sample. A possible explanation for this finding is that the opportunity cost of going to school tends to be higher for older children who are more likely to help their parents at home or in farms, and work and the percentage of children engaged in child labor is substantially higher in Ethiopia than in Zambia and in South Africa (UNICEF 2019). Finally, the only significant correlation found between attrition and household wealth was in the Zambia study.

Table 1. Correlates with attrition

	MA/Zambia Eastern province	READ CO/Ethiopia				SPS/South Africa	
		Amhara		Oromia			
		School -based	School + Community	School -based	School + Community	Eastern Cape	KwaZulu-Natal
Treatment	-0.01	-0.02	0.00	0.01	0.11*	0.03	0.01
	(0.02)	(0.04)	(0.05)	(0.05)	(0.05)	(0.02)	(0.02)
Reading score at baseline	-0.02*	-0.11***	-0.10***	-0.11***	-0.09***	-0.02***	-0.02**
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
Female	0.02	-0.05	-0.08**	-0.07**	-0.03	-0.01	0.01
	(0.01)	(0.03)	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)
Age	0.01	0.02**	0.03***	0.05***	0.07***	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)
Asset index	0.02*	-0.01	0.03	0.00	0.01	0.01	-0.00
	(0.01)	(0.03)	(0.03)	(0.03)	(0.03)	(0.01)	(0.00)
N	2254	1452	1473	1278	1247	4784	5331

Note: Marginal effects evaluated at the mean of the independent variables, after running logistic regressions. Regressions for MA/Zambia and SPS/South Africa include grade-fixed effects (READ CO/Ethiopia is just one grade). Standard errors are clustered at the community level for SPS/South Africa and at the school level for all other experiments. Asterisks denote significance at 0.1% (***), 1% (**) and 5% (*) level.

To explore non-linear relationships between baseline characteristics and attrition, we also estimate attrition models using quantiles for baseline reading skills, assets, and age. The results, shown in Annex I, confirm a negative relationship between reading skills and attrition, a positive correlation between age and attrition, and no or very weak relationship between household assets and attrition.

These results have a few important implications. First, if attrition rates do not vary much by treatment status, Lee bounds should be relatively tight and probably informative. This would not apply to

interventions that affect enrollment and/or attendance and therefore the likelihood that students are surveyed at endline. In those cases, and as long as data are collected at the school rather than at home, it is very likely that attrition will be correlated with treatment status, hence Lee bounds will probably be wide and non-informative.

Second, the fact that baseline reading skills and attrition are negatively correlated indicates that external validity of the evaluation may be compromised due to attrition, in the sense that the evaluation results do not necessarily apply to low-performing students. Note that this is true even if differential attrition between treatment groups is negligible. Moreover, it is interesting that while reading skills are strongly correlated with attrition, household wealth, as measured by the asset index, is not.

Third, the fact that age and sex are correlated with attrition indicates that overaged and male children are underrepresented in endline samples, compromising the external validity of the evaluation results. Older children are more likely to have repeated grades and probably have a higher opportunity cost to go to school than younger children, which would explain their greater likelihood to attrite from the sample. It is perhaps more surprising to find that males are relatively more likely to attrite. This corroborates findings that conditioned on enrollment, girls are more likely to stay in school than boys in developing countries (Grant and Behrman 2010).

4.2 Implications of attrition

In this section we discuss the results of using different methods to address attrition, namely IPW, and Lee and Manski bounds.

To calculate IPW estimates we run weighted regressions, using as weights the inverse of one minus the predicted probabilities of logit regressions modeling attrition. In the previous section, we present attrition regressions, using a limited set of covariates common across the experiments. In this section, our purpose is to model attrition using all available information, so for each experiment we include variables in the attrition model that are not necessarily across all data sets. For example, for Zambia, we have household size and birth order; for Ethiopia self-reported attendance, and for the South Africa height and whether there are books to read at home. [4,5]

[4] Children's height is considered a good indicator of long run and underlying health, as it reflects cumulative linear growth, and deficiencies show the effects of past or chronic inadequacies in nutrition or exposure to disease.
[5] Results for the full attrition models are available upon request.

By adding more variables to the attrition models, we expect to improve their performance to predict attrition. Because in each data set there are several variables that can be included, we follow the stepwise approach proposed by Doyle et al. (2016) to select a subset of covariates. As a robustness check we also used LASSO methods (Tibshirani 1996) to select predictors for the attrition model, following Molina and Macours (2019) and Ahrens, Hansen, and Schaffer et al. (2018). In **Annex II** we explain how we implement these methods, and we include the list of variables considered in the models. The main results are summarized in Table 2, which shows the adjusted R-squared for each model and experiment. The first row shows results for the models that include the reduced set of covariates, shown in Table 1; the second row shows results for the stepwise approach and in the third row we add school fixed effects to the models in the second row. Finally, the fourth row shows results for the LASSO method. Notably, all these approaches produce similar levels of adjustment, and even when we include school-fixed effects in the regressions, the R-squared is relatively low.

