Mobility's Effect on Girls Education and Empowerment:
Lessons from Zimbabwe

Zachary Robb

An essay submitted to the Department of Economics
in partial fulfillment of the requirements
for the degree of Master of Arts

Queens Economics Department
Queen's University
Kingston Ontario, Canada
September 2021

Acknowledgments

First, I would like to thank my supervisor, Dr. Chris Cotton, for his support and guidance during the development of this essay. I would also like to thank Ardyn Nordstrom, my secondary supervisor, for her insight, guidance, and help navigating the data.

Lastly, I would like to thank the boys, André Labossière, Brian Chow, Deaglan Jakob, Léo Leroux, Joel Jackson, and Trevor "the fish" Knetchel, for your continued moral and emotional support during this crazy year. All omissions and errors are solely their fault.

Abstract

Distance to school is a major barrier to education in the developing world. This barrier can be much more prominent for girls, who often report more substantial safety concerns. This paper uses primary data from World Visions IGATE-T program and propensity score matching to determine the effect that distributing bicycles had on girls' feelings of empowerment, test scores, and school attendance. Overall bicycles led to an increase in literacy and numeracy test scores by .23 and .28 standard deviations respectively. There was no clear effect on attendance or empowerment. I also find evidence that bicycles increased the number of days the student reported being late; which may be explained by an increase in attendance.

Contents

Introduction

Improving girls' access to education, with the goal of attaining gender equality, is a crucial component in both economic development and achieving the United Nations Millennium Development goals. While many programs have focused on increasing the supply of education or providing incentives to parents to send their daughters to school, relatively few have focussed on a significant barrier to education, distance. Distance is particularly pertinent for girls, many of whom potentially face more dangers and safety concerns on commutes. The World Bicycle Relief's (WBR) Bicycles to Empower and Educate program (BEEP) aims to mitigate this barrier by providing bicycles to improve equitable access to education. This paper analyses BEEP's impact on Zimbabwean girls' educational outcomes, feelings of empowerment, and attendance rates.

Similar to Benhassine et al. (2015) and Muralidharan and Prakesh (2017), this paper focuses on assessing the impact of a labelled kind transfer (LKT). LKT's differ from conditional cash transfers in two ways. They do not provide cash to recipients and instead provide a good (in this case, a bicycle). They are also not conditional on behaviour and are merely suggested to be used for certain activities. In BEEP, the recipients are given the bike and told to use it mainly for school (Fiala, 2017).

A previous analysis of a BEEP program used a clustered randomized control trial to estimate bicycles' impact on girls' education in Zambia (see Fiala, 2017). My analysis differs in a few major ways. First, the participants in the Zambian study were restricted to those who were in grades 5 through 7 and walked at least 3 kilometres to school. My analysis contains girls from grades 3 through 8 and includes individuals who live at least a 15-minute walking commute to school. Therefore, it expands on the Zambian analysis by providing more insights on both younger and older students, and, more importantly, for a much wider range of distance from school, which will allow me to asses how the programs' impact depends on distance. Second, because BEEP was launched in conjunction with World Vision Zimbabwe's (WVZ) IGATE-T program, the BEEP program was implemented in a subset of communities that were already undertaking efforts to support girls' education and empowerment through the IGATE-T program. Therefore, my estimates reflect the marginal impact of including a bicycle component on an otherwise more typical girls' education project. This study also utilizes IGATE-T's evaluation data, which is extremely rich and contains thousands of variables, including information on test scores, school attendance, leadership skills, transition outcomes, and various household characteristics. Finally, BEEP was implemented during a chaotic period in which Zimbabwe faced a cyclone, a severe drought, a global pandemic, teacher strikes and various economic crises. Thus, any results demonstrate the program's resilience during uncertain times.

Since BEEP was implemented within the IGATE-T treatment group, no proper control group was formed. To overcome this, I use propensity score matching to identify a suitable counterfactual. This method ensures that the treatment and control group have similar characteristics at the baseline, mimicking the features of a randomized control trial. This paper uses optimal full matching, hereafter referred to as full matching, due to the small sample size and the failure of traditional matching techniques to produce a balanced sample. Rosenbaum and Rubin (1983) have shown that under certain conditions, conditioning on the propensity score can allow one to obtain unbiased estimates of the average treatment effect.

Using a difference-in-differences model, I analyze the effect of treatment on literary and numeracy scores, attendance rates, and feelings of empowerment. Overall, the results indicate that the BEEP significantly affected test scores but had no clear impact on Youth Leadership Index scores and school attendance rates. The treatment group scored on average .23 standard deviations higher in literacy and .28 standard deviations higher in numeracy. This paper provides the first evidence that decreasing the distance cost through the provision of bicycles can lead to improved test scores. There is also evidence that bicycle's led to an increase in the number of days the student reported being late; but this may be explained by a potential decrease in days of school missed.

Furthermore, I perform subgroup analysis examining the impact that receiving bicycles had on those who face heavy chore burdens, have disabilities, and live far from school. These analyses severely restrict the sample size meaning that any subgroup analysis results are primarily suggestive. I find, unsurprisingly that the effect of receiving bicycles is significantly higher for those who face longer commute times.

In a similar fashion to Muralidharan and Prakesh (2017), I analyze the effect that bicycles have relative to the distance to school, or in my case, commute times. Muralidharan and Prakesh found that bicycles had an inverted-U effect on enrollment, meaning that at close distances, the bicycles had little to no effect, medium distances had a large effect, and at far distances had a very low effect. I fail to find a similar result for attendance rates, with bicycles having the largest significant impact for those who face a commute of 46 minutes to an hour, with no significant effects for the others. However, this may be a power issue as the sample size for each group is relatively small.

The rest of the paper is as follows. Section 2 examines the current literature on reducing barriers to education; section 3 describes the context and the program; section 4 describes the data, propensity score methods, and the creation of the control group; section 5 presents the main results and sub-group analysis; section 6 concludes.

Literature Review

Programs aimed at reducing barriers to education have typically taken one of two approaches: to increase the benefits of attending school or reduce the costs associated with attending school. The former typically takes the form of conditional cash transfers (CCTs), while the latter focuses on subsidies and increasing the number of schools in each area. The review below examines both methods and reviews studies examining similar interventions. I direct the reader to Kremer and Miguel (2008) for a complete review of the literature.

Increasing Benefits to Education

Traditionally policies to improve educational outcomes in both girls and boys have centred around subsidies to education. One of the largest and most notable of these programs is Mexico's Progressa. Progressa, the seminal conditional cash transfer program, provided monthly grants to mothers whose children maintained an attendance rate of 85%. Studies on this program found that Progressa increased enrollment by 3.4-3.6 percentage points for all children in grades 1-8 and increased older girls' enrollment by 14.8 percentage points, which is significantly higher than boys' enrollment which grew by 6.5 percentage points (Schultz, 2004). While the benefits that Progressa and other CCT programs have had cannot be denied, in terms of cost-effectiveness, they are not a very efficient way to improve girl's education. This is no doubt because these programs aim to provide income support for the poor and are not focused solely on girls' education.

To address this issue of the cost-effectiveness of CCTs, Benhassine et al. (2015) examined the effect that removing the conditionality had on school attendance. Using what the authors called a labelled cash transfer LCT, they compared the differences this had on school attendance rates to a traditional CCT. The results indicate that an LCT is just as effective as a CCT in improving attendance rates and potentially more effective at improving numeracy scores.

Decreasing costs of Education

While the demand side approach typically involves increasing the benefits of attending school, the supply side primarily centres around improving access to education through the creation of new schools, thereby reducing the distance cost of attending school. Several studies have studied the effect of school creation on educational outcomes. In a 2001 study, Esther Duflo (2001) studied the impact of the Indonesian government's massive school construction program. The program, which began in 1973, involved the creation of over 61,000 primary schools throughout the country. Duflo used variation in the timing and number of schools created to estimate primary school construction's effect on education and earnings. She found that children aged 2 to 6 in 1974 received .012 to 0.19 more years of education and estimated the increase in wages for these individuals to be from 6.8 to 10.6 per cent.

A randomized control trial conducted by Burde and Linden (2013) examined the effect of placing schools in Afghanistan villages. They found that introducing a village-based school program led to an increase in girls' enrollment by 52 percentage points and an increase in their average test scores by 0.65 standard deviations. The impact was large enough to effectively eliminate the gender gap in enrollment and dramatically reduce test scores' gender disparity.

Health based Interventions

Several health-based interventions have been implemented to increase educational outcomes and enrollment rates in the developing world. For example, parasitic worm infections are quite common in school-aged children in developing countries. Kremer and Miguel (2004) find that mass deworming exercises in Kenya reduced absence rates by 7 percentage points. It also led to positive health and education spillover. Those not treated in schools that received treatment were 8 percentage points more likely to participate in school than children in the control group, likely resulting from reduced transmission within schools.

