WPS7485

Policy Research Working Paper 7485

When Do In-service Teacher Training and Books Improve Student Achievement?

Experimental Evidence from Mongolia

Habtamu Fuje
Prateek Tandon

WORLD BANK GROUP
Education Global Practice Group
November 2015

Policy Research Working Paper 7485

Abstract

This study presents evidence from a randomized control trial (RCT) in Mongolia on the impact of in-service teacher training and books, both as separate educational inputs and as a package. The study tests for the complementarity of inputs and non-linearity of returns from investment in education as measured by students' test scores in five subjects. It takes advantage of a national-scale RCT conducted under the Rural Education and Development project. The results suggest that the provision of books, in addition to teacher training, raises student achievement substantially. However, teacher training and books weakly improve test scores when provided individually. Students whose teachers have received training and whose classrooms have acquired books improved their cumulative score (totaled across five tests) by 34.9 percent of a standard deviation, relative to a control group. Students treated only with books improved their total score by 20.6 percent of a standard deviation relative to a control group of students. On the other hand, extra teacher training did not have a statistically significant effect on the total test score. In addition, providing both inputs jointly improved test scores in most subjects, which was not the case when either input was provided individually. This study sheds light on the relevance of supplementing teacher training schemes with appropriate teaching materials in resource-poor settings. The policy implication is that isolated education investments, in settings where complementary inputs are missing, could deliver minimal or no return.

This paper is a product of the Education Global Practice Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at hfuje@worldbank.org and ptandon@worldbank.org.

When Do In-service Teacher Training and Books Improve Student Achievement? Experimental Evidence from Mongolia

Habtamu Fuje*
Columbia University

Prateek Tandon
The World Bank

Keywords: In-service teacher training, RCT, matching, impact
JEL Classification: I28, I21, O15

* Corresponding author: Habtamu.Fuje@columbia.edu or Habtamu_Fuje@post.harvard.edu
The study also benefited from discussion with faculty and graduate students at Columbia University. Yabibal Walle, University of Göttingen, Andinet Woldemichael, Georgia State University, and Kefyalew Endale, National Graduate Institute for Policy Studies, proofread the draft version. Charles Abelmann, Cristobal Ridao-Cano, and Katherine Nesmith of the World Bank originally designed the evaluation. D. Khishigbuyan, Project Coordinator of READ, provided assistance throughout project implementation and follow up. Deon P. Filmer and David Evans, from the World Bank, graciously provided invaluable comments and suggestions. We thank you all.

1 INTRODUCTION

Policy makers and practitioners in developing and developed countries often invest heavily in brief in-service teacher training to enhance education outcomes. Spurred by the targets of the Millennium Development Goals (MDGs), developing countries have also rapidly expanded school infrastructure in the past decade and ramped up in-service teacher training. These investments have aimed to satisfy the growing demand for teachers and help improve educational quality (GOM, 2007; Bunyi et al., 2013; Kidwai et al., 2013). However, conclusive evidence on the impact of in-service teacher training on student achievement—as measured by a comprehensive set of test scores—scarcely exists, particularly in developing countries. Moreover, the differential impact of such training on achievement when students and teachers have access to appropriate books to effectively implement the lessons learned during training—versus when they do not—has not been investigated. Previous studies have focused on the individual provision of either teacher training or books and have not examined a potential complementarity between these inputs.

Properly documenting the impact of such investments on student outcomes can address this gap. The few rigorous evaluations of teacher training programs conducted to date suggest a moderate potential to improve student outcomes, but the evidence is mixed. A recent systematic review by Glewwe et al. (2013), which examined impact evaluation studies from 1990-2010, concluded that there is only modest evidence that teacher training improves student test scores. Specifically, 11 of the 29 estimates included in their analysis demonstrate positive, significant impacts (one is significant and negative). But, only three of these studies were well identified, experimental or based on natural experiments. Other works on the impacts of teacher training also do not provide conclusive positive evidence: improvements in test scores were documented by some (see Jacob and Lefgren (2004); Zhang et al., 2013; Raudenbush et al., 1993), while others find no evidence (see Angrist and Lavy, 2001; Harris and Sass, 2011; and Lai et al., 2011). Evans and Popova (2014) noted that the type of teacher training also matters; a one-time in-service training might be as effective as long-term peer mentoring/coaching.

With regards to the impacts of books, the same review by Glewwe et al. (2013) revealed that, in general, there is strong, but non-unanimous, evidence for the positive

impact of textbooks and workbooks on student learning. However, when considering well identified studies only, they noted weak evidence. Older studies suggest that books improve achievement (Heyneman et al., 1984; Jamison et al., 1981), while more recent studies in Kenya (Glewwe et al., 1998 and Glewwe et al., 2009) and in Sierra Leone (Sabarwal et al., 2014) contradict these findings.

Most of these previous studies, however, have had some methodological limitations. The most serious methodological issue with observational studies is the non-random assignment of teachers to in-service training programs or students to book provision. A few quasi-experimental studies have attempted to address these issues (Rothstein, 2010; Jacob and Lefgren, 2004; Angrist and Lavy, 2001). A number of issues arising from non-random assignment need to be addressed. For instance, factors like self-initiation, relationships with supervisors, personal connections and political participation confound with a teacher's decision to attend in-service training as well as her general motivation and capacity to teach (see Jacob and Lefgren (2004)). Similarly, a student's access to books confound with a number of other covariates such as parental education, wealth, and school resources, which directly affect student outcomes.

This study uses data from the randomized assignments of teachers into a training program or the provision of books to randomly selected primary schools in Mongolia under the Rural Education and Development (READ) project to examine the impacts of these interventions on student achievement. The randomization is nationally representative–it covers the entire rural population of the whole country, as opposed to a typical small-scale randomization study from which generalization to national population is not feasible. This enables us to address limitations arising from non-random assignment and provides a basis to generalize about the impact of the interventions.

In addition, this study investigates the differential impact of in-service teacher training or book provision as a stand-alone intervention vis-à-vis in-service training accompanied by provision of age-appropriate books. Some previous evidence on the topic suggests that provision of education inputs as a bundle is more effective in improving outcomes (see McEwan (2014); Evans and Popova (2014); and Conn (2014) for detailed review). The evaluation of these interrelated investments sheds light on the potential complementarity of educational inputs, and non-linearity of returns to

education investment by comparing returns to provision of books or teacher training alone against returns from training teachers along with the provision of books. This addresses the question of whether the sum of returns from "extra teacher training" and "books only" interventions is lower or higher than the return from training complemented by books. If the sum of returns from the individual interventions is lower than return from the joint investment, then evidence for complementarity of books and training in education production exists.

We ask two questions that have fundamental policy relevance: (1) Do short in-service teacher training and books improve students' test scores when provided individually in a resource-poor setting? (2) How does the return from the joint provision of these inputs compare with sum of returns from providing each input individually? We find significant, positive effects on student outcomes when books and training were provided together as a package, rather than as individual inputs. Books only and extra teacher training marginally improved test scores in some, but not all, subjects. The magnitude of impact of either input was not academically significant. However, when teachers are trained and students are provided with books, the test scores of a treatment group of students increased substantially, relative to a control group of students.

The rest of the paper is organized as follows. Section 2 presents a brief description of the context, and detailed discussion of the survey design, instruments and interventions. Section 3 outlines the framework and identification strategies employed. Section 4 presents descriptive and analytic results, an investigation of heterogeneity in treatment effects, and robustness checks. A discussion of results is provided in Section 5.

2 CONTEXT, SURVEY DESIGN AND INTERVENTION

2.1 Context

The Ministry of Education, Culture and Science (MECS) developed a new Education Sector Master Plan (ESMP2) for 2006-2015 that built on the General Guideline for Socio-Economic Development of Mongolia (GGSEDM) for 2006-2008. The GGSEDM identified five priority actions for education: (1) reduce school dropout and provide elementary education for all; (2) transform the education system into an 11-year

system by 2006 and then into a 12-year system by 2007; (3) improve the learning environment, physical facilities supply of teachers and textbooks at secondary schools ; (4) lower gender inequality in primary and secondary school enrollment; and (5) increase accessibility of schools for children with disabilities. The ESMP2 sought to sequence the government priorities by: (1) upgrading education quality at all levels of schooling; (2) providing education services to children in all parts of the country, including rural areas, and to the poor and vulnerable groups; and (3) improving the management capacity of central and local educational institutions. The government acknowledged that low levels of educational attainment were key determinants of poverty, and that poverty could be a key factor that limited access to and quality of schooling. These efforts were in response to the dramatic decline in support for the country's education system after its transition to a free market economy in the early 1990s. Enrollment in rural schools declined rapidly, and access to high-quality learning materials diminished. Schools in rural areas had few textbooks and little or no supplementary reading books (World Bank, 2013).