These results can be interpreted in two different ways. One interpretation is that while attrition is largely driven by different factors that may or may not be correlated with reading skills, the variables that we are using do not capture this data-generating process. This would cast doubts on the validity of IPW as a way to correct for attrition, as the model is not controlling for all the relevant variables. An alternative interpretation is that there is substantial randomness driving attrition in these contexts given that despite including a relatively large set of covariates, and even after saturating the model with school-fixed effects, the R-squares barely move. This interpretation is credible as absenteeism is high and common across students in these samples. Unfortunately, we cannot test whether attrition is random after controlling for the covariates we include in the model.

Table 2. Adjusted R-squared for Linear Probability models on attrition

	MA/Zambia	READ CO/Ethiopia				SPS/South Africa	
		Amhara		Oromia			
	Eastern province	School-based	School + Community	School-based	School + Community	Eastern Cape	KwaZulu-Natal
Reduced set of covariates	0.003	0.053	0.048	0.056	0.068	0.005	0.004
Stepwise approach	0.022	0.083	0.074	0.079	0.076	0.026	0.013
Stepwise approach + school FE	0.056	0.134	0.139	0.128	0.139	0.096	0.069
LASSO	0.021	0.089	0.074	0.073	0.063	0.014	0.017

Figure 2 shows IPW treatment effects using the attrition model derived from the stepwise approach aforementioned (the IPW results using LASSO, shown in **Annex II**, are very similar to the ones presented below), as well as Lee bounds, Manski bounds, and bounds à la Kling and Liebman. For each experiment we also show the "raw" ITT as a benchmark. Across all experiments IPW and ITT are very similar. For the experiments for which the ITT parameter is statistically significant, namely the Zambia/MA program and the school-based program in Amhara, Ethiopia, the IPW estimates are basically the same as the ITT, while for the other five experiments, where the ITTs are not significant, the IPW estimates are not significant either. Although these results suggest that attrition does not have major implications on treatment effect estimation, the fact that the attrition models perform relatively poorly casts doubts on the validity of the assumption that attrition is driven by the observable characteristics required for the IPW approach.

We found similar results using Lee bounds and the modified Lee bounds we propose, in the sense that for the two experiments for which we found significant ITT effects, we also found informative bounds; and conversely, for the interventions for which we do not find an ITT effect, Lee (and Lee modified) bounds are non-informative. These results are hardly surprising given that attrition rates are relatively similar between treatment and control groups within each experiment, except for the READ CO School + Community experiment in Oromia, Ethiopia, where the attrition rate for the treatment group is 10 percentage points higher than for the control group. In this case, the bounds are not informative but there is no significant ITT anyway.

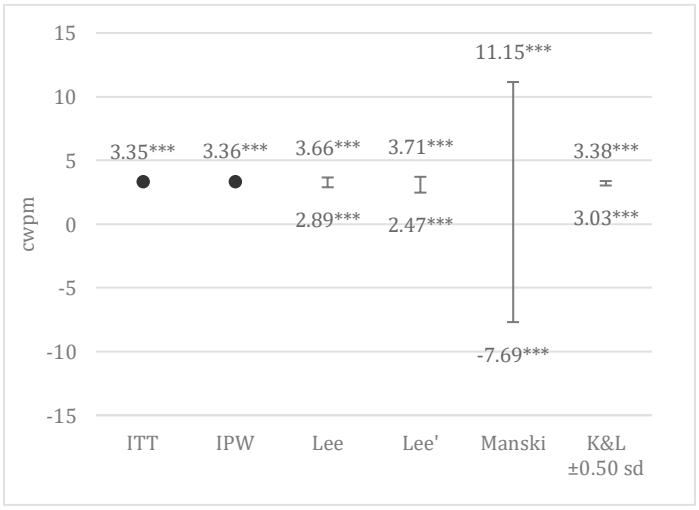
Lee bounds only document the treatment effects on those "always" observed, not on the ones that are never observed at endline regardless of their treatment status, or the ones that are observed contingent on their treatment status. Even the modified Lee bounds we propose do not document the impact on the never observed. Given that we showed that underperforming students are less likely to be observed across all experiments, knowing whether the program has a heterogeneous impact across baseline reading skills would shed light on whether the Lee bounds are under or over estimating the impact of the program over the children initially sampled. In the next section, we explore this type of treatment heterogeneity and discuss implications for the validity of the Lee bounds.