Reducing Distance Costs

While interventions such as mass school creation and conditional cash transfer programs have been proven to be effective ways at increasing overall education outcomes from a cost standpoint, they are not the most desirable. Government budgets in developing countries are limited, and thus, cheaper alternatives for increasing female education outcomes must be developed. One of these more affordable alternatives may be the distribution of bicycles.

A recent study conducted by Muralidharan and Prakash (2017) analyzed the impact that the Indian state of Bihar's Bicycle Program had on female enrollment rates. The Bicycle Program provided bikes for all girls enrolled in grade 9 to enable them to get to school more easily. Using data from the Indian District Level Health Survey combined with school-level secondary school enrollment data and official data on the number of students who took and passed the secondary school certificate examination. Employing a triple difference-in-differences model, they use 16–17-year-old girls in the same state as the initial control group, then the boys as the second control group to mitigate any confounding effects by any changes that took place in Bihar during the same period. Overall, they found that the cycle program led to a 5.2 percentage point increase in the probability that girls aged 14-15 are enrolled in or have already completed grade 9. When examining the results by distance to school, they find that the results have an inverted-U shape, meaning that the program had small effects on those very close to school and those very far from school, with the largest gains being those a medium distance from the school. As they point out in the paper, this makes intuitive sense; those that are very far from the school face a large distance cost and bicycles likely do not make up for that cost. To account for distance, they use a quadruple difference-in-differences model using the same model as before but adding a dummy for long-distance (which equals 1 if the participant lives more than 3 km from school). Unsurprisingly, the results yield a higher increase in probability, with girls living more than 3 km away being 8.7 percentage points more likely to be enrolled in or have completed grade 9. When analyzing the program's impact on education outcomes, they found that the program led to an increase in the number of girls who took the secondary school certificate exam by 18.4 percentage points and an increase in the number of candidates who passed the exam by 12.2 percentage points. Analyzing the program from a cost-effectiveness perspective, they find that it ranks much better than CCT's based in the same region, with the CCT's costing $3 per month per recipient and increasing enrollment by 4 percentage points. The cycle program cost on average $1 per month and led to an increase of 5.6 percentage points. While the paper did cover female empowerment, it was solely done through qualitative analysis with no direct causal effect being calculated.

An analysis of another BEEP program in Zambia conducted by Fiala (2017) used a clustered randomized control trial to determine the impact of bicycle access on girls' educational and empowerment outcomes. The study split 100 schools into three control groups. The first control group received the standard bicycle program where all eligible girls were offered a bicycle on the condition that the bicycle is used primarily to travel to school. A field mechanic was also trained for each school to provide repairs and maintenance for a fee. Each student was also required to pay a start-up fee of around 5 USD. The second group received a bicycle and were not required to pay the start-up fee for spare parts. Then a comparison group that received no bicycles. The study found that giving girls access to bicycles reduced their commute time by around 35 minutes each way and decreased the number of days participants were late to school by 1.43. They also found that levels of empowerment (score on the locus of control index) were significantly larger (.66 compared to .5). The researchers did not find any statistically significant impact on grade transition, dropout rates or test scores.

Program Description and Context

Zimbabwe

BEEP was implemented over a chaotic period in Zimbabwe. The project began in 2017, the same year that Robert Mugabe resigned after a military takeover. The new president Mnangagwa was elected in 2018 and promised to stabilize the economy and increase foreign investment. The government implemented several fiscally conservative measures, including a severe decrease in spending and a 130% increase in fuel prices.

Before the COVID-19 pandemic, Zimbabwe was already in a deep recession. Cyclone Idai, combined with a severe drought, particularly affected the agriculture, water and electricity sectors causing a ripple effect for other sectors. Combined with the central government's tight control of public finances, these factors led to a large amount of inflation, with the local currency depreciating more than 70 percent against the US dollar. By the end of 2019, extreme poverty had risen to 42 per cent, compared to 30 per cent in 2017 (World Bank, 2021). This rate continued to grow during the pandemic and estimates by The World Bank (2021) say that the number of extreme poor has reached 7.9 million, or around 49% of the total population in 2021.

In response to COVID-19, the Zimbabwean government closed schools in March of 2020 and implemented a strict lockdown restricting movement and business activities. While online learning was available, most households did not have adequate internet access and could not participate. Schools reopened in September of 2020, but teachers were absent. After reopening, teachers went on strike, citing severely depreciated wages and a lack of PPE. After two months of strikes and negotiations, the government agreed to increase the teachers' salaries by 41%, ending the strike. Schools have been operating with enhanced COVID-19 measures ever since.

IGATE-T: BEEP

IGATE-T is a World Vision project designed to empower Zimbabwe's most vulnerable youth by increasing education quality and improving attitudes towards education. It was part of the UK government's Girls' Education Challenge working to improve education outcomes of marginalized girls around the world. It involves a series of interventions designed to help the most marginalized. The interventions include the formation of leadership clubs, teacher development programs, community-based education programs, and BEEP.

Since November 2017, the WBR has been working with World Vision Zimbabwe (WVZ) to distribute 7,400 bicycles to schools across Zimbabwe. Bicycles were distributed randomly amongst the girls. The selection of the BEEP treatment group was conducted at the midline. The IGATE-T enumerators asked the participants:

> "Have you received a Bicycle from IGATE in the last two years?"

If the girl answered yes to this question and attended one of the IGATE-T treatment schools, they were added to the treatment group.

Since the BEEP participants also were also members of the IGATE-T treatment group it is impossible to completely separate the effect of the two programs. The analysis in this paper show the effect that bicycles have when combined with other programs designed to increase girls' education and empowerment.

Data, Propensity Score Matching, and Estimating Equations

Data Overview

The data used is from World Vision's IGATE-T evaluation. The IGATE-T data was collected in three periods, the baseline in 2017, the midline in 2019, and the endline in 2021. The data contains 1700 individuals and over 4000 covariates. Since the BEEP program operated in conjunction with IGATE, BEEP participants also received IGATE-T treatment. In order to isolate the effect that bicycles had from that of other IGATE-T's interventions, the control population is selected from IGATE-T's treatment group. This ensures that any difference in outcomes is attributable to receiving a BEEP bicycle and not a difference in any other IGATE-T interventions received. This leaves us with a treatment group that contains 195 individuals spread out across grades 1-8. After filtering out those in the IGATE-T control group, 531 individuals remained, including the 195 that received bikes and 336 that did not. The enumerators also interviewed the girls' teachers, headteachers, and heads of their respective households.

There are three main focus areas of this analysis, test scores, attendance, and reported feelings of empowerment. The test scores came from learning assessments administered to students by enumerators and include basic mathematics and literary questions designed to assess the learner's comprehension of the subject. Unfortunately, due to the COVID-19 pandemic, endline data was not collected for test scores; thus, only midline effects are available. When calculating attendance rates, I use the student surveys conducted by the IGATE-T enumerators who asked the students:

"In the last 20 days, how many days of school did you miss?"

The empowerment scores come from the CARE's Youth Leadership Index (YLI). The YLI was specifically designed to measure changes in self-perceptions of leadership among youth, specifically those aged 10-17 (CARE, 2014). The survey involves several statements and asks the participant to choose how often the statement is true for them. The scores are calculated by adding up the responses and the numerical values assigned to them. As is common practice, all test scores are standardized and all effects reported are in standard deviations.

Propensity Score Matching

To determine the causal effect of an intervention, it is imperative to compare treated outcomes to a counterfactual (Rubin, 1974). The counterfactual is defined as what would have happened to the treatment group in the absence of treatment. Since it is impossible to both apply and withhold treatment to an individual, researchers rely on creating a control group. The control group should display similar characteristics to the treatment group, with the only difference being that the treatment group received treatment whilst the control did not. A control group is ideally created before an experiment where the researchers could randomize treatment over a population to ensure that any change in outcome could be attributable to the treatment itself and not some unmeasured variable. While many studies in economics utilize the randomized control trial, there are many cases where it is inappropriate or impossible to randomize access to treatment. In these cases, researchers must use different quantitative methods to determine the causal effect. Propensity score matching is one such method. Developed by Rosenbaum (Rosenbaum & Rubin, 1983), propensity score matching involves matching individuals based on the probability of receiving treatment conditional on their baseline characteristics. Propensity score matching simplifies traditional multivariate matching, which matches individuals based on several different variables, leaving it up to the researcher to specify the most important characteristics to match on. Matching on the propensity score collapses all the characteristics of an individual into one number reducing both the computational costs and the potential researcher bias.

The propensity score acts as a sort of balancing score: conditional on the propensity score, the distribution of the baseline characteristics should be similar between the treatment and control groups. Rosenbaum and Rubin (1983) set out two conditions that must be met for propensity score matching to produce unbiased estimates. First, the treatment assignment must be independent of the potential outcomes conditional on the baseline characteristics, meaning that there must be no measured confounders. The second states that every subject must have a non-zero chance of receiving treatment.