2.2 Intervention and design

To improve the quality of primary education in rural Mongolia, MECS, with technical and financial support from the World Bank, implemented a comprehensive rural education program, the READ project. READ's main policy instruments were availing high-quality children's books and improving teachers' skills through in-service training schemes. Under this project, primary schools received grade-specific classroom libraries, which entailed equipping classrooms with grade-appropriate books and shelves for these books. These books were used during class hours, and students were also occasionally allowed to borrow them for use at home. Each classroom received about 160 books. These education materials were provided at a very low cost. The average costs (in 2008 US$) of a single book and a set of shelves were $2.1 and $71.5, respectively (World Bank, 2013).

Primary school teachers participated in an intensive training to improve their skills to support students in math, reading and writing activities. The training was rolled out in a cascade model: the training of the trainer-teachers was implemented first. Afterwards, these trainers trained fellow teachers on how to improve their students'

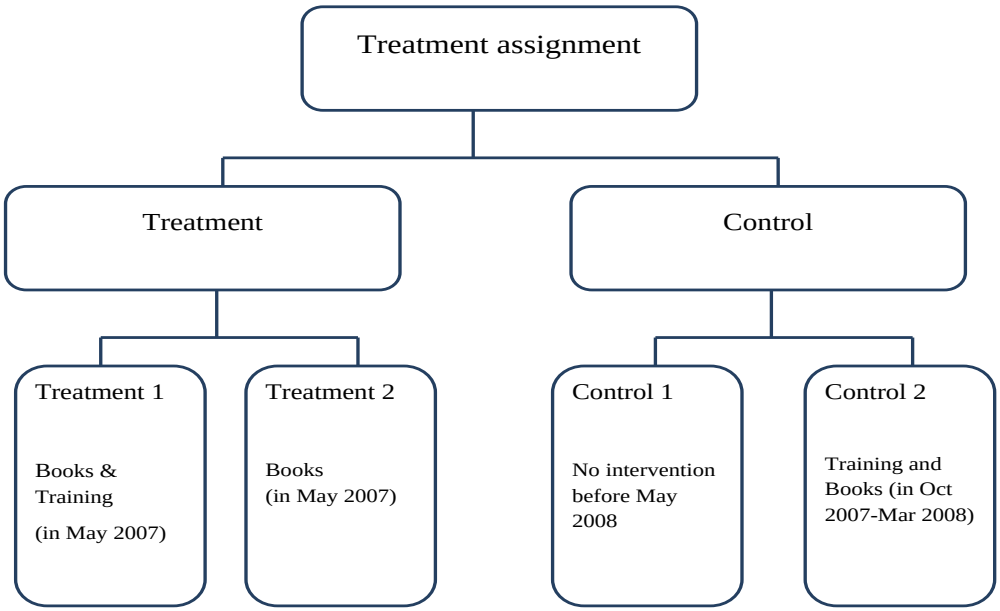
math, reading and writing skills. About 178 mentor/trainer teachers were trained for four days by well qualified national trainers, and then they conducted an average of 2.26 visits per school to mentor fellow teachers. The training of fellow teachers lasted for three days. This cascading of training enabled the delivery of teacher training in a more cost effective manner than other teacher training projects. The training cost was $ 3.14 per day per teacher under READ, relative to $ 7.62 for other similar training schemes in the country (World Bank, 2013).

To evaluate the impact of teacher training or books alone as well as teacher training complemented by books, a national-scale randomization was carried out. The initial design of the evaluation strategy was such that schools in the 21 provinces/*aimags* would be randomly assigned to Treatment One ($T1$), Treatment Two ($T2$) or a control group (C) (see Figure 1). The control group was later divided into two: Control One ($C1$) and Control Two ($C2$).[1,2] $T1$ includes primary schools randomly selected in five provinces (Arkhangai, Bulgan, Zavkhan, Sukhbaatar and Tov), and these schools received classroom libraries and in-service teacher training in May 2007. $T2$, schools in Ovurkhangai and Govi-Altai provinces, were provided classroom libraries, but not teacher training, in May 2007. $C1$, which includes schools in Dornogovi, Omnogovi, Uvs, Khovd, Khovsgol, Khentii and Govisumber provinces, was originally to receive classroom libraries and teacher training at the end of the experiment (in May 2008), but the plan was changed later and it received books in October 2007 and training between October 2007 and March 2008. $C2$ encompasses schools in Bayan-Olgii, Bayankhongor, Dornod, Dondgovi, Selenge, Darkhan-Uul and Orkhon provinces, and these schools received treatment at the end of the experiment (books in May 2008 and training between May and September 2008). Figure A.1 (see annex) shows *aimags* in which the four groups of schools are located, and Table 1 presents the timeframe of the survey and interventions, and the number of schools and students surveyed.

[1] $C1$ received treatment halfway through the study period. Therefore, direct comparison of $T1$ and $T2$ with C will not be feasible. In addition, the 'pure' control group ($C2$) has smaller sample size. Hence, the follow up survey included additional schools in the sample.

[2] Administratively, Mongolia is divided into 21 *aimags* and the capital city, Ulaanbaatar. These *aimags* are further divided into *soums*, and then into *bags* (NSO, 2006).

Figure 1: Assignment to treatment and control groups



To reduce spillover effects and ensure the political feasibility of providing schools with different inputs, a given province was allowed to have either treatment or control schools. Then, in these selected schools, a class from two grades (specifically, from third and fourth grade) was randomly selected and surveyed.[3,4] In the upcoming sections, we discuss the limitation of confining treatment and control schools in selected provinces, instead of allowing each province to have both types of schools, and we cluster standard error at province level to correct for this limitation. Finally, students within a class were randomly selected if the class size was more than 20; otherwise the whole class was surveyed.

The baseline survey was conducted during April-May 2007, just before the end of the academic year, and it encompassed 137 schools, 141 teachers and 2,612 students. A follow-up survey was conducted in April 2008. In the follow-up survey, additional schools, classes, teachers and students were surveyed to address initial imbalances in

[3]The classes were sorted alphabetically, and if there were at least 20 students in the first class of each grade, it was selected as a sample. Otherwise, next class with at least 20 students was selected. If such class did not exist, the class with the highest number of students was selected.

[4]In 2004, Mongolia has began a transiting from a ten-grade education system, with four primary school grades, to a twelve-grade system (Yang and Sato, 2009).

the number of observations in the treatment and control groups during the baseline survey. It covered 172 schools, 311 classes, 308 teachers and 5,322 students (see Table 1). The follow-up survey covered all students and teachers who were surveyed in the baseline, but also included additional teachers and students. The cause of imbalance and how this additional observation is leveraged to address the imbalances is discussed under the 'identification strategy' subsection.

Table 1: Treatment assignment, timeframe and number of schools and students in each arm

	Treatment 1	Treatment 2	Control 1	Control 2
April-May 2007		**Baseline**		
May-2007	Books & Training	Books	-	-
Oct-2007	-	-	Books	-
October 2007-March 2008	-	-	Training	-
Number of Schools and Students				
Schools	50	41	26	20
Students	946	784	505	377
Apr-2008		**Endline**		
May-2008	-	-	-	Books
May-Sept 2008	-	-	-	Training
Number of Schools and Students				
Schools	48	41	49	34
Students	1665	1432	1326	899

2.3 Instrument

The data collection required a significant number of survey staff. For the baseline survey, 32 people were deployed. Each survey team included three people (a team leader and two enumerators), who spent a full day in each school implementing the survey instrument (MEC and LRCM, 2008). The survey staff used measures to ensure that assessment items were appropriately translated, used transparently documented assessment procedures, including quality control procedures, and availed procedures to ensure that assessments were implemented in a standardized manner across all participating schools.

The survey instrument encompasses two sets of questionnaires: the first regarding

students and the second about teachers, classrooms and schools. Under the first instrument, students were tested in language (reading, writing and listening), numeracy skills (math), and scholastic and verbal aptitude (Peabody) using test instruments adopted from international testing standards and piloted by a team of international and local researchers. Under the second questionnaire, observation sheets for schools, classrooms and teachers were completed to collect information on school resources, classroom conditions, and teacher qualifications.

As mentioned above, five assessments were administered: a Peabody vocabulary test adapted to the Mongolian context; a mathematics assessment, based on questions from the Trends in International Mathematics and Science Study (TIMSS); and listening, reading, and writing assessments based on the Mongolian curriculum. Validation measures for the mathematics and Peabody tests were carried out under the READ project. Prior to the mathematics assessment, an investigation of construct equivalence with Grade 4 TIMSS items was undertaken. A panel of Mongolian math experts, MECS staff and an international technical assistant reviewed the TIMSS 2003 mathematics items and identified items that were suitable for Mongolia. The panel used test-curriculum matching analysis to evaluate the degree of congruence between the international mathematics assessment and the Mongolian national curriculum. Since an item might have been in the curriculum for some but not all students in the country, an item was determined appropriate if it was in the intended curriculum for more than 50 percent of the students (World Bank, 2006).