None of the bounds for the Manski bounds are informative, even when overall attrition is relatively small, like in MA/Zambia. Given that we are looking not at a binary outcome variable but at one that measures correct words read per minute that can go from 0 to over a hundred, it is not surprising that imputing these extreme values would lead to such wide bounds. To analyze whether looking at a binary variable results in informative bounds, in **Annex III** we estimate Manski bounds for an indicator variable for whether

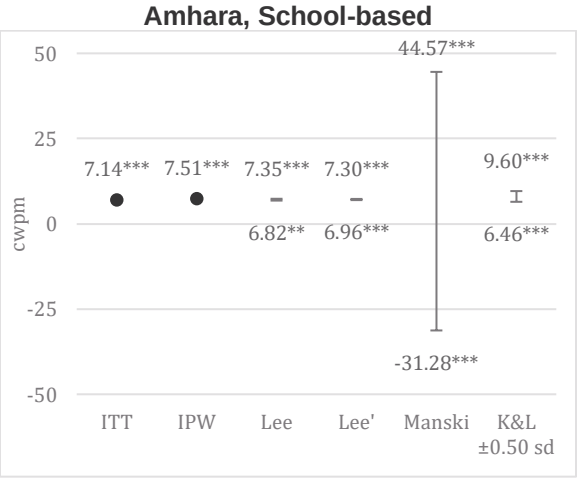
students can read at least one word that, albeit not the focus of this study, is an outcome of interest in and of itself, often reported in early reading interventions in developing countries, where there are large fractions of students that cannot read a single word. However, even when we use this binary variable, Manski bounds do not produce informative bounds for any experiment.

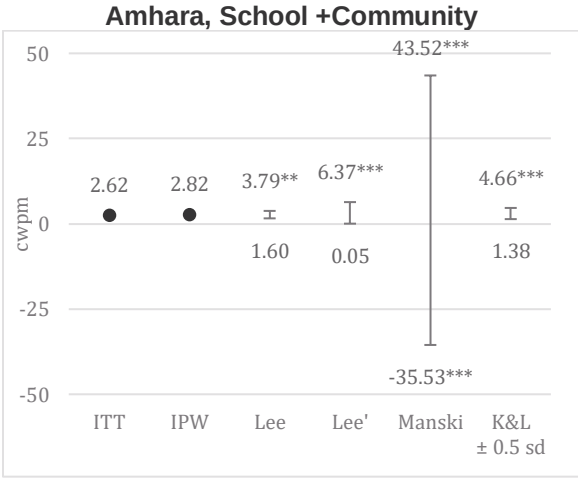
Finally, bounds à la Kling and Liebman produce informative bounds for the two experiments for which we find a significant ITT effect, and non-informative for the rest. These bounds follow the same idea as Manski's but rather than using the minimum and maximum possible values for the imputations, use functions of the mean and standard deviation that are much less "extreme," producing bounds tighter than Manski's by construction. However, whether the unobserved outcome means fall within the bounds proposed by Kling and Liebman is obviously untestable.

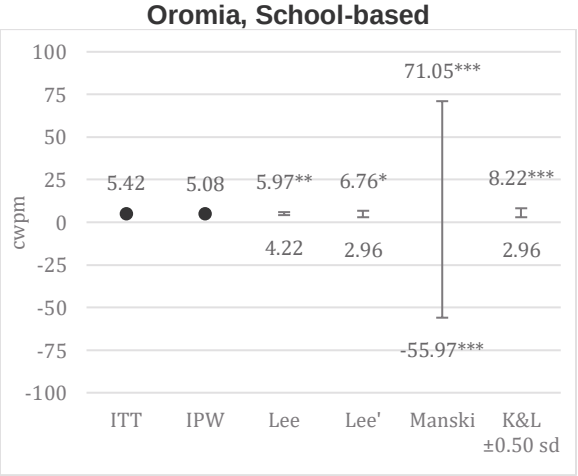
Figure 2. Treatment impacts under different attrition corrections – Oral Reading Fluency (CWPM)