The propensity score is typically estimated by logistic regression but can be estimated by boosting, neural networks or random forests (Lee et al., 2010; Setoguchi et al., 2008). Once the propensity score has been estimated, matches are formed of treated and untreated subjects who share a similar propensity score. Currently, no restrictions exist on the appropriate level of distance between matched sets. When examining the optimal calliper width, Austin (2011) found that a calliper width of .2 standard deviation of the logit of the propensity score minimized the mean squared error of the estimated treatment effect in several scenarios.

Propensity score matching has many desirable attributes. Unlike other quasi-experimental techniques, it ensures that characteristics are similarly distributed. It also separates the analysis from the creation of the control group. Rubin (2001) mentions that when conducting regression analysis, the temptation to work towards the desired outcome is always present. Propensity score matching prevents this by separating the design and analysis sections of a study.

Matching Identification Equations

Variable selection for the propensity score model is quite important. Unfortunately, there is a lack of consensus in the literature as to which method is best. Austin (2011), lays out four different sets of variables that could be included in the propensity score model: all measured baseline covariates, all baseline covariates that affect treatment assignment, all baseline covariates that affect the outcome, and all covariates that affect both the outcome and treatment assignment. The most common of these approaches is only including the variables that affect treatment assignment. This makes intuitive sense as the propensity score is defined as the probability of receiving treatment assignments.

Austin (2007) examined the benefits of including different sets of baseline characteristics. They found that including all measured baseline covariates and all covariates that affect treatment to balance characteristics slightly better than solely including covariates that affect treatment. Including these covariates also did not induce any increase in bias leading to more precise estimates of the treatment effect. In creating the estimation equation, I follow the advice of Austin and include all potential confounders. Following Rosenbaum (1984), I convert NA responses into dummy variables to include in the estimating equation. This ensures that the distribution of NA answers is similar between both groups. When the districts are included in the estimating equation, the model performs worse. Thus, I do not include it in the final estimation equation and account for this by using regression adjustment (Nguyen et al., 2017). The final estimating equation for the propensity score is:

$$Treatment_i = \beta_0 + \hat{\beta}_i \hat{X}_i + \epsilon_i$$

Where treatment is a binary variable indicating whether the individual received a BEEP bicycle, and X_i is a vector of variables including, commute times, grade, the head of the household's gender, disability status etc. To provide some robustness checks and analyse the impact that different matching methods have on control group selection, I calculate the propensity score, create control groups using traditional 1:1 matching and full matching.

Matching Techniques

The most common approach to propensity score matching is 1:1 matching, where a single treated subject is matching with a single untreated subject with a similar propensity score. This has several benefits, the most notable of which is its simplicity. Researchers have shown that this method has several issues. First, it reduces the overall sample size since matches are formed 1:1 if the number of untreated subjects is larger. This results in some untreated subjects being excluded from the study, potentially reducing the variance and imposing some bias. Second, it may result in some poor matches as some treated individuals may not be similar (in terms of propensity score) to their matches or vice versa. One can overcome this by specifying the calliper for the distance between propensity scores, but this may reduce the researcher's ability to interpret the effect, as it may no longer reflect the impact for all treated individuals (Rosenbaum & Rubin, 1985).

Similar to 1:1, there is k:1 matching which involves creating matched groups with k untreated and 1 treated subject. Studies have shown this method to reduce bias compared to 1:1 matching, but it still runs into the issue of discarding observations and potentially poor matches.

Ming and Rosenbaum (2000, March) developed many-to-one matching; wherein M untreated subjects are matched to 1 treated subject. They proved that when allowing this M to vary by group, bias was greatly reduced when compared to traditional 1:1 matching. Full matching (Rosenbaum, 1991; Gu and Rosenbaum, 1993) takes this one step further by allowing the number of untreated and the number of treated to vary for each set.

Full matching works by forming a series of matched sets in which each group has at least one treated and multiple controls, or at least one control and multiple treated. Full matching optimally creates these sets, meaning that treated individuals who have many similar untreated individuals will be matched with more untreated individuals, whereas treated individuals who are not similar to many other untreated will be matched with relatively fewer. This provides full matching with much more flexibility than standard k:1 matching, which requires each set to have the same number of controls regardless of how well they fit. Full matching has two attractive features compared to other approaches. First, it uses all the subjects in the original sample. This differs from other matching techniques where subjects can be excluded from the final sample Rosenbaum (1983). This may impose a bias/variance trade-off on the researcher, who must balance reducing the bias due to selection of the most similar individuals but increasing the variance due to fewer observations. Full matching circumvents this trade-off by using all available observations and creating more optimal matches. Secondly, full matching allows the estimation of the average treatment effect or the average treatment effect of the treated. Traditional pair matching only allows for the former.

In an observational study, Hansen (2004) used full matching to create a control group to examine the effect of coaching on SAT scores. Before matching, the control and treatment groups were separated on the propensity score by 1.1 standard deviations. Full matching reduced this to .01-.02 standard deviations.

Assessing Balance

The propensity score is a balancing score. Thus, after matching, the distribution of baseline covariates, conditional on the propensity score, should be the same if the identification equation is adequately specified. If, however, baseline characteristics are not balanced after conditioning on the propensity score, this may be evidence of misspecification of the identification equation. Assessing balance also ensures that your treatment and control group are identical at baseline, allowing you to proceed as if it were a randomized control trial.

In order to compare the baseline characteristics of control and treatment groups, Austin (2009) suggests that you begin with a comparison of means and medians for continuous covariates. The standardized differences compare the difference in means in units of the pooled standard deviation. Statistical tests, such as t-tests, must not be used to assess balance as they are confounded by sample size. The matched sample is smaller than the original, and thus when relying on statistical tests, the results may be misleading (Imai et al., 2008).

Comparing the standardized mean difference is a good start as the propensity score should, in theory, balance the distribution of the covariates. While there is no universally accepted threshold of what standardized mean difference indicates imbalance, Normand et al. 2001, April demonstrate that an absolute standardized difference less than 0.1 means the difference between covariates is negligible. In this analysis, since there are several variables slightly above .1 I use a threshold of .15 to indicate a significant difference between groups (see Stuart and Green, 2008).

Several authors have also used graphical methods such as boxplots cumulative distribution functions and empirical nonparametric density plots to compare the distribution of baseline covariates between treatment and control groups (Austin, 2009). Ho et al. (2007) suggest that comparing the estimated propensity score between treatment and control may be a way to complement standardized mean comparisons. Figures 1 and 2 contain a graphical comparison of the propensity score distribution for both one-to-one matching and optimal full matching.

Table 1 has the means of the baseline characteristic by group. As you can see, while some of the characteristics are relatively balanced before matching, the distance and grade indicators are quite imbalanced. Table 2 contains the standardized mean differences for one-to-one matching, and Table 3 shows the standardized mean differences for full matching.

As you can see one-to-one matching did not do well at balancing some of the baseline covariates there are many above the .15 threshold. There are two reasons why this could be. First, the identification equation may be misspecified, meaning that there may be unobserved characteristics that influence the likelihood of receiving treatment. The second reason that one-to-one matching does not balance covariates is the lack of similar individuals. One-to-one matching without calliper restriction matches treated subjects with the untreated with the closest propensity score; thus, if the nearest untreated individual is significantly different than the treated, this may result in a poor match. Due to the relatively small sample size of the study, I suspect the latter to be the case.

Looking at the full matching control group, I find that full matching better balances baseline covariates. There are some imbalances in the covariates. The indicator for when the head of the household is male has a standardized mean difference of .55, which is significantly over the .15 threshold. The indicators for Insiza and Mberengwa (two districts in Zimbabwe) are also above this threshold at .25 and .68, respectively. To account for this difference when estimating treatment effect, I follow Ho et al., (2007) advice and use regression adjustment with the imbalanced indicators as control variables.

Estimating Equations

As mentioned before, there are three main outcomes of interest, literary and numeracy scores, school attendance rates, and feelings of empowerment (YLI scores). To analyze the change of the outcomes of interest, I first calculate the difference between endline and baseline, endline and midline, and midline and baseline, then regress those differences on treatment and other remaining imbalances. Since full matching creates weights in order to create better matches all of the regression using the full matching sample are done with weighted ordinary least squares. The weights assigned to treated subjects are 1, while the weights of the control subjects vary to match the treated better. The estimating equation is as follows:

$$D_{i_t} = \beta_0 + \beta_{1_i} T_i + \hat{\beta}_{2_i} \hat{X}_i + \epsilon_i$$

Where D_{it} is the difference in the outcomes for individual i at either baseline, midline or endline. β_{1_i} is the treatment variable indicating whether the individual received a BEEP bicycle, and $\hat{X}_i$ is a vector of control variables pertaining to the remaining imbalances between treatment and control groups. Following Abadie and Speiss (2021), and Austin and Small (2014, October) all standard errors are cluster robust, being clustered at the matched sets level. The regressions are conducted in R using the 'estimatr' package which allows the user to specify the weights of the sample, and the type of standard error. In this case, I use the 'stata' standard error which creates heteroskedastic-consistent variance estimates for the clustered case. It also contains a special finite sample correction.