The Peabody test administered was a norm-referenced instrument for measuring the listening vocabulary of children. For each item, the assessor would say a word, and the student responded by selecting the picture that best illustrates that word's meaning. Items were reviewed by the MECS panel to ensure they were appropriate for the Mongolian curriculum. The mathematics, reading, and writing tests used a balanced incomplete block design, with different item content across different test booklets. Different test booklets were then randomly assigned to to different students. Items were grouped into blocks, and each block was repeated in more than one test booklet to ensure balance across test booklets (World Bank, 2006).

An international assessment expert hired by the project used construct equivalence analysis to confirm that the assessments measured the same constructs between boys

and girls, and the assessment frameworks applied to both genders

3 FRAMEWORK AND IDENTIFICATION STRATEGY

3.1 Conceptual Framework

Comprehensive frameworks for linking student achievement to any single education input remain elusive. For instance, the impact of an intervention that provides books to students in the third grade is a dynamic function of current and past covariates, including qualifications of current teachers, family's socioeconomic status and school attributes, as well as historical records prior to the current year (pre-school to second grade) of these covariates, and the student's performance in the previous grades. Capturing these dynamic relationships using a static framework and lacking historical data on relevant covariates makes empirical estimation of an input's impact challenging.

Moreover, the impact of an education investment, say teacher training, on a student's performance depends on the availability of other complementary inputs, like appropriate books. Whether increases in such inputs, say through in-service teacher training, matter for student outcomes is an area of ongoing research and limited clarity (Hanushek, 2004; Hanushek and Rivkin, 2010; Todd and Wolpin, 2003). The potential non-linearity in education production also suggests that returns from packaged inputs could be substantially different from the sum of returns from applying these same inputs individually (Hanushek, 2004). The complementarity of educational inputs also suggests that an individual input would have different impacts on outcomes when it is provided alone versus when it is provided in conjunction with other inputs (Linden, 2008). This is particularly pertinent in resource-poor settings, where many complementary education inputs may be missing, and availing one without the other may provide little or no return from such investment.

Lacking relevant historical covariants, this study relies on a static model of education production and also allows for the possibility of testing the complementarity of inputs. This static econometric specification of an education production function entails representing the association between a student's classroom achievement (test scores) on the one hand, and current teacher's qualifications (formal education, in-service training, experience, motivation etc.), student-specific characteristics (gender, age, appetite to

read and the like), her family's socioeconomic status (asset/income, education, housing conditions and so on) and school resources (school type, inputs, general hygiene, location, facilities etc.), on the other. Specifically, student $i's$ achievement in class c of school s ($Y_{i,c,s}$)—for the current study, scores in math, reading, writing, listening and Peabody tests—is a function of student- and family-specific characteristics ($X_{i,c,s}$); and classroom- and school-specific covariates ($R_{c,s}$) and the qualifications of her teacher, j, ($Q_{j,c,s}$). In line with previous studies (Hanushek and Rivkin, 2010; Todd and Wolpin, 2003), by assuming a static relationship, we specify a model of education production function as:

$$Y_{i,c,s} = \beta_0 + \beta_1 Q_{j,c,s} + \beta_2 X_{i,c,s} + \beta_3 R_{c,s} + \epsilon_{i,c,s} \tag{1}$$

...where $\epsilon_{i,c,s}$ is an error term.

The empirical challenge in identifying the causal impact of an educational input (or set of inputs) on student achievement is the non-randomness of input choices. For example, an in-service teacher training program that is intended to improve teachers' competence may not be attributable to a change in student test scores in a non-experimental setting because of the non-random assignment of teachers to students and teacher training opportunities to teachers. Students from families with better socioeconomic status tend to get matched with better trained, motivated and well-paid teachers, and hence teachers' qualifications tend to confound with unobserved achievement determinants (Clotfelter et al., 2006). In addition, a teacher's access to in-service training may depend on her motivation and/or personal connection with education administrator or school director (Jacob and Lefgren, 2004). This non-randomness implies that $cov(Q, \epsilon) \neq 0$ and $cov(R, \epsilon) \neq 0$—leading to biases in observation-based studies. Therefore, devising a valid identification strategy to discern the impact of improvement in teacher qualifications on test scores is important. The subsection below is devoted to discussing how the current research handles this identification challenge.

An in-service teacher training intervention presumably affects student achievement through its impact on teacher quality ($\Delta Q_{j,c,s}$). For an extensive discussion on how teacher training might improve quality by enhancing pedagogical skills as well as subject-matter understanding see Mullens et al. (1996). For instance, an experimental

evaluation of a teacher training scheme has also documented heterogeneous impacts on the teachers' own English test score, where only teachers with university degree benefited well from in-service training (Zhang et al., 2013). On the other hand, providing classroom-library materials could change the resources available in treated schools ($\Delta R_{c,s}$). This is particularly the case in a resource-poor setting, such as Mongolia, where essential teaching aids such as textbooks and workbooks were lacking.

In the context of the current study, the interventions that could improve education productivity have been randomly assigned with the intention to increase test scores in the treatment group: in-service teacher training to improve teacher quality and/or classroom libraries to ease resource scarcity. A control group of students, on the other hand, have not been exposed to any treatment until the experiment was finalized. Analysis of the baseline survey data confirms that the 'initial randomization' was properly done: there were no systematic differences between test scores (and other covariates) of students in the treatment and control groups.[5]

3.2 Identification strategy and empirical approach

The treatment effects from the above three interventions are estimated as follows: (1) *Training and Books*: To identify the causal impact of providing books and in-service training as complementary education inputs, students in treatment group one ($T1$) and control group that has not received any treatment ($C2$) were matched, and mean achievement-gaps between students in these groups provided the estimated impact of the two interventions. (2) *Extra Training*: Identification of the impact of extra in-service teacher training, on top of books, is based on the comparison of the differences in student achievement between (matched) treatment one ($T1$), which received training and books, and treatment two ($T2$), which received books (but not teacher training). It is important to note that this identification of the contribution of teacher training likely includes the returns from the joint provision of training and books as well as the contribution of each input plus any complementarity between them. In the context of rural Mongolia, where education inputs were lacking prior to READ interventions, it is most likely that education production function to exhibit increasing return for

[5]The 'initial randomization' refers to the initial randomization undertaken before the control group was divided into two, following change in policy regarding the interventions.

addition inputs. As a result, when teacher training is added to books, the increase in student outcomes is likely to improve education at least as much the effect of teacher training provided as a stand-alone intervention. More concisely, we argue that in this resource constrained setting, the education production function is likely to exhibit increasing returns to scale. Therefore, the estimated treatment effect of training should be considered as the maximum possible contribution of providing a short training to teachers, without providing complementary books.[6] (3) *Books only*: Impact of the classroom libraries intervention is identified by comparing outcomes of (matched) $T2$ against $C2$.

Due to the change in the intervention plan from the initial evaluation design halfway through the treatment period, specifically the exposure of part of the control group ($C1$) into unplanned treatment,[7] the application of the standard randomized control trial (RCT) estimation technique—through direct comparisons of differences in mean outcomes between $T1$ and $T2$ on the one hand, and the remaining control group ($C2$) on the other—is not feasible. Therefore, to ensure that the counterfactuals are properly set, and the treatment effect is consistently estimated, propensity score matching is used as an alternative identification strategy. More specifically, this approach involves two steps: (1) estimating propensity scores (PS) by matching control with treatment group of students on relevant covariates, using endline survey data only; and (2) applying a regression of student outcomes on covariates using the matched data in step one, with PS serving as weighting factor and standard error clustered at *aimag* level.

In the first step, we estimate PS. $P(X_i) = p(T_i = 1|X_i)$ is the likelihood that student i would be exposed to treatment (T_i) conditional on covariates, X_i (Rosenbaum and Rubin, 1983; Becker and Ichino, 2002)). As the number of students in treatment arm is larger than those in control group, in this step, some students that do not satisfy the matching criteria are excluded.[8] Specifically, observations off-common support are

[6]The ideal scenario would be to have another treatment group of students whose teachers were provided training alone. The comparison of test scores of these students with a control group that has not received any treatment could have provided the impact of training only, which is anticipated to be lower than the treatment effect estimated using the above setting.

[7]Part of the control group was given the books and shelves ahead of schedule because of community demand.

[8]The typical application of propensity score matching method is when there are larger number of observations in the control group to be matched with fewer observations in the treatment group. In this case, we have many more treated than control students. Therefore, each student in control group

excluded, and for both groups students with PS in the top and bottom 1% are trimmed off.