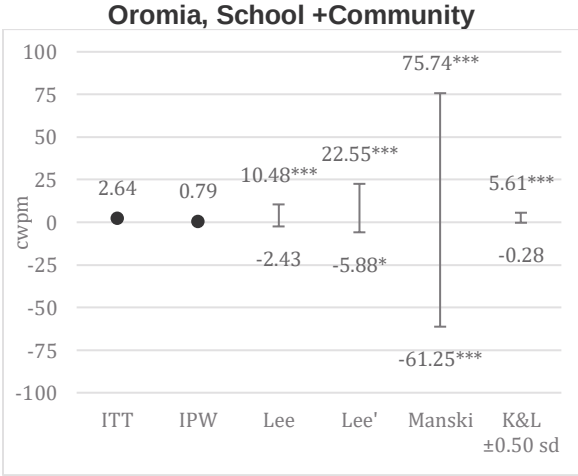


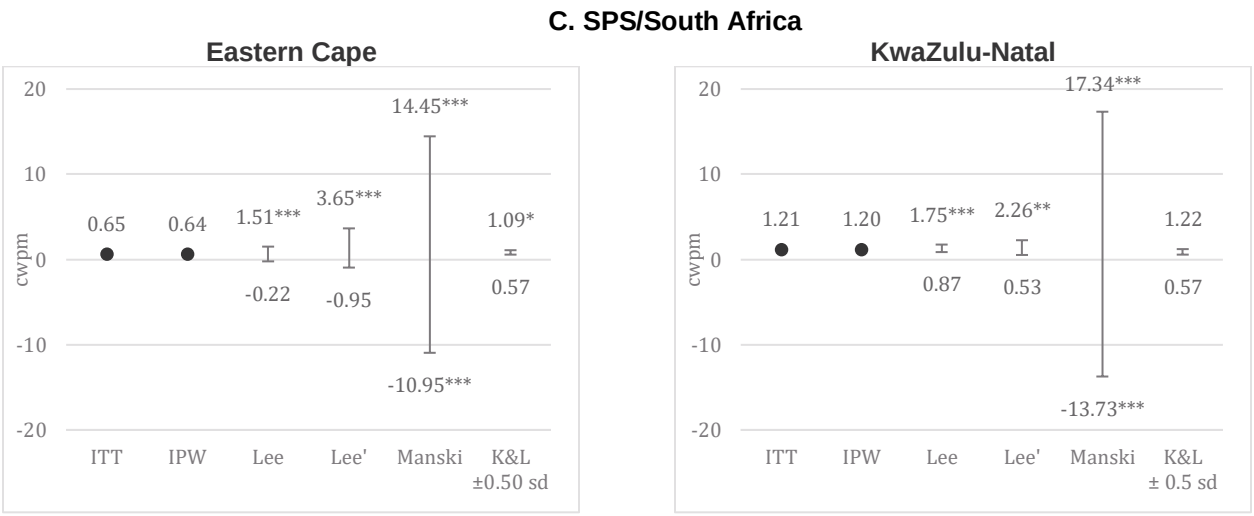
B. READ CO/Ethiopia











Notes: The outcome of interest is the difference between endline and baseline for oral reading fluency. Asterisks denote significance at 0.1% (***), 1% (**) and 5% (*) level.

4.3 Lee bounds and treatment heterogeneity

Lee bounds only inform the impact of the program for those always observed. If treatment heterogeneity is correlated with attrition, Lee bounds could grossly over or underestimate the impact of the program for the initial sample. This is especially important in the case of early reading interventions because, as we show above, attrition is correlated with reading performance at baseline, so it would be valuable to know whether Lee bounds estimates are over or underestimating the impact on the baseline sample. To explore this, we estimate IPW outcome regressions including interaction terms between the treatment dummy and dummies for quartiles of baseline reading skills.[6] Table 3 shows results of this analysis. The left out interaction is the one between treatment status and the lowest quartile, so if programs had a greater impact on students with higher baseline performance, the shown interaction terms should be positive and increase in the quartiles.

For Zambia, the results suggest that the program had a greater impact on high-performing students at baseline, as the interaction terms are more or less increasing in the quartiles, although they are not statistically significant. The patterns are less clear for Ethiopia. For both interventions in Amhara and the School + Community one in Oromia the signs of the interactions are not even the same across quartiles within each experiment. For the school-based intervention in Oromia the results suggest that the low-

[6] We use quartiles of the first principal component of the EGRA subtasks rather than of the oral reading fluency score because large fractions of children cannot read a single word, so quartiles of the oral reading fluency actually do not split the sample in quartiles very well.

performing students benefited more from the program, as all interactions are negative, although all the parameters are quite imprecisely estimated. The results for SPS/South Africa indicate that high-performing students learned to read more words per minute than low- performing students from the program in both Eastern Cape and KwaZulu-Natal (just as in MA/Zambia), as the coefficients on the interactions between treatment and baseline quartiles are positive, increasing in the quartiles in absolute values, and actually statistically significant for the highest quartile.