Results

Overall Results

Table 4 shows the results of the regression with days of school missed as the outcome of interest. Since this is a difference-in-differences model, a negative coefficient means a decrease in reported days missed. Columns 1-2 display the difference-in-differences model from baseline to midline, columns 3-4 show the results for midline to endline, and 5-6 show the treatment's overall effect from baseline to endline. The baseline to midline results shows a negative but statistically insignificant effect on days of school missed. The treatment effect from midline to endline is .85, meaning that receiving a bicycle led to, on average, .85 more missed days. This effect is statistically significant at the 10% level but becomes insignificant after controlling for remaining imbalances between the treatment and control group. The overall effect of treatment (baseline to endline), show a positive but statistically insignificant result meaning that overall, receiving a bicycle led to no change in days of school missed.

Table 5 shows the effect of receiving a bicycle on girls' literacy and numeracy scores. Due to the COVID-19 pandemic, WVZ was unable to collect test scores for the endline; therefore, this analysis is conducted only on the baseline to midline period. As is normal when reporting test scores all of the scores have been standardized, thus the coefficients represent changes in standard deviation. Starting with numeracy, we see that without accounting for the remaining imbalance in the control and treatment group, there is a treatment coefficient of .32, which is statistically insignificant. After controlling for the differences between the treatment and control group, the magnitude of this coefficient drops to .28 but becomes significant at the 5% level. A similar effect is found for literacy scores; when not accounting for the imbalance, a positive but insignificant result of .28 is found. After adjusting for the imbalance, there is a slightly smaller but statistically significant effect of .23.

The following table, Table 6, shows the effect of treatment on YLI scores. After adding the control variables, there is a negative but statistically insignificant result for all periods.

The final table (7) shows the effect that treatment had on the number of days a student reported being late to school. The treatment coefficient after adding control variables is 1.59 and is significant at the 1% level, meaning that receiving a bicycle increased the number of late days by 1.59. This is in stark contrast to Fiala (2017), whose analysis on a BEEP program in Zambia found a reduction in reported late days. This result is discussed further in the distance analysis.

Disability Subgroup Analysis

The following analyses involve those who reported having a disability at baseline, reducing the sample size of the analyses to around 30 individuals. A quick power test indicates that in order to see statistically significant results, we would need each group to contain around 182 individuals. Thus, while this group's results cannot be used as concrete evidence of the treatment effect, they may provide insights into how treatment affect those with disabilities.

Table 8 shows the difference-in-differences model's results for those who reported having a disability on the number of school days missed. The results show a much more extreme treatment effect than those found when analyzing the total sample. This makes intuitive sense as those with disabilities face larger barriers to attending school, and thus, treatment would likely have a larger effect. Similar to the overall analysis, I find a negative treatment effect (a decrease in days missed) when analyzing from baseline to midline. The coefficients for treatment indicate a positive but statistically insignificant treatment effect. When looking at the effect from midline to endline, I find that treatment positively affects days missed. The treatment coefficient of 1.51 is statistically significant but loses significance after imbalances are accounted for. The overall effect from baseline to endline is statistically insignificant, indicating that treatment did not affect the number of school days missed.

The effect of treatment on literacy and numeracy for those who reported having a disability is reported in Table 9. The treatment coefficient for numeracy scores is insignificant both before and after, accounting for the imbalance between groups. The effect on literacy scores (.46) is significant without control variables but becomes negative and insignificant after controlling for imbalances.

The change in YLI scores for those who reported having a disability is reported in table 10. There is a positive impact (.85) on YLI scores from baseline to midline, but this becomes statistically insignificant when added controls. From midline to endline, there is a negative and statistically significant impact on YLI scores. The coefficient -.84 is significant at the 5% level.

Heavy Chore Burden Analysis

The following subgroup analyzed is those who face a heavy chore burden, which I define as spending more than 4 hours of their day doing chores. Household chores are a significant barrier to education, and girls who must complete hours of chores may not attend school if the commute takes up a large portion of their day.

Table 11 shows the effect on days of school missed for those with a heavy chore burden. From baseline to midline, the treatment led to an increase in the number of days of school missed by .71, which is significant at the 10% level. I find no discernable treatment effect when looking at the effects from midline to endline and from baseline to endline periods.

I find smaller treatment effects for literacy and numeracy scores than those found when analyzing the total sample. Table 12 displays these results. The effect on numeracy with and without controls is .09 and .19, respectively. Again, these coefficients are statistically insignificant, and thus, the treatment appears to have no effect. The same goes for literacy scores, where the treatment effect is .09 and .01. Again, these results are statistically insignificant, and thus, the treatment effect is statistically no different than zero.

As for YLI scores, Table 13 shows that there are no discernable effects, with all coefficients being statistically insignificant.

Far and Distance Analysis

The final group I am analyzing is those who reported having a commute time of one hour or longer. Intuitively one would expect the treatment to have the largest effect on these individuals as they face the highest distance cost to education.

Starting with attendance, we see a similar effect to that of the overall sample (see Table 14). The baseline to midline results shows a .42 decrease in days missed, midline to endline shows an increase of 1.12, and endline to baseline displays an increase of .61; all these results are statistically insignificant when control variables are added.

Looking at the effect of treatment on literacy and numeracy scores of those with a longer commute time, we see a significant increase in both literacy and numeracy scores. Table 15 shows that treatment led to an average increase of .57 standard deviations without controls, which is significant at the 5% level. With controls, the magnitude of this effect drops to .54 but remains statistically significant at the 5% level. As for literacy scores, we see an increase of .43 without controls, which is statistically insignificant. But after accounting for imbalances between the groups, this effect becomes .37, and is statistically significant at the 5% level.

As for YLI scores, Table 16 shows that the treatment effect from baseline to midline is insignificant. The midline to endline results show a large negative coefficient but are statistically insignificant when controls are added.

Further Distance Analysis

Muralidharan and Prakesh's (2017) study on the Indian bicycle program found that bicycles' effect on enrollment conditional on the distance to school showed an inverted-U effect. Those who lived both very close and very far saw little to no impact, while those in medium distances accounted for most of the increase in enrollment. The IGATE-T data has reported commute times and I can therefore perform a similar analysis using the reported commute time. Table 17 shows the effect of treatment on attendance by each of the reported commute time bins. Due to sample size issues, I present the results without controls. The results indicate that there is no inverted-U effect when looking at attendance rates. The largest effect (-2.51) is for the group who faces a commute of 16-30 minutes but is statistically insignificant with a p-value of .1233. Those who have a reported commute time of 31-45 minutes see a positive but statistically insignificant coefficient of .3. The groups who face a commute time of 46 minutes to an hour and greater than 3 hours see negative and significant results of -.89 and -1.11, respectively. However, the results for the latter are primarily suggestive as the sample size is 6.

Table 18 shows the effect of treatment on the reported number of reported late days by each distance subgroup. The only significant effect is for those who face a commute time of 16-30 minutes. The coefficient of 2.14 is statistically significant at the 10% level after controlling for the remaining sample imbalances. As mentioned in the paragraph above, the change in days of school missed for the same group is -2.51; thus, it is likely that an increased number of reported late days is due to a decrease in days of school missed. In other words, these girls are now showing up late rather than not showing up at all.

Conclusion

Distance to school is a significant barrier to receiving education for children in developing countries. The Bicycles for Education and Empowerment program aims to mitigate this barrier by providing bicycles for those who most need them. Two major studies have analyzed the impact of bicycle programs on girls' education. This paper adds to the literature by finding evidence of bicycles' positive impact on test scores. Secondly, unlike Muralidharan and Prakesh (2017) and Fiala (2017), my sample contains girls in grades 3-8 demonstrating the programs ability to impact the outcomes of elementary students and secondary students.

In this paper, I use primary data collected over 5 years after the distribution of bicycles and use propensity score matching to create a counterfactual. Using a regression adjustment, I estimate the impact of BEEP on girls' test scores, attendance rates and YLI scores. The literacy and numeracy scores for those who received a BEEP bicycle were .28 and .23 standard deviations higher than those who did not receive a bicycle. There is also evidence that this was driven by those who face commute times of over an hour. I fail to find any overall impact on attendance rate and YLI scores after accounting for differences in the control and treatment groups. Unlike Fiala (2017), there is a positive treatment effect on the number of reported late days, with those in the treatment group reporting on average 1.59 more late than those in the control group. However, this may be partially explained by increases in attendance.

When conducting further subgroup analysis, I fail to find any significant impacts on test scores, attendance, YLI scores or lateness from baseline to endline (although this is likely due to a lack of power). The one exception being the group with a commute time greater than one hour, which saw significantly higher test scores than the control group at .54 and .37.