In the second step, we use the matched dataset to estimate the treatment effects (of in-service teacher training and books as a package, extra in-service training on top of books, and books alone) by running the following regressions:

$$Training\ \&\ Books: \ Y_{i,c,s} = \alpha_0 + \alpha_1 T1\_C2j, c, s + \alpha_2 Q_{j,c,s} + \alpha_3 X_{i,c,s} + \alpha_4 R_{c,s} + e_{i,c,s} \quad (2)$$

$$Training: \quad Y_{i,c,s} = \gamma_0 + \gamma_1 T1\_T2_{j,c,s} + \gamma_2 Q_{j,c,s} + \gamma_3 X_{i,c,s} + \gamma_4 R_{c,s} + \epsilon_{i,c,s} \quad (3)$$

$$Books: \quad Y_{i,c,s} = \omega_0 + \omega_1 T2\_C2_{c,s} + \omega_2 Q_{j,c,s} + \omega_3 X_{i,c,s} + \omega_4 R_{c,s} + u_{i,c,s} \quad (4)$$

...where $e_{i,c,s}$, $\epsilon_{i,c,s}$ and $u_{i,c,s}$ are error terms. T1_C2, T1_T2 and T2_C2 are dummy variables indicating whether student i is in one or the other group. For instance, T1_C2 is equal to one if she is in group T1 and zero if she is in group C2. Estimated coefficients of the corresponding these dummies (i.e. $\hat{\alpha}_1$, $\hat{\gamma}_1$ and $\hat{\omega}_1$) are the impacts of the respective intervention(s) on students' test scores in five areas.

The empirical estimation of these equations is conducted by using PS as a probability weight. The standard errors are clustered at *aimag* level—allowing heteroskedasticity and within-cluster error correlation—to account for the fact that each *aimag* has either treatment or control schools, which might create within-group dependence. In addition, there are few clusters in each group of interventions and hence the large sample property of cluster standard error might not be satisfied. Accordingly, we resort to "wild cluster bootstraping" for asymptotic refinement (see Cameron et al. (2008)).

As a robustness check, we also combine all groups of students—those who received training and books (T1), books only (T2), and control group (C2)—and estimate the follow equation:

$$Y_{i,c,s} = \delta_0 + \delta_1 T1_{j,c,s} + \delta_2 T2_{j,c,s} + \delta_3 Q_{j,c,s} + \delta_5 R_{c,s} + \delta_4 X_{i,c,s} + e_{i,c,s} \quad (5)$$

is matched with one student in treatment group, and logistic distribution is assumed.

...where T1 and T2 are dummies equal to one for groups that received training and books and books only (and zero otherwise), respectively. The coefficients (δ_1 and δ_2) are the corresponding impacts of each set of interventions. The impact of extra-training is estimated as the difference between these coefficients (i.e., $\delta_1 - \delta_2$), and test for statistical significance of this difference is conducted.

4 RESULTS

4.1 Descriptive results

In this section, we briefly discuss the implementation of propensity score matching and discuss descriptive results. As described above, the control and the treatment group of students were matched using endline survey. Observations that happen to be off-common support were dropped, and the data is further trimmed by removing observations with probabilities in the top and bottom 1% for the corresponding group. The densities of propensity scores are resented in Figure A.2 (see annex). The matching results for the three groups of interventions are presented in Table 2 below, and Table A.1-A.2 (see annex). As Table 2, A.1 and A.2 show, there were statistically significant mean differences in some covariates before matching, and these systematic differences have been addressed after matching (i.e. balancing is achieved). In other words, the factors that could attenuate or amplify the impact of the interventions, such as students' characteristics, their families' socioeconomic status, teachers' qualification and school resources, do not exhibit statistically significant differences between the treatment and control groups.

In addition, we present mean student outcomes after matching (both during baseline and endline surveys) in Table A.3 (see annex). The (matched) treatment and control groups, for all of the three interventions, do not exhibit systematic baseline differences in outcome indicators. Below is a brief discussion for each intervention.

Training and Books: Before matching, half of the covariates that could potentially affect test scores had statistically significant mean differences between the control and the treatment group of students. The PSM has taken care of these differences in these covariates. Teachers' qualifications (formal training and years of experience) are similar—the majority of teachers have had formal education and about 14 years of

professional experience. Students' book ownership at home, the difference in which could bias the impact of books provided at school, was similar for both groups during the baseline and follow-up. The same holds true for students' characteristics (age, gender, frequency of taking extra lessons per week, number of days per week in which the students had to accomplish household chores before and after work, distance from school, and residing with mother/father or others), and their families' socioeconomic status (whether both or either parent have completed high school education, residence type and ownership of telephone at home). The treatment and control schools also had similar characteristics in terms of the existence of infrastructure like toilet/hand-washing facilities (Table 2).

A detailed description of the achievement difference between treatment and control groups of students is presented in Table A.3 (see annex). The baseline achievement gap between treatment and control groups of students is not appreciable.[9] For instance, mean total score in the five tests for students in the treatment and control groups are 24.8 and 25.3, respectively. Similar results holds true when tests are considered individually. During the follow-up survey, the mean of the total score for the treatment and control groups increased to 31.9 and 29.2, respectively. The scores in individual tests have also exhibited a similar widening gap between students in treatment and control groups.

[9]The baseline data is not included in analytic results, and we are presenting it as a descriptive information only.

Table 2: Mean values of covariates and t-test for mean-difference (before and *after* matching), for books and training (April 2008)

Variable	**Control** Unmatched (Matched)	**Treated** Unmatched (Matched)	**%bias**	**% reduct bias**	**t-test** **t**	**p>t**
Gender (=1 for boys)	0.52	0.50	3.7		0.83	0.41
	(0.52)	(0.53)	(-1.3)	63.3	(-0.26)	(0.80)
Age	10.37	10.39	-2		-0.45	0.65
	(10.37)	(10.34)	(3.2)	-62.8	(0.60)	(0.55)
Number of books at home	2.16	2.17	-1.4		-0.32	0.75
	(2.16)	(2.10)	(6.8)	-373.1	(1.32)	(0.19)
Extra lesson (frequency)	2.35	2.27	7.7		1.77	0.08*
	(2.35)	(2.31)	(3.8)	50.1	(0.74)	(0.46)
Chores before school (frequency)	2.92	2.94	-1.6		-0.36	0.72
	(2.92)	(2.98)	(-5.6)	-253	(-1.10)	(0.27)
Chores after school (frequency)	2.95	2.98	-2.7		-0.61	0.54
	(2.95)	(2.94)	(1.6)	40.6	(0.30)	(0.76)
Reside far from school	0.03	0.04	-0.3		-0.07	0.94
	(0.03)	(0.03)	(4.4)	-1279.8	(0.90)	(0.37)
HH size	5.44	5.09	21.8		5.11	0.00***
	(5.44)	(5.39)	(3.2)	85.3	(0.59)	(0.55)
Living arrangement	0.56	0.52	7.5		1.69	0.09*
	(0.56)	(0.52)	(8.3)	-11.7	(1.61)	(0.11)
Residence type	1.85	1.74	11.5		2.64	0.01**
	(1.85)	(1.89)	(-5.3)	54.4	(-0.97)	(0.33)
Telephone at home	0.50	0.64	-30.2		-6.91	0.00**
	(0.50)	(0.48)	(4.1)	86.4	(0.78)	(0.44)
Mother/father has secondary edu	0.50	0.47	6		1.36	0.17
	(0.50)	(0.47)	(5.6)	5.9	(1.09)	(0.28)
Teacher's rank	2.82	2.90	-9.4		-2.21	0.03**
	(2.82)	(2.85)	(-3.4)	63.6	(-0.67)	(0.50)
Teacher's experience (year)	14.26	15.13	-8.1		-1.90	0.06*
	(14.26)	(13.75)	(4.8)	40.9	(0.95)	(0.34)
Hand washing facility exists	0.78	0.83	-10.5		-2.43	0.02**
	(0.78)	(0.79)	(-2.4)	77.5	(-0.44)	(0.66)
School has toilet	0.63	0.44	38.6		8.73	0.00***
	(0.63)	(0.63)	(-1.4)	96.5	(-0.27)	(0.79)

Note: Living arrangement refer to whether the child resides with his mother and/or father, grandparents, other relatives or school dormitory. Residence type includes 'ger', house, apartment or school dormitory. Chore frequency refers to number of days per week the child has to do household chores before/after school.

Extra Training: For the groups of students that received extra training, on top of books, similar matching results are presented in Table A.1 in the annex. After matching, the covariates that could influence test scores, such as students' and their families' characteristics, teachers' qualifications, and school features also did not differ significantly between the treatment and control groups. In addition, the baseline test scores are generally equivalent among the treated and control groups of students, with a mean total score of 25.2 and 24.0, respectively. Scores on individual tests are also comparable. During the endline survey, students in both groups improved their total mean score, but there is no pronounced widening of the gap in the mean score between treatment and control groups (Table A.3).

Books only: For this intervention, after matching, there was no systematic difference in teachers' qualifications, students' and their families' characteristics as well as school conditions between treatment and control groups (see Table A.2). Similarly, there was no systematic difference between control and treatment groups at baseline in terms of achievement. The mean of total test scores for treatment and control groups of students were 23.6 and 24.6 points, respectively. Baseline scores on individual tests are also similar across the two groups. The means of total scores on the follow-up tests for treatment and control students were 31.6 and 29.0 points, respectively (Table A.3).