Table 3. Treatment heterogeneity across baseline reading skills – Oral reading fluency (CWPM)

		READ CO/Ethiopia				SPS/South Africa	
		Amhara		Oromia			
	MA/Zambia	School-based	School + Community	School-based	School + Community	Eastern Cape	KwaZulu-Natal
Treatment	1.78**	6.32	5.25	5.34*	1.29	-0.72	-0.70
	(0.65)	(3.19)	(4.13)	(2.65)	(2.65)	(0.98)	(1.14)
Interactions between treatment status and baseline reading skills quartiles							
Quartile 2	1.08	5.02	0.58	-1.43	0.03	0.48	1.04
	(1.30)	(2.84)	(4.41)	(3.79)	(3.59)	(1.07)	(1.22)
Quartile 3	1.94	-0.63	-4.45	-0.49	2.63	1.50	2.64
	(1.18)	(3.70)	(4.64)	(4.18)	(4.63)	(1.12)	(1.43)
Quartile 4	1.77	-1.25	-3.66	-2.67	-1.00	2.90*	3.80*
	(1.67)	(3.30)	(4.40)	(3.52)	(3.79)	(1.33)	(1.85)
	1961	933	929	681	592	4110	4450

Note: The outcome of interest is the difference between endline and baseline for oral reading fluency. All regressions include dummies for baseline reading skills quartiles. Standard errors are clustered at the community level for SPS/South Africa and at the school level for all other experiments. Asterisks denote significance at 0.1% (***), 1% (**) and 5% (*) level.

The results for MA/Zambia and SPS/South Africa indicate that high-performing students learned to read more words per minute thanks to the program than low performing students. This, combined with the fact that performance at baseline is negatively correlated with attrition, suggests that Lee bounds overstate the impact of the program over the original baseline sample.

5. Conclusions

Survey attrition is inevitable in longitudinal studies. In this paper we present evidence that attrition implications in early reading interventions may be minor, at least according to the IPW and Lee bounds approaches. However, the assumption attached to the former, selection on observables, is strong and may not correctly describe the data generation process. The assumption inherent to Lee bounds, monotonicity, seems weaker but ultimately as untestable as IPW's. Manski bounds impose the least restrictions to the data generation process but fail to provide informative bounds across the analyzed experiments, even when binary outcomes were considered. Manski bounds will probably produce not informative bounds in other applications, unless overall attrition is very small, in which case it is probably inconsequential to begin with. Finally, Manski bounds derivatives, the Kling and Liebman bounds, are restrictive in the sense that they assume that mean outcomes of those that attrit will fall within essentially arbitrary ranges.

Aside from Manski bounds, Lee bounds are probably the most conservative method to address attrition in program evaluations. Therefore, we recommend that researchers consider this approach to bound the treatment effect in the presence of attrition.

However, in addition to assuming monotonicity, Lee bounds only document the impact on those always observed. This is problematic in the context of early reading interventions because the resulting bounds down-weight the impact on low performers, given that, as we showed, reading skills at baseline and attrition are negatively correlated. The extension we propose to Lee bounds documents the impact on the compliers too, but still leaves out of the analysis the impact on the never observed.

Researchers interested in documenting treatment impacts on low-performers should plan for this type of student to be more likely to attrit. To tackle this problem, one strategy is to oversample low performing students at baseline. This may be difficult because in most cases researchers only learn the reading skills of students after baseline data are collected, but information available ex-ante could be used to oversample low performing students, for example if there are test score data on standardized exams, even at the school level, researchers could oversample schools that underperform in these exams. Another strategy is to allocate additional resources to track down and assess at-home low-performers at endline that would otherwise attrite from the sample.[7] In addition, researchers should explore whether there is treatment effect heterogeneity across baseline reading skills (as well as other determinants of attrition). This exercise would inform whether Lee bounds may produce a biased estimate of the impact the program

[7] Molina and Macours (2019) propose methodologies to use intense tracking information to correct for selection bias.

would have on the entire original sample. The results for MA/Zambia and SPS/South Africa suggest that these programs benefited more high-performing students, which combined with the fact that reading skills are negatively correlated with attrition imply that Lee bounds may overestimate the impact the program would have on the original sample.

Finally, Lee bounds will probably fail to be informative when attrition rates are very different between treatment and control, which was not the case in most of the experiments examined in this paper. When treatment status is expected to affect school attendance (e.g., school feeding programs, conditional cash transfers), it would be recommendable that student surveys are conducted not only at school but at students' homes when needed, so differences in attrition rates by treatment status would be lower than if surveys were conducted at the schools.