While Muralidharan and Prakesh's (2017) study on the Indian state of Bihar's bicycle program found an inverted-U effect on enrollment by distance, I fail to find such effects for attendance. With the only significant effect being for those who live 46 minutes to an hour away. When conducting the same analysis on reported number of late days the group who faces a 16-30 minute commute saw a statistically significant increase of 2.14 days. This matches very closely to the average increase in attendance for the same group of -2.51. Meaning it is likely that the increase in late days is due to an increase in days attended.

This paper holds one important insight for policymakers. While the previous BEEP analysis found that bicycles did not affect test scores, I show that test scores improve if bicycles are distributed in communities actively promoting girls' education. Bicycles themselves may not improve test scores but, they may be an important mechanism for facilitating change when combined with other interventions.

References

Abadie, A. & Spiess, J. (2021). Robust Post-Matching Inference. *https://doi.org/10.1080/01621459.2020.1840383*. https://doi.org/10.1080/01621459.2020.1840383

Austin, P. C. (2007). Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. *The Journal of Thoracic and Cardiovascular Surgery*, *134*(5), 1128–1135. https://doi.org/10.1016/J.JTCVS.2007.07.021

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioural Research*. https://doi.org/10.1080/00273171.2011.568786

Austin, P. C. & Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine*, *33*(24), 4306–4319. https://doi.org/10.1002/SIM.6276

Benhassine, N., Devoto, F., Duflo, E., Dupas, P. & Pouliquen, V. (2015). Turning a Shove into a Nudge? A &quot;Labeled Cash Transfer&quot; for Education. *American Economic Journal: Economic Policy*, *7*(3), 86–125. https://doi.org/10.1257/POL.20130225

Burde, D. & Linden, L. L. (2013). Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools. *American Economic Journal: Applied Economics*, *5*(3), 27–40. https://doi.org/10.1257/APP.5.3.27

CARE. (2014). *CARE Education CARE's Youth Leadership index* (tech. rep.).

Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. *American Economic Review*, *91*(4), 795–813. https://doi.org/10.1257/AER.91.4.795

Fiala, N. (2017). The Impact of Bicycles on Girls' Education and Empowerment Outcomes in Zambia.

Gu, X. S. & Rosenbaum, P. R. (1993). Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405–420. https://doi.org/10.1080/10618600.1993.10474623

Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT. https://doi.org/10.1198/016214504000000647

Ho, D. E., Imai, K., King, G. & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(3), 199–236. https://doi.org/10.1093/PAN/MPL013

Imai, K., King, G. & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *171*(2), 481–502. https://doi.org/10.1111/J.1467-985X.2007.00527.X

Kremer, M. & Holla, A. (2008). Improving Education in the Developing World : What Have We Learned From Randomized Evaluations? 1.

Lee, B. K., Lessler, J. & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337–346. https://doi.org/10.1002/SIM.3782

Miguel, E. & Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, *72*(1), 159–217. https://doi.org/10.1111/J.1468-0262.2004.00481.X

Ming, K. & Rosenbaum, P. R. (2000). Substantial Gains in Bias Reduction from Matching with a Variable Number of Controls. *Biometrics*, *56*(1), 118–124. https://doi.org/10.1111/J.0006-341X.2000.00118.X

Muralidharan, K. & Prakash, N. (2017). Cycling to School: Increasing Secondary School Enrollment for Girls in India. *American Economic Journal: Applied Economics*, *9*(3), 321–50. https://doi.org/10.1257/APP.20160004

Nguyen, T.-L., Collins, G. S., Spence, J., Daurès, J.-P., Devereaux, P. J., Landais, P. & Le Manach, Y. (2017). Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. *BMC Medical Research Methodology 2017 17:1*, *17*(1), 1–8. https://doi.org/10.1186/S12874-017-0338-0

Normand, S. L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D. & McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, *54*(4), 387–398. https://doi.org/10.1016/S0895-4356(00)00321-8

Rosenbaum, P. R. (1991). A Characterization of Optimal Designs for Observational Studies. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(3), 597–610. https://doi.org/10.1111/J.2517-6161.1991.TB01848.X

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/BIOMET/70.1.41

Rosenbaum, P. R. & Rubin, D. B. (1985). *Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score* (tech. rep. No. 1).

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/H0037350

Rubin, D. B. (2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology 2001 2:3*, *2*(3), 169–188. https://doi.org/10.1023/A:1020363010465

Schultz, T. P. (2004). School subsidies for the poor: evaluating the Mexican Progresa poverty program. *Journal of Development Economics*, *74*(1), 199–250. https://doi.org/10.1016/J.JDEVECO.2003.12.009

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J. & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, *17*(6), 546–555. https://doi.org/10.1002/PDS.1555

Stuart, E. A. & Green, K. (2008). Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental psychology*, *44*(2), 395–406. https://doi.org/10.1037/0012-1649.44.2.395

World Bank. (2021). *Overcoming Economic Challenges, Natural Disasters, and the Pandemic : Social and Economic Impacts* (tech. rep.). Zimbabwe Economic Update;no. 3 Washington, D.C. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/563161623257944434/overcoming-economic-challenges-natural-disasters-and-the-pandemic-social-and-economic-impacts

Figures

Figure 1

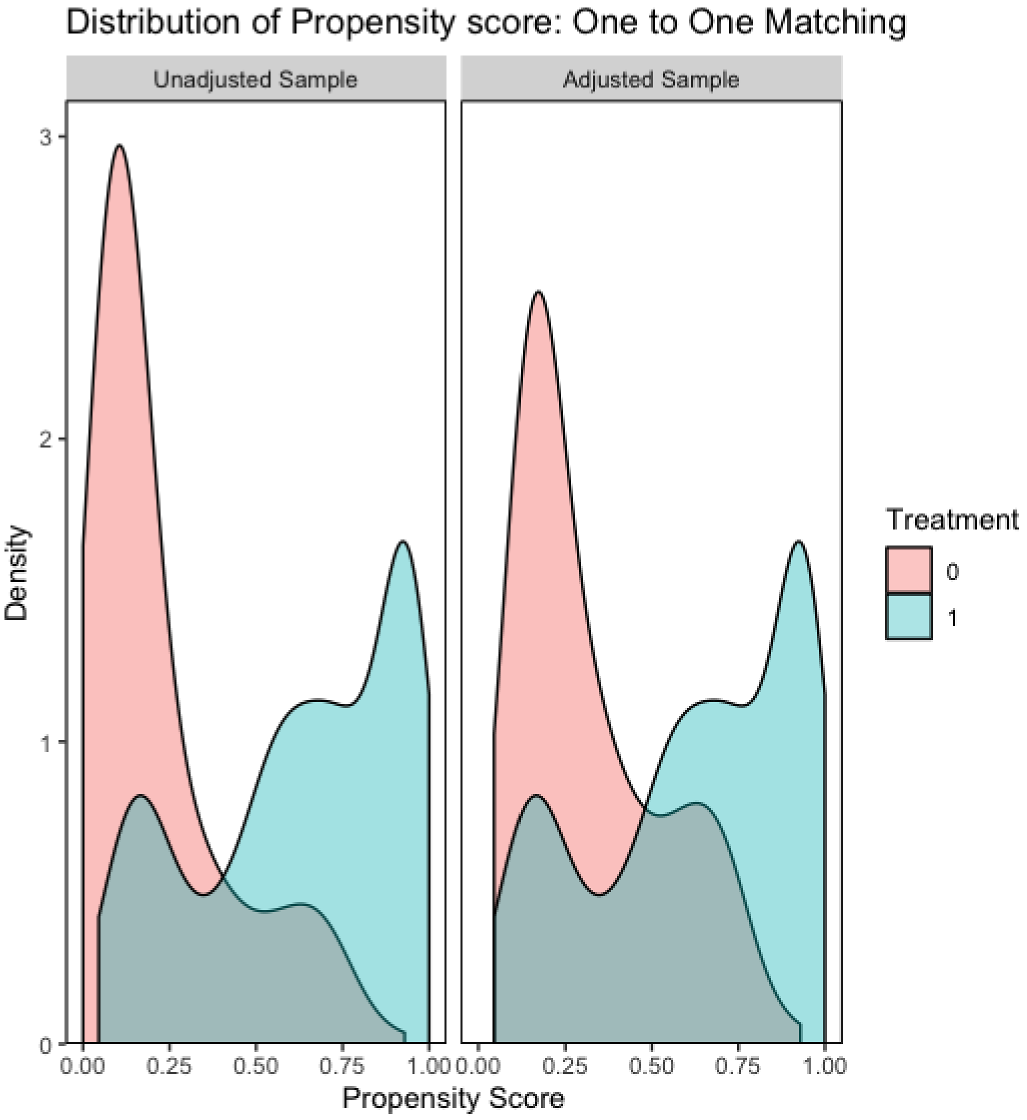
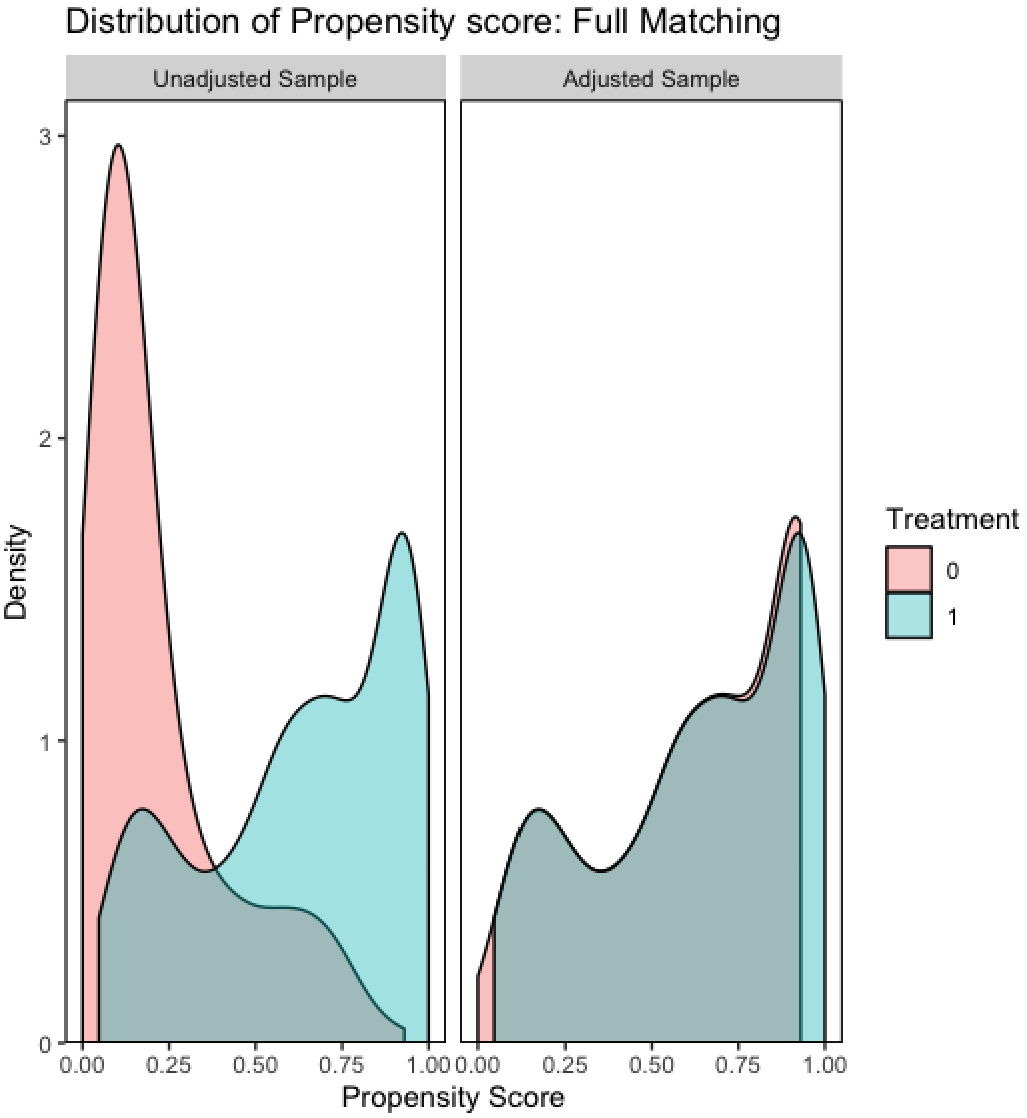