4.2 Analytic results

This section presents estimated treatment effects using the empirical approach outlined in subsection 3.2. In the subsequent section, we assess heterogeneity in treatment effects, and also present robustness checks by re-estimating ATEs under different specifications. The results show that when in-service teacher training and books are are provided individually, they weakly improve test scores on some, though not all, subjects. However, when teachers are trained and students are provided with the necessary books to facilitate the implementation of knowledge acquired during training, test scores improve considerably. The impact of each intervention is discussed below.

Training and Books Intervention: For the group of students who accessed books through classroom libraries and whose teachers participated in training, test scores on almost all tests improved substantially. Table 3 presents ATE on individual test scores as well as on the total score. The total test score increased by equivalent to 34.9 percent

of a standard deviation. As shown in Figure 2 (panel A and B), the kernel densities of the treatment and control groups generally overlap during the baseline survey. During the endline survey, the mean test score of the treatment group of students was higher than that of the control group. Considering each test individually, the intervention improved writing and math test scores the most (by 27.1 and 25.9 percent of standard deviation, respectively). Reading and Peabody test scores increased, respectively, by 25.6 and 20.9 percent of standard deviation, respectively.[10] The interventions did not improve performance on listening test.

Table 3 Impact of teacher training and books on test score

	(1) Peabody	(2) Math	(3) Listening	(4) Reading	(5) Writing	(6) Total Score
ATE	0.481***	0.617***	0.225	0.989***	0.555**	2.867***
	(0.00)	(0.00)	(0.10)	(0.00)	(0.02)	(0.00)
N	2424	2424	2424	2424	2424	2424

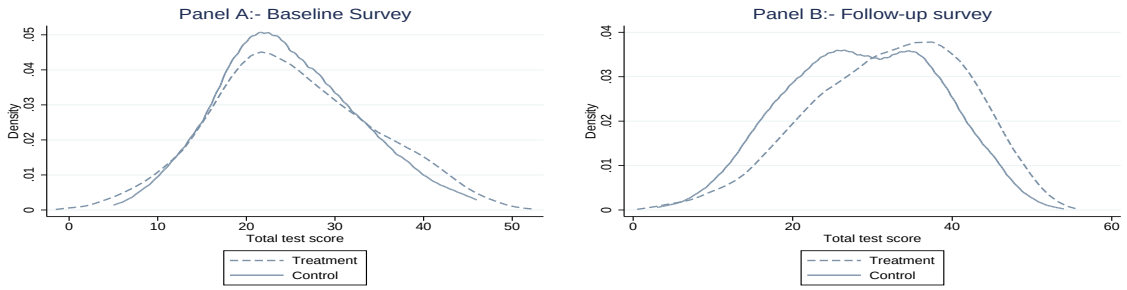
b coefficients; p in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The standard errors are clustered at *aimag* level, with "wild cluster boostraping" (Cameron et al., 2008). The matching variables are characteristics of the student, household, teacher and school. Student's characteristics includes gender, age, number of books she owns, frequency of extra lessons after school, frequency of accomplishing household chores before and after school and a dummy for residing more than one hour walk from school and commuting by foot. Characteristics of the student's household encompasses household size, the household head's relationship with the child, wealth indicators (such as housing condition, dummy for phone and car ownership) and education level. Teacher's formal education and years of experience makeup characteristics of the student's current teacher. Teacher's formal education and years of experience makeup characteristics of the student's current teacher. School characteristics encompasses dummy for existence of hygiene infrastructure (like toilet and hand-washing facilities).

[10]For training and books intervention group of students, the standard deviations of test scores in Peabody, math, listening, reading, writing and total score are 2.30, 2.38, 1.93, 3.85, 2.05 and 8.19, respectively (see Table A.3 in the annex).

Figure 2: Density of total test score for teacher training and books intervention



Note: Total test score is the sum of scores in math, reading, writing, listening and Peabody tests.

Extra Teacher Training: As discussed above, the comparison of test scores of students who were treated with in-service teacher training and books against those who received books only is the basis of estimating ATE of teacher training only. This comparison shows that the extra teacher training has weaker impacts on test scores. Due to the extra teacher training intervention, total test scores did not improve (Table 4). Figure 3 presents the kernel density of total test score, which reveals a similar result: both the treatment (training and books receivers) and control (books only receivers) groups of students performed similarly during the baseline and follow-up—even if the mean of total score improved during the follow-up survey for both groups. Out of the five tests, only score in writing has improved by 15.3 percent of a standard deviation, and this is smaller in magnitude when compared to training, complemented with books.[11] No impact on Peapody, math, reading and listening test scores due to the extra in-service teacher training was found.[12]

These findings lie at the heart of the contentious literature on effectiveness of brief in-service teacher training schemes in improving test scores. Some previous studies find training improved test score, other do not. For instance, Jacob and Lefgren (2004), employing a quasi-experimental method based on the school reform program in Chicago, established that in-service teacher training had no statistically significant or academically meaningful impact on reading and math achievement of students in elementary school. Similarly, Zhang et al. (2013) undertook a randomized control trial and documented that short-term in-service teacher training in Beijing's migrant

[11] As discussed in the 'identification strategy' section, the impact training only is likely to be overestimated as it might also include any complementarity effects between these inputs.

[12] For students in training only intervention group, the standard deviations of test scores in Peabody, math, listening, reading, writing and total score are 2.14, 2.58, 2.23, 4.15, 2.10 and 8.55, respectively.

schools did not improve scores in an English proficiency test. Using observational data from rural primary schools of Thailand, teachers' exposure to in-service training has been shown not to predict instructional quality or student achievement in Thai language, math, social and natural studies, character development and work orientation tests (Raudenbush et al., 1993). However, others find that teacher training enhances students' performance in these subjects. For instance, Angrist and Lavy (2001) documented that in-service training has had a significant impact on students' achievement in math and reading in non-religious elementary schools in Jerusalem, whereas the impact on the achievement of students in religious schools was inconclusive. Similarly, Harris and Sass (2011) and Lai et al. (2011) found that teachers' qualifications and on the job training improve student outcomes. These results from previous studies are consistent with the findings of this study—extra teacher training, on top of books, weakly improves test score in some subjects. However, when training is provided along with appropriate books, it strongly improves student outcomes.

After all, the circumstances under which training becomes effective could be diverse. Among other factors, whether the teachers have the necessary teaching aids to implement any pedagogical technique they acquire from training could be crucial. Especially in countries where essential education inputs may be missing, in-service teacher training could render ineffective. In fact, as we have documented above, when training is combined with book provision, test scores in most subjects improve substantially.

Table 4: Impact of extra teacher training, on top of books, on test score

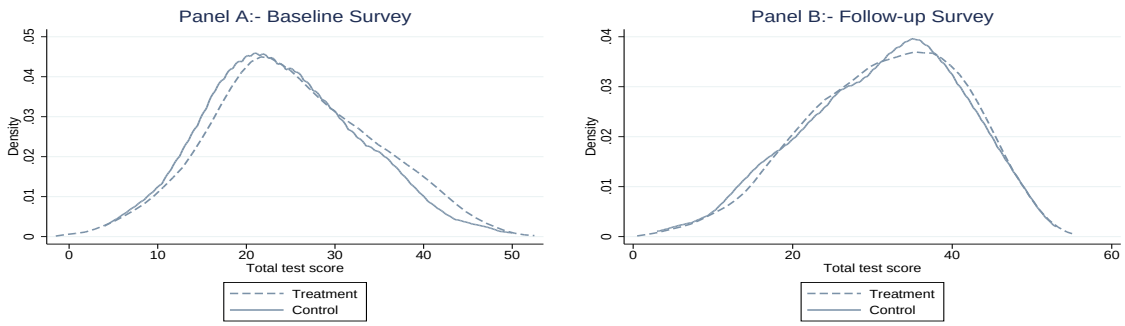
	(1) Peabody	(2) Math	(3) Listening	(4) Reading	(5) Writing	(6) Total Score
ATE	0.243	-0.0563	-0.0491	-0.229	0.321*	0.229
	(0.38)	(0.84)	(0.72)	(0.48)	(0.08)	(0.88)
N	2968	2968	2968	2968	2968	2968

b coefficients; p in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The matching variables are characteristics of the student, household, teacher and school. For the full list of covariates see Table A.1 in the annex.

Figure 3: Density of total test score for teacher training only intervention



Note: Total test score is the sum of math, reading, writing, listening and Peabody test scores.

Books only: Providing books had a strong impact on test scores. Books alone greatly increased test scores more than teacher extra training, but the books intervention still had a much weaker impact than training and books provided as a package. It improved scores in many more subject tests. For instance, it increased the total score by 20.6 percent of a standard deviation (Table 5). The density of the total test score for the treatment and control group of students exhibits a mildly stronger shift in mean score among the treated groups of students (Figure 4). The intervention improved the scores in two of the five tests. It increased scores in reading and math tests by 22.2 and 25 percent of standard deviation, respectively. These improvements in test scores due to book provision are lower than the impacts under the joint provision of training and books.[13]

The findings that books improve test scores in some subjects, even when provided alone, is in line with a general narrative provided in the systemic review by Glewwe et al. (2013): when considering all the evidences holistically, textbooks and workbooks improve weakly learning outcomes. In addition, we find that the return from the provision of books increases when it is jointly provided with teacher training. The latter result, along with the fact that training also works better when provided along with books, is evidence of the complementarity of education inputs.