References

1. Ahrens, Achim, Christian Hansen, and Mark Edwin Schaffer. (2018). LASSOPACK: Stata Module for LASSO, Square-root LASSO, Elastic Net, Ridge, Adaptive LASSO Estimation and Cross-validation.
2. Alderman, Harold and Donald Bundy. (2012). "School Feeding Programs and Development: Are We Framing the Question Correctly?" *The World Bank Research Observer* 27, no. 2: 204–21, https://doi-org.proxy.uchicago.edu/10.1093/wbro/lkr005
3. Aurino, Elisabetta, Aulo Gelli, Clement Adamba, Isaac Osei-Akoto, and Harold Alderman. (2020). "Food for Thought? Experimental Evidence on the Learning Impacts of a Large-Scale School Feeding Program." *J. Human Resources* 1019-10515R1; published ahead of print December 14, 2020, doi:10.3368/jhr.58.3.1019-10515R1
4. Bando, Rosangela, Francisco Gallego, Paul Gertler, and David R. Fonseca. (2017). "Books or Laptops? The Effect of Shifting from Printed to Digital Delivery of Educational Content on Learning." *Economics of Education Review* 61: 162-73.
5. Behrman, Jere R., Susan W. Parker, and Petra E. Todd (2005): "Long-term Impacts of the Oportunidades Conditional Cash Transfer Program on Rural Youth in Mexico." IAI Discussion Papers, no. 122, Georg-August-Universität Göttingen, Ibero-America Institute for Economic Research (IAI), Göttingen.
6. Bundervoet, Tom. (2018). "Internal Migration in Ethiopia: Evidence from a Quantitative and Qualitative Research Study." World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/32097 License: CC BY 3.0 IGO.
7. Cilliers, Jacobus, Brahm Fleisch, Cas Prinsloo, and Stephen Taylor. (2018). "How to Improve Teaching Practice? Experimental Comparison of Centralized Training and In-classroom Coaching." RISE-WP-18/024
8. Drake, Lesley, Meena Fernandes, Elisabetta Aurino, Josephine Kiamba, Boitshepo Giyose, Carmen Burbano, Harold Alderman, Lu Mai, Arlene Mitchell, and Aulo Gelli. (2017). "School Feeding Programs in Middle Childhood and Adolescence." In *Disease Control Priorities* 3, ed. D. Bundy, N. De Silva, S. Horton, D. Jamison, and G.C. Patton. Washington, D.C.: The World Bank. http://dcp-3.org/sites/default/files/chapters/DCP3CAHD_Ch 12.pdf
9. Doyle, Orla, Colm Harmon, James J. Heckman, Caitriona Logue, and Seong Hyeok Moo. (2016). "Early Skill Formation and the Efficiency of Parental Investment: A Randomized Controlled Trial of Home Visiting." *Labour Economics*. http://dx.doi.org/10.1016/j.labeco.2016.11.002.
10. Fiszbein Ariel and Norbert R. Schady. (2009). "Conditional Cash Transfers: Reducing Present and Future Poverty." World Bank Policy Research Report. Washington, DC: World Bank.
11. Fitzgerald, John, Peter Gottschalk, and Robert Moffitt. (1998). "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics." *The Journal of Human Resources* 33, no. 2: 251–99.
12. Ghanem, Dalia, Sarojini Hirshleifer, and Karen Ortiz-Becerra. (2020). Testing Attrition Bias in Field Experiments. Working Papers 202010, University of California at Riverside, Department of Economics.
13. Glewwe, Paul and Karthik Muralidharan. (2016). "Improving School Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." In *Handbook of the Economics of Education* 5, 653–743. Amsterdam: Elsevier.
14. Grant, Monica J. and Jere R. Behrman. (2010). "Gender Gaps in Educational Attainment in Less Developed Countries." *Population and Development Review* 36, no. 1, 71-89.
15. He, Fan, Leigh Linden, and Margaret MacLeod. (2009). "A Better Way to Teach Children to Read? Evidence from a Randomized Controlled Trial." Unpublished manuscript.