Figure 2


Distribution of Propensity score: Full Matching
Unadjusted Sample
Adjusted Sample
Density
3
2
1
0
0.00
0.25
0.50
0.75
1.00
Propensity Score
Treatment
0
1

Tables

Table 1

Standardized Mean Differences Before Matching

	Treatment		
	0	1	SMD
n	336	195	
diff_food_BL (mean (SD))	0.33 (0.47)	0.25 (0.43)	0.186
diff_food_BL_no (mean (SD))	0.67 (0.47)	0.75 (0.43)	0.186
diff_food_BL_na (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
orphan_BL (mean (SD))	0.12 (0.33)	0.12 (0.33)	0.003
b_hoh_m (mean (SD))	0.45 (0.50)	0.34 (0.47)	0.228
b_hoh_f (mean (SD))	0.38 (0.48)	0.25 (0.43)	0.269
b_mother (mean (SD))	0.00 (0.00)	0.01 (0.07)	0.101
b_mother_no (mean (SD))	0.88 (0.33)	0.62 (0.49)	0.623
b_mother_NA (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
b_married (mean (SD))	0.00 (0.00)	0.01 (0.07)	0.101
b_married_no (mean (SD))	0.88 (0.33)	0.62 (0.49)	0.623
b_married_NA (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
b_hoh_ed (mean (SD))	0.90 (0.30)	0.90 (0.30)	0.008
b_safe_travel (mean (SD))	0.78 (0.42)	0.79 (0.41)	0.04
b_chores (mean (SD))	0.15 (0.36)	0.10 (0.30)	0.156
b_chores_no (mean (SD))	0.68 (0.47)	0.48 (0.50)	0.412
b_chores_NA (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
b_disab (mean (SD))	0.08 (0.27)	0.06 (0.24)	0.062
b_disab_no (mean (SD))	0.78 (0.42)	0.60 (0.49)	0.395
b_insiza (mean (SD))	0.10 (0.30)	0.06 (0.23)	0.166
b_chivi (mean (SD))	0.49 (0.50)	0.30 (0.46)	0.38
b_mangwe (mean (SD))	0.17 (0.37)	0.24 (0.43)	0.173

Table 1

Standardized Mean Differences Before Matching

	Treatment		
b_mberengwa (mean (SD))	0.25 (0.43)	0.13 (0.34)	0.307
b_dist_0.25 (mean (SD))	0.15 (0.36)	0.07 (0.25)	0.267
b_dist_0.5 (mean (SD))	0.17 (0.38)	0.10 (0.30)	0.221
b_dist_0.75 (mean (SD))	0.08 (0.27)	0.07 (0.25)	0.041
b_dist_1 (mean (SD))	0.17 (0.38)	0.18 (0.38)	0.018
b_dist_2 (mean (SD))	0.19 (0.39)	0.20 (0.40)	0.032
b_dist_NA (mean (SD))	0.20 (0.40)	0.09 (0.29)	0.306
grade_3_BL (mean (SD))	0.29 (0.46)	0.08 (0.27)	0.582
grade_4_BL (mean (SD))	0.28 (0.45)	0.06 (0.23)	0.625
grade_5_BL (mean (SD))	0.21 (0.41)	0.11 (0.31)	0.285
grade_6_BL (mean (SD))	0.08 (0.28)	0.24 (0.43)	0.425
grade_7_BL (mean (SD))	0.07 (0.25)	0.21 (0.40)	0.405
form1_BL (mean (SD))	0.01 (0.09)	0.01 (0.10)	0.014
grade_NA_BL (mean (SD))	0.04 (0.20)	0.31 (0.46)	0.746
far (mean (SD))	0.43 (0.50)	0.59 (0.49)	0.326

Table 2

Standardized Mean Differences: One to one Matching

	Treatment		
	0	1	SMD
n	195	195	
diff_food_BL (mean (SD))	0.30 (0.46)	0.25 (0.43)	0.115
diff_food_BL_no (mean (SD))	0.70 (0.46)	0.75 (0.43)	0.115
diff_food_BL_na (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
orphan_BL (mean (SD))	0.13 (0.34)	0.12 (0.33)	0.015
b_hoh_m (mean (SD))	0.47 (0.50)	0.34 (0.47)	0.273
b_hoh_f (mean (SD))	0.32 (0.47)	0.25 (0.43)	0.148
b_mother (mean (SD))	0.00 (0.00)	0.01 (0.07)	0.101
b_mother_no (mean (SD))	0.84 (0.37)	0.62 (0.49)	0.523
b_mother_NA (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
b_married (mean (SD))	0.00 (0.00)	0.01 (0.07)	0.101
b_married_no (mean (SD))	0.84 (0.37)	0.62 (0.49)	0.523
b_married_NA (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
b_hoh_ed (mean (SD))	0.92 (0.27)	0.90 (0.30)	0.066
b_safe_travel (mean (SD))	0.77 (0.42)	0.79 (0.41)	0.065
b_chores (mean (SD))	0.10 (0.30)	0.10 (0.30)	0.017
b_chores_no (mean (SD))	0.68 (0.47)	0.48 (0.50)	0.413
b_chores_NA (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
b_disab (mean (SD))	0.07 (0.26)	0.06 (0.24)	0.041
b_disab_no (mean (SD))	0.76 (0.43)	0.60 (0.49)	0.357
b_insiza (mean (SD))	0.14 (0.35)	0.06 (0.23)	0.293
b_chivi (mean (SD))	0.45 (0.50)	0.30 (0.46)	0.31
b_mangwe (mean (SD))	0.17 (0.38)	0.24 (0.43)	0.152
b_mberengwa (mean (SD))	0.23 (0.42)	0.13 (0.34)	0.269