[13]For group of students in books-only intervention, the standard deviations of test scores in Peabody, math, listening, reading, writing and total score are 2.39, 2.36, 1.80, 3.93, 2.02 and 8.13, respectively.

Table 5: Impact of books only on test score

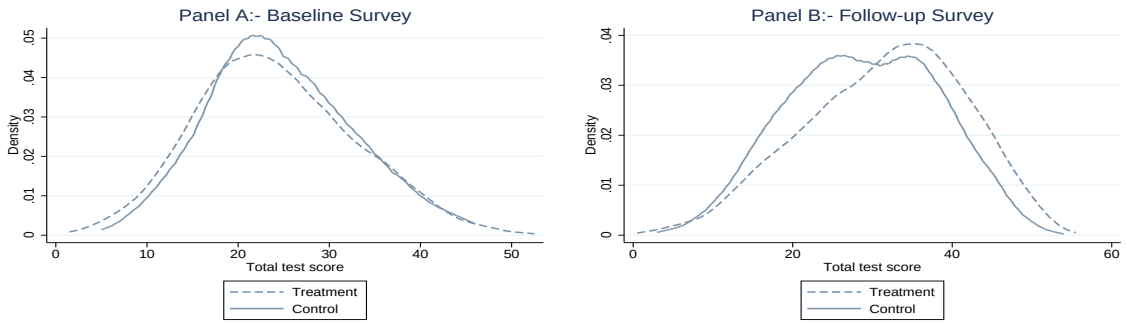
	(1) Peabody	(2) Math	(3) Listening	(4) Reading	(5) Writing	(6) Total Score
ATE	-0.105	0.525**	0.186	0.982***	0.124	1.712*
	(0.78)	(0.02)	(0.20)	(0.00)	(0.72)	(0.08)
N	2111	2111	2111	2111	2111	2111

b coefficients; p in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The matching variables are characteristics of the student, household, teacher and school. For the full list of covariates see Table A.2 in the annex

Figure 4: Density of total test score for books only intervention



Note: Total test score is the sum of math, reading, writing, listening and Peabody test scores

4.3 Heterogeneity in treatment effects

This subsection investigates any heterogeneity in treatment effects using three subsamples of students, based on their gender, access to extra lessons, and parental education. On the bases of each of the above characteristics, the sample was divided into two subgroups: students who have taken at least one extra lesson per week versus those who did not; students whose either (or both) parent have completed secondary education against those whose parents have not completed high school; and boys or girls. It is reasonable to expect that students who have taken extra lessons or have educated parents could benefit differently from these interventions.

For students who did not have access to extra lessons, provision of these inputs, either individually or as a package, improved their performance meaningfully. Especially, books only and training and books as a package increased the test score of this group. On the other hand, students that have taken extra lessons outside school have performed better in some subjects when they were treated with these interventions. However, the overall improvements in the performance of this group is relatively smaller than those students who did not have access to extra lesson (see Table A.4, annex).

Returning to parental education, we find that students whose parents have not completed secondary education have benefited from books and training, and books only interventions more than those with educated parents. In addition, these students improved their performance more when books and training were provided together. Moreover, training teachers does not seem to help students with less educated parents and educationed parents alike (Table A.5). In terms of the student's gender, there are differences in treatment effects of the three interventions. The provision of packaged inputs (training and books) improved girls' score more than boys. But books alone do not seem to improve girls' test scores significantly (Table A.6). The general message from these results is that providing packaged inputs helps groups of students who might be disadvantaged (i.e. those who do not have access to extra-lesson sessions, with less educated parents, and girls).

4.4 Robustness check

In this section, we check the robustness of the results presented in the preceding subsection by re-estimating the impacts of each intervention under different specifications. To assess how the estimated impacts could change with changes in matching variables, the propensity score matching estimation is implemented by progressively including characteristics of students, their families, teachers and schools in four specifications. In addition, we estimate the treatment effect on the total test score by matching on all possible combinations of covariates (by adding and dropping regressors), while including the students' characteristics as 'core variables' in all the regressions. Despite the limitations of using this method (see Lu and White (2014)), this provides reasonable checks as to whether the treatment effect is appropriately estimated. Table A.7 (in the annex) presents the average treatment effects (ATEs), for the three interventions, with various sets of matching variables. In specification 1, we present ATEs by

matching students based on their own characteristics only. In subsequent specifications, we progressively include characteristics of their families, teachers and their schools' resources. The results, in general, support the main findings—teacher training provided along with teaching aids improves test scores substantially, while the interventions implemented individually have weak impacts and improve scores only in some subjects.

In addition, we estimate the treatment effects by pooling the three groups together and estimating Equation 5. The result, presented in Table 6, is consistent with main result. It shows that inputs provided as a package improve test scores significantly, relative to isolated input provision. In this approach, we find that teacher training has no effect on all test scores (even on writing, which was statistically significant in the main specification).

Table 6 Impact of teacher training and books, and books only on test score

	(1) Peabody	(2) Math	(3) Listening	(4) Reading	(5) Writing	(6) Total Score
Training and Books	0.557**	0.772***	0.268	1.210***	0.614**	3.420***
	(0.04)	(0.00)	(0.12)	(0.00)	(0.04)	(0.00)
Books only	0.0335	0.578***	0.176	1.048***	0.145	1.980*
	(1.00)	(0.00)	(0.34)	(0.00)	(0.74)	(0.00)
Extra-training*	.523	.194	.092	.162	.469	1.44
	[0.221]	[0.745	[0.737]	[0.87]	[0.181]	[0.952]
N	5038	5038	5038	5038	5038	5038

Note: The standard errors are clustered at *aimag* level, with "wild cluster boostraping" (Cameron et al., 2008). P-values in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The matching variables are those used in the main results.
*The impact of extra-training is calculated using post-estimation test for the difference between coefficients of training and books and books only estimations. P-values of the chi-squared test for the differences are in brackets.

5 CONCLUSION

Policy makers around the world are keenly interested in the potential of in-service teacher training programs and the provision of high-quality learning materials to help improve schooling outcomes. Surprisingly few evaluations have used a randomized controlled trial approach to examine the impacts of introducing these types

of interventions—either individually or jointly—in developing countries. Limited conclusive evidence exists about the impact of these interventions on primary school programs, and most of this evidence comes from small pilot projects. Even less evidence is available regarding their impact as part of a nationwide education program.

This work fills a gap in the literature. While other studies have provided inconclusive evidence as to the impact of teacher training or book provision on student outcomes when inputs are provided individually, no previous work has attempted to explore the differential impact of providing these two critical education inputs individually versus jointly to test for any input complementarity in education investments. This study thus provides interesting, new, and important insights. The evaluation found significant, positive effects on student outcomes when books and training were provided together as a package, rather than as individual inputs. Books only and extra teacher training marginally improved test scores in some, but not all, subjects. The magnitude of impact of either input was not academically significant. However, when teachers are trained and students are provided with books, the test scores of a treatment group of students increased substantially, relative to a control group of students.

The findings from this study provide information to education policy makers in developing countries on how their input allocation choices could result in significantly different outcomes. Isolated education investments in settings where complementary inputs are missing could deliver minimal or no return. On the other hand, coordinated investments could improve student outcomes substantially, beyond and above the sum of returns from the same investments undertaken individually. These coordinated interventions are very cost effective. Equipping a classroom with 160 books and a set of shelves costs only $353.5 (in 2008 US$). Similarly, as noted above, the cost of training teachers was relatively low. This makes the cost of these joint interventions per student substantially lower.

To inform the design and implementation of future teacher training and book provision schemes, other research should focus on exploring the impacts of providing packaged inputs versus isolated inputs in settings with different levels of resource availability (classroom, school, household, and region). It may be likely that heterogeneity in treatment effects based on the existence of complementary school- and household-resources will prevail, while the result may not hold in areas where a

reasonable amount of education resources are already in place. Additional work should also investigate the impact of different types of teacher training programs, including methods, pedagogical strategies, and rollout of these interventions, on test scores. Detailing these outcomes would have significant implications for policy makers with limited resources who are seeking improved efficiency and better student outcomes.

References

Angrist, J. D., Lavy, V., 2001. Does teacher training affect pupil learning? evidence from matched comparisons in jerusalem public schools. Journal of Labor Economics 19 (2), 343–369.

Becker, S. O., Ichino, A., 2002. Estimation of average treatment effects based on propensity scores. Stata Journal 2 (4), 358–377.

Bunyi, G. W., Wangia, J., Magoma, C. M., Limboro, C. M., 2013. Teacher preparation and continuing professional development in kenya: Learning to teach early reading and mathematics.

Cameron, A. C., Gelbach, J. B., Miller, D. L., 2008. Bootstrap-based improvements for inference with clustered errors. The Review of Economics and Statistics 9 (3), 414–42.

Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., 2006. Teacher-student matching and the assessment of teacher effectiveness. The Journal of Human Resources 41 (4), 778–820.

Conn, K. M., 2014. Identifying effective education interventions in sub-saharan africa: A meta-analysis of rigorous impact evaluations.

Evans, D. K., Popova, A., 2014. What works to improve learning in developing countries? an analysis of divergent findings in systematic reviews.

GHIN, 2011. Mongolia: Provincial boundaries.
URL http://ghin.pdc.org/mde/

Glewwe, P., Kremer, M., Moulin, S., 1998. Textbooks and test scores: Evidence from a prospective evaluation in kenya.

Glewwe, P., Kremer, M., Moulin, S., 2009. Many children left behind? textbooks and test scores in kenya. American Economic Journal: Applied Economics 1 (1), 112–135.

Glewwe, P. W., Hanushek, E. A., Humpage, S. D., Ravina, R., 2013. School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010. Education Policy in Developing Countries, pp. 13–64.

GOM, G. o. M., 2007. Millennium development goals based comprehensive national development strategy of mongolia.

Hanushek, E. A., 2004. What if there are no 'best practices'? Scottish Journal of Political Economy 51 (2), 156–172.

Hanushek, E. A., Rivkin, S. G., 2010. Generalizations about using value-added measures of teacher quality. The American Economic Review 100 (2), 267–271.

Harris, D. N., Sass, T. R., 2011. Teacher training, teacher quality and student achievement. Journal of Public Economics 95 (7), 798–812.

Heyneman, S. P., Jamison, D. T., Montenegro, X., 1984. Textbooks in the philippines: Evaluatin of the pedagogical impact of a nationwide investment. Educational Evaluation and Policy Analysis 6 (2), 139–150.

Jacob, B. A., Lefgren, L., 2004. The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in chicago. The Journal of Human Resources 39 (1), 50–79.

Jamison, D. T., Searle, B., Galda, K., Heyneman, S. P., 1981. Improving elementary mathematics education in nicaragua: An experimental study of the impact of textbooks and radio on achievement. Journal of Educational Psychology 73 (4), 556–567.

Kidwai, H., Burnette, D., Rao, S., Nath, S., Bajaj, M., Bajpai, N., 2013. In-service teacher training for public primary schools in rural india: Findings from district morigaon (assam) and district medak (andhra pradesh).

Lai, F., Sadoulet, E., Janvry, A. d., 2011. The contributions of school quality and teacher qualifications to student performance: Evidence from a natural experiment in beijing middle schools. Journal of Human Resources 46 (1), 123–153.

Linden, L. L., 2008. Complement or substitute? the effect of technology on student achievement in india.

Lu, X., White, H., 2014. Robustness checks and robustness tests in applied economics. Journal of Econometrics 178, Part 1, 194–206.

McEwan, P. J., 2014. Improving learning in primary schools of developing countries a meta-analysis of randomized experiments. Review of Educational Research.

MEC, LRCM, 2008. Follow-up survey for READ project: Some results of the survey.

Mullens, J. E., Murnane, R. J., Willett, J. B., 1996. The contribution of training and subject matter knowledge to teaching effectiveness: A multilevel analysis of longitudinal evidence from belize. Comparative Education Review 40 (2), 139–157.

NSO, 2006. Mongolian statistical year book 2006.

Raudenbush, S. W., Eamsukkawat, S., Di-Ibor, I., Kamali, M., Taoklam, W., 1993. On-the-job improvements in teacher competence: Policy options and their effects on teaching and learning in thailand. Educational Evaluation and Policy Analysis 15 (3), 279–297.

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.

Rothstein, J., 2010. Teacher quality in educational production: Tracking, decay, and student achievement. The Quarterly Journal of Economics 125 (1), 175–214.

Sabarwal, S., Marshak, A., Evans, D. K., 2014. The permanent input hypothesis : the case of textbooks and (no) student learning in sierra leone.

Todd, P. E., Wolpin, K. I., 2003. On the specification and estimation of the production function for cognitive achievement. The Economic Journal 113 (485), 3–33.

World Bank, W., 2006. Mongolia: Rural education and development project, project files, client connection.

World Bank, W., 2013. Implementation completion and results report: Rural education and development project.

Yang, A., Sato, Y., 2009. Secondary education regional information base, country profile mongolia.

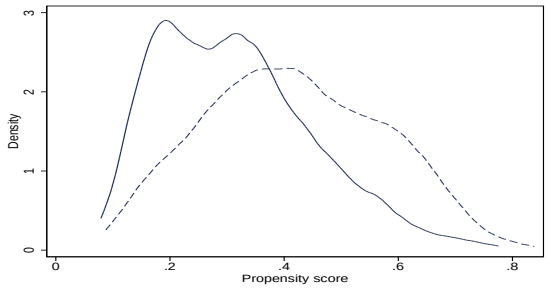
Zhang, L., Lai, F., Pang, X., Yi, H., Rozelle, S., 2013. The impact of teacher training on teacher and student outcomes: evidence from a randomised experiment in beijing migrant schools. Journal of Development Effectiveness 5 (3), 339–358.

6 Annex

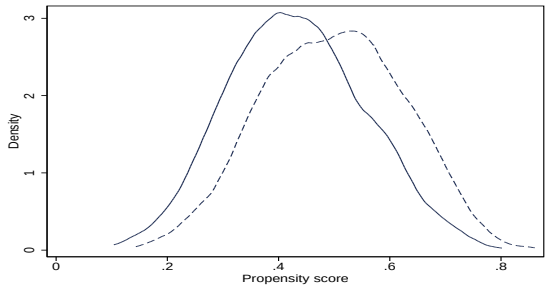
Figure A.1: Provinces with treatment and control schools

Note: Boundary coordinates of provinces are taken from United Nations Office for the Coordination of Humanitarian Affairs (cited in: GHIN (2011)).[14]

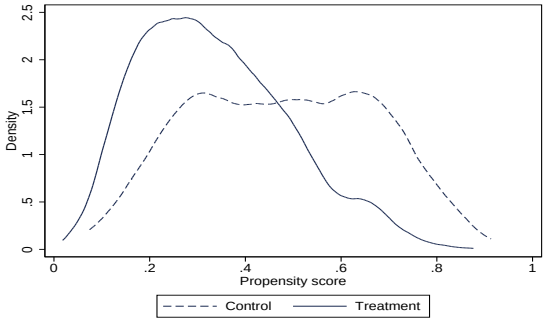
Figure A.2: Density of propensity scores from matching of treatment and control groups (endline survey), observation off- and on-common support



(a) Books and Training



(b) Training



(c) Books

Note: Observation off-support were excluded. Further, observations with propensity score in the top and bottom 1% were trimmed-off/excluded.

Table A.1: Mean values of covariates and t-test for mean-difference (before and *after* matching), for extra teacher training (April 2008)

Variable	**Control** Unmatched *Matched*	**Treated** Unmatched *Matched*	**%bias**	**% reduct bias**	**t-test** **t**	**p>t**
Gender (=1 for boys)	0.51	0.50	1.7		0.46	0.65
	(0.51)	(0.49)	(4.4)	-164	(1.15)	(0.25)
Age	10.67	10.39	30.4		8.32	0.00***
	(10.67)	(10.70)	(-2.8)	91	(-0.73)	(0.47)
Number of books at home	2.26	2.17	9.2		2.52	0.01**
	(2.26)	(2.25)	(0.8)	91	(0.22)	(0.83)
Extra lesson (frequency)	2.42	2.27	14.1		3.87	0.00***
	(2.42)	(2.39)	(2.2)	84	(0.58)	(0.56)
Chores before school (frequency)	2.95	2.94	0.6		0.17	0.87
	(2.95)	(2.94)	(0.8)	-25	(0.20)	(0.84)
Chores after school (frequency)	2.99	2.98	0.9		0.24	0.81
	(2.99)	(3.00)	(-1.3)	-49	(-0.34)	(0.73)
Reside far from school	0.03	0.04	-2.6		-0.71	0.48
	(0.03)	(0.03)	(0)	100	(0.00)	(1.00)
HH size	5.14	5.09	3.2		0.86	0.39
	(5.14)	(5.14)	(-0.2)	93	(-0.05)	(0.96)
Living arrangement	0.53	0.52	1.6		0.44	0.44
	(0.53)	(0.56)	(-5.1)	-220	(-1.34)	(0.18)
Residence type	1.66	1.74	-8.4		-2.31	0.02**
	(1.66)	(1.65)	(1.1)	87	(0.28)	(0.78)
Telephone at home	0.61	0.64	-7.4		-2.03	0.04
	(0.61)	(0.61)	(-0.9)	88	(-0.24)	(0.81)
Family owns car	0.39	0.40	-2.2		-0.60	0.55
	(0.39)	(0.39)	(-0.3)	86	(-0.08)	(0.94)
Mother/father has se u	0.48	0.47	3.5		0.96	0.34
	(0.48)	(0.51)	(-5)	-43	(-1.30)	(0.19)
Teacher's experience (year)	17.11	15.13	20.9		5.71	0.00***
	(17.11)	(16.73)	(4)	81	(1.00)	(0.32)
School yard has litter	0.02	0.06	-18.4		-4.93	0.00***
	(0.02)	(0.02)	(2.6)	86	(0.94)	(0.35)
School has toilet	0.47	0.44	7.7		2.10	0.04**
	(0.47)	(0.47)	(1.5)	81	(0.38)	(0.70)

Note: Living arrangement refer to whether the child resides with his mother and/or father, grandparents, other relatives or school dormitory. Residence type includes 'ger', house, apartment or school dormitory. Chore frequency refers to number of days per week the child has to do household chores before/after school.