16. Heckman, James. J. (1979). "Sample Selection Bias as a Specification Error." *Econometrica* 47, no. 1, 153–61.
17. Horowitz, Joel L. and Charles F. Manski. (2000). "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data." *Journal of the American Statistical Association* 95: 77-84.
18. Huber, Martin and Giovanni Mellace. (2015). "Sharp Bounds on Causal Effects under Sample Selection." *Oxford Bulletin of Economics and Statistics*, 77, 0305–9049. doi: 10.1111/obes.12056
19. Kling, Jeffrey R and Jeffrey B Liebman. (2004). "Experimental Analysis of Neighborhood Effects on Youth." Unpublished manuscript.
20. Kotze, Janeli, Brahm Fleisch, and Stephen Taylor. (2019). "Alternative Forms of Early Grade Instructional Coaching: Emerging Evidence from Field Experiments in South Africa." *International Journal of Educational Development* 66, 203-13.
21. Lee, David S. (2002). "Trimming for Bounds on Treatment Effects with Missing Outcomes." NBER Working Paper, 0277.
22. Lee, David S. (2009). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76, 1071–1102.
23. Molina Millan, Teresa and Karen Macours. (2019). "Attrition in Randomized Control Trials: Using Tracking Information to Correct Bias." Unpublished manuscript.
24. Murnane R. J. and Alejandro J. Ganimian. (2014). "Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations." NBER Working Paper, no. 20284, Inter-American Development Bank, Washington, DC.
25. NORC (2015). USAID/Ethiopia Reading for Ethiopia's Achievement Developed Community Outreach (READ CO) Program, Evaluation Design Report.
26. NORC (2018). USAID/Ethiopia Impact Evaluation of Reading for Ethiopia's Achievement Developed Community Outreach (READ CO) Program. Endline Evaluation Report.
27. Ome, Alejandro and Alicia Menendez. (2020). "Using SMS and Parental Outreach to Improve Early Grade Reading Skills in Zambia." Unpublished Manuscript.
28. Parker, Susan, Luis Rubalcava, and Graciela Teruel. (2007). "Evaluating Conditional Schooling and Health Programs." In *Handbook of Development Economics*, Volume 4, Chapter 62, 3963-4035. Elsevier.
29. RTI International. (2015). Early Grade Reading Assessment (EGRA) Toolkit, Second Edition. Washington, DC: United States Agency for International Development.
30. Tibshirani, Robert. (1996). "Regression Shrinkage and Selection via the LASSO." *Journal of the Royal Statistical Society: Series B (Methodological)* 58, no. 1: 267-88.
31. UNICEF 2019 https://data.unicef.org/topic/child-protection/child-labour/ accessed online on 12/03/20
32. Wooldridge, Jeffrey. M. (2002). *Econometric Analysis of Cross Section and Panel Data*.. Cambridge, MA, and London, The MIT Press.

Annex I. Non-linearities between attrition and baseline characteristics

	MA/Zambia Eastern province	READ CO/Ethiopia				SPS/South Africa	
		Amhara		Oromia			
		School-based	School + Community	School-based	School + Community	Eastern Cape	KwaZulu-Natal
Treatment	-0.01	-0.02	-0.00	0.01	0.11*	0.03	0.01
	(0.02)	(0.04)	(0.05)	(0.05)	(0.05)	(0.02)	(0.02)
Female	0.01	-0.05	-0.08**	-0.07**	-0.02	-0.01	0.01
	(0.01)	(0.03)	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)
Reading score at baseline							
Quartile 2	-0.08**	-0.22***	-0.16***	-0.14**	-0.04	-0.04**	-0.04**
	(0.02)	(0.04)	(0.05)	(0.04)	(0.04)	(0.01)	(0.01)
Quartile 3	-0.02	-0.25***	-0.20***	-0.16**	-0.04	-0.02	-0.05**
	(0.02)	(0.05)	(0.05)	(0.05)	(0.05)	(0.01)	(0.01)
Quartile 4 (highest score)	-0.07***	-0.32***	-0.27***	-0.34***	-0.23***	-0.05**	-0.06**
	(0.02)	(0.05)	(0.05)	(0.05)	(0.05)	(0.02)	(0.02)
Asset index							
Quartile 2	-0.01	-0.05	-0.04	-0.00	-0.03	-0.02	-0.01
	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)
Quartile 3	-0.00	-0.03	-0.07	-0.09	-0.04	-0.00	-0.00
	(0.02)	(0.04)	(0.04)	(0.05)	(0.04)	(0.02)	(0.01)
Quartile 4 (most assets)	0.04	-0.04	-0.06	-0.02	-0.03	0.01	-0.01
	(0.02)	(0.04)	(0.04)	(0.06)	(0.05)	(0.02)	(0.01)
Age groups[(a)]							
Group 1	-0.01	-0.05	-0.03	0.05	0.11**	-0.01	-0.01
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.01)	(0.01)
Group 2	-0.01	-0.03	0.02	0.14**	0.14***	0.02	-0.02
	(0.05)	(0.04)	(0.05)	(0.04)	(0.04)	(0.02)	(0.02)
Group 3 (oldest)	0.02	0.08	0.12**	0.26***	0.31***	0.04*	0.01
	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)	(0.02)	(0.02)
N	2253	1452	1473	1278	1247	4784	5331

[(a)]Age groups are different between experiments. For MA/Zambia the excluded category is 6-year-olds and the shown groups are 7-8-year-olds, 9-10 year olds, and 11-year-olds and older; for READ CO/Ethiopia the excluded category is 8- year-olds and younger, and the shown groups are 9-year-olds, 10-year-olds, and 11-year-olds and older; for SPS/South Africa the excluded category is 7-year-olds and younger, and the shown groups are 8-year-olds, 9-year-olds, and 0-year- olds and older.

Note: Marginal effects evaluated at the mean of the independent variables, after running logistic regressions. Regressions for MA/Zambia and SPS/South Africa include grade fixed effects (READ CO/Ethiopia is just one grade). Standard errors are clustered at the community level for SPS/South Africa and at the school level for all other experiments. Asterisks denote significance at 0.1% (***), 1% (**) and 5% (*) level.