Table 2

Standardized Mean Differences: One to one Matching

	Treatment		
b_dist_0.25 (mean (SD))	0.11 (0.32)	0.07 (0.25)	0.162
b_dist_0.5 (mean (SD))	0.11 (0.32)	0.10 (0.30)	0.05
b_dist_0.75 (mean (SD))	0.07 (0.26)	0.07 (0.25)	0.02
b_dist_1 (mean (SD))	0.26 (0.44)	0.18 (0.38)	0.198
b_dist_2 (mean (SD))	0.24 (0.43)	0.20 (0.40)	0.087
b_dist_NA (mean (SD))	0.16 (0.37)	0.09 (0.29)	0.215
grade_3_BL (mean (SD))	0.21 (0.40)	0.08 (0.27)	0.374
grade_4_BL (mean (SD))	0.15 (0.36)	0.06 (0.23)	0.307
grade_5_BL (mean (SD))	0.30 (0.46)	0.11 (0.31)	0.496
grade_6_BL (mean (SD))	0.14 (0.35)	0.24 (0.43)	0.236
grade_7_BL (mean (SD))	0.12 (0.32)	0.21 (0.40)	0.238
form1_BL (mean (SD))	0.02 (0.12)	0.01 (0.10)	0.045
grade_NA_BL (mean (SD))	0.07 (0.25)	0.31 (0.46)	0.648
far (mean (SD))	0.44 (0.50)	0.59 (0.49)	0.3

Table 3

Standardized Mean Differences Full Matching

	Treatment		
	0	1	SMD
n	336	195	
diff_food_BL (mean (SD))	0.29 (0.46)	0.25 (0.43)	0.109
diff_food_BL_no (mean(SD))	0.71 (0.46)	0.75 (0.43)	0.109
diff_food_BL_na (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
orphan_BL (mean (SD))	0.11 (0.32)	0.12 (0.33)	0.028
b_hoh_m (mean (SD))	0.61 (0.49)	0.34 (0.47)	0.557
b_hoh_f (mean (SD))	0.25 (0.43)	0.25 (0.43)	0.013
b_mother (mean (SD))	0.00 (0.00)	0.01 (0.07)	0.101
b_mother_no (mean (SD))	0.59 (0.49)	0.62 (0.49)	0.061
b_mother_NA (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
b_married (mean (SD))	0.00 (0.00)	0.01 (0.07)	0.101
b_married_no (mean (SD))	0.59 (0.49)	0.62 (0.49)	0.061
b_married_NA (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
b_hoh_ed (mean (SD))	0.95 (0.21)	0.90 (0.30)	0.194
b_safe_travel (mean (SD))	0.77 (0.42)	0.79 (0.41)	0.062
b_chores (mean (SD))	0.10 (0.30)	0.10 (0.30)	0.017
b_chores_no (mean (SD))	0.46 (0.50)	0.48 (0.50)	0.042
b_chores_NA (mean (SD))	0.00 (0.00)	0.00 (0.00)	<0.001
b_disab (mean (SD))	0.08 (0.27)	0.06 (0.24)	0.07
b_disab_no (mean (SD))	0.55 (0.50)	0.60 (0.49)	0.106
b_insiza (mean (SD))	0.13 (0.34)	0.06 (0.23)	0.257
b_chivi (mean (SD))	0.28 (0.45)	0.30 (0.46)	0.043
b_mangwe (mean (SD))	0.19 (0.39)	0.24 (0.43)	0.116
b_mberengwa (mean (SD))	0.40 (0.49)	0.13 (0.34)	0.643

b_dist_0.25 (mean (SD))	0.07 (0.26)	0.07 (0.25)	0.013
b_dist_0.5 (mean (SD))	0.08 (0.28)	0.10 (0.30)	0.053
b_dist_0.75 (mean (SD))	0.08 (0.27)	0.07 (0.25)	0.055
b_dist_1 (mean (SD))	0.12 (0.33)	0.18 (0.38)	0.159
b_dist_2 (mean (SD))	0.19 (0.39)	0.20 (0.40)	0.037
b_dist_NA (mean (SD))	0.12 (0.33)	0.09 (0.29)	0.092
grade_3_BL (mean (SD))	0.07 (0.26)	0.08 (0.27)	0.015
grade_4_BL (mean (SD))	0.07 (0.25)	0.06 (0.23)	0.047
grade_5_BL (mean (SD))	0.10 (0.31)	0.11 (0.31)	0.01
grade_6_BL (mean (SD))	0.22 (0.42)	0.24 (0.43)	0.033
grade_7_BL (mean (SD))	0.16 (0.36)	0.21 (0.40)	0.123
form1_BL (mean (SD))	0.01 (0.08)	0.01 (0.10)	0.037
grade_NA_BL (mean (SD))	0.37 (0.48)	0.31 (0.46)	0.127
far (mean (SD))	0.64 (0.48)	0.59 (0.49)	0.112

Table 4

Treatment effect on Days of School Missed

	Midline - Baseline		Endline - Midline		Endline - Basline	
	1	2	3	4	5	6
Intercept	0.32	−0.31	−0.94**	0.29	−0.14	−0.43
	(0.36)	(0.58)	(0.40)	(0.39)	(0.23)	(0.51)
Beep Treatment	−0.73	−0.44	0.85*	0.39	−0.40	−0.31
	(0.47)	(0.60)	(0.44)	(0.37)	(0.33)	(0.44)
Control Variables	No	Yes	No	Yes	No	Yes
R^2	0.02	0.06	0.04	0.14	0.01	0.02
Adj. R^2	0.01	0.04	0.04	0.12	0.00	−0.00
Num. obs.	357	297	452	340	355	299
RMSE	2.54	2.60	2.01	2.05	2.05	2.14
N Clusters	91	83	93	87	90	83

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 5

Treatment effect on Literacy and Numeracy scores

	Midline - Baseline			
	Numeracy Scores	Numeracy Scores	Literacy Scores	Literacy Scores
Intercept	-0.43^{**}	-0.15	-0.31^{**}	-0.13
	(0.17)	(0.14)	(0.14)	(0.09)
Beep Treatment	0.32	0.28^{**}	0.28	0.23^{**}
	(0.20)	(0.13)	(0.18)	(0.11)
Control Variables	No	Yes	No	Yes
R^2	0.04	0.15	0.05	0.22
Adj. R^2	0.03	0.14	0.05	0.20
Num. obs.	479	392	479	392
RMSE	0.78	0.74	0.56	0.52
N Clusters	94	88	94	88

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 6

Treatment effect on YLI

	Midline - Baseline		Endline - Midline		Endline - Basline	
	YLI Scores	YLI Scores	YLI Scores	YLI Scores	YLI Scores	YLI Scores
Intercept	−0.19	0.30	0.18	0.11	0.01	0.36
	(0.17)	(0.28)	(0.13)	(0.21)	(0.19)	(0.33)
Beep Treatment	0.10	−0.03	−0.20	−0.15	−0.11	−0.19
	(0.21)	(0.20)	(0.16)	(0.20)	(0.20)	(0.24)
Control Variables	No	Yes	No	Yes	No	Yes
R^2	0.00	0.11	0.01	0.13	0.00	0.13
Adj. R^2	−0.00	0.09	0.01	0.11	−0.00	0.11
Num. obs.	443	368	531	392	443	368
RMSE	1.20	1.16	1.15	1.07	1.26	1.23
N Clusters	94	87	94	88	94	87

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 7

Treatment effect on Reported Days Late

	Endline - Midline	
	Days Late	Days Late
Intercept	-1.21^{*}	-0.68
	(0.70)	(0.81)
Beep Treatment	1.37^{**}	1.59^{***}
	(0.64)	(0.56)
Control Variables	No	Yes
R^2	0.04	0.17
Adj. R^2	0.04	0.16
Num. obs.	442	333
RMSE	3.17	2.66
N Clusters	93	86

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 8

Treatment effect on Days of School Missed: Disabled Subgroup

	Midline - Baseline		Endline - Midline		Endline - Basline	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	0.29	−0.30	-1.40^{**}	0.08	-0.99^{*}	-1.98^{**}
	(0.67)	(1.12)	(0.49)	(0.77)	(0.47)	(0.68)
Beep Treatment	−1.62	−0.66	1.51^{**}	0.52	−0.24	0.73
	(1.02)	(0.86)	(0.58)	(0.66)	(0.84)	(0.55)
Control Variables	No	Yes	No	Yes	No	Yes
R^2	0.16	0.34	0.29	0.52	0.00	0.43
Adj. R^2	0.12	0.08	0.26	0.36	−0.05	0.19
Num. obs.	23	19	28	25	21	18
RMSE	1.83	1.72	1.08	1.07	1.52	1.45
N Clusters	19	16	19	18	16	15

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 9

Treatment effect on Numeracy and Literacy Scores: Disabled Subgroup

	Midline - Baseline			
	Numeracy Scores	Numeracy Scores	Literacy Scores	Literacy Scores
Intercept	-0.66^{**}	-0.12	-0.58^{***}	0.02
	(0.24)	(0.19)	(0.20)	(0.22)
Beep Treatment	0.36	-0.05	0.46^{**}	-0.07
	(0.31)	(0.17)	(0.21)	(0.24)
Control Variables	No	Yes	No	Yes
R^2	0.06	0.77	0.13	0.66
Adj. R^2	0.03	0.71	0.11	0.58
Num. obs.	38	30	38	30
RMSE	0.67	0.36	0.56	0.42
N Clusters	25	21	25	21

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 10

Treatment effect on YLI : Disabled Subgroup

	Midline - Baseline		Endline - Midline		Endline - Basline	
	YLI	YLI	YLI	YLI	YLI	YLI
Intercept	−0.31	−0.07	0.36	0.25	0.05	0.16
	(0.34)	(0.72)	(0.33)	(0.26)	(0.34)	(0.72)
Beep Treatment	0.85^{*}	0.91	-0.69^{*}	-0.84^{**}	0.15	0.08
	(0.47)	(0.70)	(0.39)	(0.35)	(0.38)	(0.63)
Control Variables	No	Yes	No	Yes	No	Yes
R^2	0.15	0.50	0.12	0.36	0.00	0.46
Adj. R^2	0.12	0.35	0.09	0.20	−0.03	0.30
Num. obs.	35	27	38	30	35	27
RMSE	0.99	0.92	0.91	0.63	1.03	0.86
N Clusters	23	19	25	21	23	19

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 11

Treatment effect on Days of School Missed: Heavy Chore Burden Subgroup

	Midline - Baseline		Endline - Midline		Endline - Basline	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	1.26^{**}	-0.62	-0.96	1.13^{*}	0.04	0.72
	(0.52)	(0.40)	(0.80)	(0.57)	(0.40)	(0.75)
Beep Treatment	-1.01	0.71^{*}	0.80	-0.85	0.02	-0.32
	(0.66)	(0.37)	(0.91)	(0.51)	(0.53)	(0.56)
Control Variables	No	Yes	No	Yes	No	Yes
R^2	0.07	0.47	0.04	0.40	0.00	0.13
Adj. R^2	0.05	0.39	0.02	0.32	-0.02	-0.01
Num. obs.	53	45	61	53	52	45
RMSE	1.58	1.02	1.75	1.14	1.12	1.15
N Clusters	32	29	37	35	31	29

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 12

Treatment effect on Literacy and Numeracy scores: Heavy Chore Burden Subgroup

	Midline - Baseline			
	Numeracy Scores	Numeracy Scores	Literacy Scores	Literacy Scores
Intercept	−0.04	−0.29	−0.12	−0.18
	(0.15)	(0.31)	(0.16)	(0.21)
Beep Treatment	0.09	0.19	0.09	0.01
	(0.30)	(0.35)	(0.25)	(0.25)
Control Variables	No	Yes	No	Yes
R^2	0.00	0.25	0.00	0.27
Adj. R^2	−0.01	0.16	−0.01	0.19
Num. obs.	72	61	72	61
RMSE	0.75	0.73	0.56	0.50
N Clusters	39	36	39	36

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 13

Treatment effect on YLI scores: Heavy Chore Burden Subgroup

	Midline - Baseline		Endline - Midline		Endline - Basline	
	YLI Scores	YLI Scores	YLI Scores	YLI Scores	YLI Scores	YLI Scores
Intercept	−0.25	0.17	−0.25	−0.34	−0.08	−0.37
	(0.38)	(0.64)	(0.38)	(0.54)	(0.44)	(0.96)
Beep Treatment	0.55	−0.29	0.55	0.32	0.39	0.25
	(0.59)	(0.52)	(0.59)	(0.53)	(0.56)	(0.64)
Control Variables	No	Yes	No	Yes	No	Yes
R^2	0.03	0.17	0.03	0.16	0.02	0.13
Adj. R^2	0.02	0.07	0.02	0.07	0.00	0.02
Num. obs.	72	57	72	61	66	57
RMSE	1.28	1.13	1.28	1.15	1.13	1.11
N Clusters	39	34	39	36	37	34

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 14

Treatment effect on Days of School Missed: Subgroup Far

	Midline - Baseline		Endline - Midline		Endline - Basline	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	0.21	−0.42	-1.42^{***}	0.25	-0.55^{*}	−0.18
	(0.67)	(0.93)	(0.38)	(0.80)	(0.31)	(0.55)
Beep Treatment	−0.44	−0.42	1.35^{***}	1.12	0.41	0.61
	(0.78)	(1.11)	(0.44)	(0.86)	(0.43)	(0.47)
Control Variables	No	Yes	No	Yes	No	Yes
R^2	0.00	0.16	0.09	0.23	0.01	0.11
Adj. R^2	−0.00	0.11	0.08	0.19	0.01	0.06
Num. obs.	144	119	220	147	147	123
RMSE	2.81	2.78	2.33	2.47	1.49	1.55
N Clusters	61	53	69	59	62	53

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 15

Treatment effect on Literacy and Numeracy scores: Subgroup Far

	Midline - Baseline			
	Numeracy Scores	Numeracy Scores	Literacy Scores	Literacy Scores
Intercept	-0.54^{**}	-0.16	-0.39^{**}	-0.09
	(0.21)	(0.23)	(0.17)	(0.14)
Beep Treatment	0.57^{**}	0.54^{**}	0.43	0.37^{**}
	(0.28)	(0.22)	(0.27)	(0.18)
Control Variables	No	Yes	No	Yes
R^2	0.09	0.30	0.09	0.40
Adj. R^2	0.08	0.27	0.08	0.37
Num. obs.	207	168	207	168
RMSE	0.91	0.82	0.68	0.58
N Clusters	77	64	77	64

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 16

Treatment effect on YLI scores: Subgroup Far

	Midline - Baseline		Endline - Midline		Endline - Basline	
	YLI Scores	YLI Scores	YLI Scores	YLI Scores	YLI Scores	YLI Scores
Intercept	-0.17	0.24	0.22	0.19	0.08	0.35
	(0.24)	(0.40)	(0.14)	(0.34)	(0.25)	(0.57)
Beep Treatment	0.33	0.38	-0.31^{**}	-0.50	0.02	-0.09
	(0.31)	(0.27)	(0.15)	(0.32)	(0.27)	(0.35)
Control Variables	No	Yes	No	Yes	No	Yes
R^2	0.02	0.19	0.02	0.17	0.00	0.12
Adj. R^2	0.01	0.15	0.02	0.13	-0.01	0.08
Num. obs.	192	157	259	168	192	157
RMSE	1.29	1.22	1.11	1.02	1.29	1.25
N Clusters	74	61	77	64	74	61

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 17

Treatment effect on Days of school missed: Distance analysis

	0-15 Mins	16-30 Mins	31-45 Mins	46-1 hour	1-2 hours	Greater than 3
(Intercept)	0.70^{*}	0.65	−0.80	0.14	−0.55	0.11
	(0.38)	(0.56)	(0.65)	(0.43)	(0.43)	(0.17)
beep_treatment	−0.40	−2.51	0.30	-0.89^{*}	0.09	-1.11^{***}
	(0.41)	(1.57)	(0.74)	(0.49)	(0.56)	(0.17)
R^2	0.02	0.10	0.01	0.02	0.00	0.64
Adj. R^2	−0.00	0.08	−0.02	0.01	−0.01	0.55
Num. obs.	46	54	32	76	78	6
RMSE	1.15	2.98	1.15	2.66	1.80	0.31
N Clusters	28	23	23	37	45	4

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 18

Treatment effect on reported late days: Distance analysis

	0-15 Mins	16-30 Mins	31-45 Mins	46-1 hour	1-2 hours	Greater than 3
(Intercept)	1.03	−1.14	−1.39	−0.78	−0.12	3.58
	(1.62)	(0.90)	(1.16)	(0.56)	(1.39)	(1.66)
beep_treatment	0.06	2.14^{*}	1.16	0.82	0.19	−4.58
	(1.66)	(1.10)	(1.31)	(0.72)	(1.45)	(1.66)
R^2	0.00	0.12	0.03	0.02	0.00	0.35
Adj. R^2	−0.02	0.11	0.00	0.01	−0.01	0.14
Num. obs.	49	69	35	79	85	5
RMSE	2.97	2.24	2.76	2.62	4.27	1.65
N Clusters	27	28	26	41	45	3

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$