Annex II. Model selection for IPW

Stepwise selection

The stepwise selection procedure follows Doyle et al. (2016). This approach is implemented in three phases: i) Estimate bivariate regressions between attrition and all potential predictors, and retain the ones that are statistically significant (p-value<=0.05); ii). An OLS regression is run using all retained predictors and the corresponding R^2 is recorded, then, to reduce further the number of covariates, multiple regressions are run to eliminate covariates iteratively using the adjusted R^2 as information criterion; iii). A logit model is estimated using the final set of covariates.

Below we list, for each data set, the variables retained in the final model, and the total number of variables considered.

MA/Zambia: Out of 27 variables considered, seven were retained: Two EGRA scores, an indicator for the child being first born, whether the household had planes to move at baseline, variables for household ownership of land plots and small livestock, and a district dummy.

READ CO/Ethiopia: 19 variables were considered across all for experiments. For school-based experiment in Amhara 12 variables were retained: Five EGRA scores and the first principal component of the six EGRA scores, child's age, gender, and interest in reading (self-reported), school attendance the previous week, if the child attended preschool, and household size. For the school + community intervention in Amhara, the same covariates were retained except that instead of school attendance a variable for whether Amharic is the language most spoken at home was retained. For school-based experiment in Oromia, seven variables were retained: One EGRA score and the first principal component of the six EGRA scores, child's age, school attendance the previous week, shift (morning, afternoon, all day), a dummy for whether Afaan Oromo is the language most spoken at home and a dummy variable for whether the child attended preschool. For the school + community intervention in Oromia the variables retained were: One EGRA score and the first principal component of the six EGRA scores, child's age, the treatment dummy and school attendance the previous week.

SPS/South Africa: In the Eastern Cape, eight variables out of 25 were retained - two EGRA scores and the first principal component of the four EGRA scores, a treatment indicator, child's height for age z-score, an indicator that the child co-resides with their mother and two district dummies. In KwaZulu-Natal, five variables out of 23 were retained – baseline oral reading fluency, child's age, child's height for age z-score, an indicator that the child co-resides with their mother and one district dummy.

LASSO

We use LASSO methods (Tibshirani 1996) to select the variables that we include in the attrition models for the IPW estimations. We follow Molina and Macours (2019) and model attrition for treatment and control groups separately, using the bias-corrected Akaike Information Criteria to select penalty levels. Then we estimate an attrition model pooling treatment and control groups, using all the covariates selected in either the treatment or control models, as well as their interactions with the treatment dummy. The estimated probabilities of this pooled model are used to construct the IPW for each experiment. Table B2 shows the IPW results using this method for the attrition model.

Table II1. IPW results using LASSO methods to select predictors

	MA/Zambia	READ CO/Ethiopia				SPS/South Africa	
		Amhara		Oromia			
	Eastern province	School-based	School + Community	School-based	School + Community	Eastern Cape	KwaZulu-Natal
Treatment	3.31***	7.29***	2.54	4.22	1.36	0.61	1.14
	(0.69)	(1.457)	(1.573)	(3.138)	(3.161)	(0.54)	(0.78)
N	1954	933	931	682	594	4109	4450

Note: The outcome of interest is the difference between endline and baseline for oral reading fluency. For predictor selection we use the Stata package LASSOPACK (Ahrens, Hansen, and Schaffer 2018). Standard errors are clustered at the community level for SPS/South Africa and at the school level for all other experiments. Asterisks denote significance at 0.1% (***), 1% (**) and 5% (*) level.

Annex III. Manski bounds for binary variables

Table III1. Manski bounds on the change between baseline and endline in the likelihood that students can read at least one word

		READ CO/Ethiopia				SPS/South Africa	
		Amhara		Oromia			
	MA/Zambia	School-based	School + Community	School-based	School + Community	Eastern Cape	KwaZulu-Natal
lower	-0.07*	-0.28***	-0.31***	-0.39***	-0.55***	-0.17***	-0.17***
	(0.03)	(0.05)	(0.06)	(0.04)	(0.05)	(0.02)	(0.02)
upper	0.19***	0.41***	0.40***	0.51***	0.45***	0.12***	0.16***
	(0.03)	(0.04)	(0.04)	(0.05)	(0.05)	(0.02)	(0.02)
N	2261	1452	1473	1384	1378	4976	5371

Note: The outcome of interest is the difference between endline and baseline for oral reading fluency. All regressions include dummies for baseline reading skills quartiles. Standard errors are clustered at the community level for SPS/South Africa and at the school level for all other experiments. Asterisks denote significance at 0.1% (***), 1% (**) and 5% (*) level.