Table A.2: Mean values of covariates and t-test for mean-difference (before and *after* matching), for books only (April 2008)

Variable	**Control** Unmatched *Matched*	**Treated** Unmatched textitMatched	**%bias**	**% reduct bias**	**t-test** **t**	**p>t**
Gender (=1 for boys)	0.53	0.51	2.2		0.49	0.62
	(0.53)	(0.55)	(-5)	-126	(-0.98)	(0.33)
Age	10.44	10.71	-28.8		-6.44	0.00***
	(10.44)	(10.50)	(-6.1)	79	(-1.20)	(0.23)
Number of books at home	2.19	2.27	-8.8		-1.94	0.05
	(2.19)	(2.17)	(1.5)	82	(0.31)	(0.76)
Extra lesson (frequency/week)	2.42	2.46	-3.1		-0.69	0.49
	(2.42)	(2.42)	(0.1)	96	(0.02)	(0.98)
Chores before school (frequency)	2.95	2.94	1.1		0.24	0.81
	(2.95)	(2.93)	(1.2)	-13	(0.24)	(0.81)
Chores after school (frequency)	2.98	2.98	-0.3		-0.07	0.95
	(2.98)	(2.95)	(2.7)	-827	(0.52)	(0.61)
Reside far from school	0.04	0.03	3.1		0.70	0.48
	(0.04)	(0.03)	(4.4)	-43	(0.88)	(0.38)
HH size	5.41	5.15	17.1		3.91	0.00***
	(5.41)	(5.34)	(4.3)	75	(0.82)	(0.41)
Living arrangement	0.55	0.53	4.1		0.90	0.37
	(0.55)	(0.53)	(4.8)	-18	(0.93)	(0.35)
Residence type	1.84	1.65	19.6		4.26	0.00***
	(1.84)	(1.95)	(-11.2)	43	(-1.98)	(0.05*)
Telephone at home	0.50	0.60	-20.3		-4.51	0.00***
	(0.50)	(0.46)	(8)	61	(1.54)	(0.12)
Family owns car	0.38	0.39	-3.2		-0.72	0.47
	(0.38)	(0.38)	(0)	100	(0.00)	(1.00)
Mother/father has secondary edu	0.49	0.49	1.6		0.36	0.72
	(0.49)	(0.47)	(4.7)	-194	(0.92)	(0.36)
Teacher has formal edu	0.99	0.99	-0.9		-0.20	0.85
	(0.99)	(0.99)	(4.6)	-423	(0.78)	(0.44)
Teacher's experience (year)	14.97	17.26	-21.7		-4.96	0.00***
	(14.97)	(15.17)	(-1.9)	91	(-0.38)	(0.71)
School has dormitory	0.95	0.92	10		2.16	0.03**
	(0.95)	(0.94)	(4.3)	57	(0.89)	(0.38)
School has toilet	0.60	0.47	27.1		5.99	0.00***
	(0.60)	(0.60)	(0.3)	99	(0.05)	0.96)

Table A.3: Mean test score of students in treatment and control groups by intervention, during baseline and follow-up

	Training & books				**Training**				**Books**			
	Baseline		*Endline*		*Baseline*		*Endline*		*Baseline*		*Endline*	
	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Control	Treat
Peabody	7.0	7.3	7.4	7.9	6.9	7.3	7.6	7.8	6.9	6.7	7.4	7.6
	(2.3)	(2.2)	(2.2)	(2.2)	(2.1)	(2.2)	(2.4)	(2.2)	(2.4)	(2.1)	(2.2)	(2.4)
Math	2.0	2.2	4.9	5.5	2.3	2.1	5.5	5.5	2.0	2.1	4.8	5.5
	(2.4)	(2.5)	(2.7)	(2.8)	(2.6)	(2.5)	(2.9)	(2.8)	(2.4)	(2.5)	(2.7)	(2.9)
Listening	7.1	6.7	6.9	7.1	6.6	6.6	7.2	7.1	7.1	6.5	6.9	7.2
	(1.9)	(2.2)	(2.3)	(2.2)	(2.2)	(2.2)	(2.3)	(2.2)	(1.8)	(2.3)	(2.3)	(2.3)
Reading	4.9	5.6	6.7	7.6	4.8	5.6	7.9	7.6	5.0	4.6	6.7	7.8
	(3.9)	(4.3)	(3.6)	(3.7)	(4.1)	(4.3)	(3.6)	(3.7)	(3.9)	(4.0)	(3.7)	(3.6)
Writing	3.8	3.5	3.3	3.8	3.7	3.5	3.5	3.9	3.6	3.6	3.2	3.5
	(2.0)	(2.0)	(2.1)	(1.9)	(2.1)	(2.0)	(2.0)	(1.9)	(2.0)	(2.1)	(2.1)	(2.0)
Total score	24.8	25.3	29.2	31.9	24.2	25.2	31.7	31.9	24.6	23.6	29.0	31.6
	(8.2)	(9.0)	(9.4)	(9.6)	(8.6)	(9.0)	(9.8)	(9.6)	(8.1)	(8.1)	(9.4)	(9.8)
N	303	924	795	1629	664	924	1343	1625	270	591	745	1366

Note: Standard deviations are in parentheses. The summary statistics is based on matched treatment and control groups. 'Treat' stands for treatment group.

Table A.4: Heterogeneity in treatment effects by the students' access to extra lessons

	Extra Lesson?	(1) Peabody	(2) Math	(3) Listening	(4) Reading	(5) Writing	(6) Total Score
Books and Training	**Yes**	0.487	0.688*	0.109	0.629	0.780**	2.692
		(0..20)	(0.08)	(0.76)	(0.38)	(0.04)	(0.12)
	N	543	543	543	543	543	543
	No	0.490	0.626**	0.212	1.047***	0.618***	2.994***
		(0.10)	(0.04)	(0.58)	(0.00)	(0.00)	(0.00)
	N	1797	1797	1797	1797	1797	1797
Training	**Yes**	0.0483	-0.267	-0.194	-0.239	0.0984	-0.554
		(0.78)	(0.56)	(0.58)	(0.54)	(0.68)	(0.82)
	N	635	635	635	635	635	635
	No	0.317	0.0407	0.0614	-0.162	0.423**	0.680
		(0.28)	(0.84)	(0.66)	(0.52)	(0.04)	(0.52)
	N	2252	2252	2252	2252	2252	2252
Books	**Yes**	0.297	0.995*	0.420	0.941	0.500	3.152
		(0.38)	(0.06)	(0.18)	(0.26)	(0.40)	(0.14)
	N	379	379	379	379	379	379
	No	-0.116	0.415*	0.148	1.090***	0.144	1.681**
		(0.46)	(0.06)	(0.62)	(0.00)	(0.60)	(0.04)
	N	1644	1644	1644	1644	1644	1644

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. 'Obs.' refers to number of observations. All covariates that were used for matching in the main results were employed as matching covariates in the estimation of ATEs.

Table A.5: Heterogeneity in treatment effects by parental education

	Educated Parent(s)?	(1) Peabody	(2) Math	(3) Listening	(4) Reading	(5) Writing	(6) Total Score
Books and Training	**Yes**	0.497**	0.466	0.0819	0.431	0.930**	2.406***
		(0.04)	(0.10)	(0.76)	(0.26)	(0.02)	(0.00)
	N	1252	1252	1252	1252	1252	1252
	No	0.423*	0.738***	0.299	1.452***	0.256	3.168***
		(0 .06)	(0.00)	(0.18)	(0.00)	(0.32)	(0.00)
	N	1072	1072	1072	1072	1072	1072
Training	**Yes**	0.376*	0.00328	-0.143	-0.461	0.350	0.125
		(0.06)	(0.84)	(0.42)	(0.10)	(0.18)	(0.82)
	N	1493	1493	1493	1493	1493	1493
	No	0.0997	-0.0774	0.0575	0.0115	0.237	0.328
		(0.72)	(0.66)	(0.84)	(0.90)	(0.14)	(0.82)
	N	1382	1382	1382	1382	1382	1382
Books	**Yes**	-0.292	0.402	0.110	0.745*	0.404	1.369
		(0.30)	(0.16)	(0.78)	(0.06)	(0.32)	(0.22)
	N	1037	1037	1037	1037	1037	1037
	No	0.0889	0.675 **	0.238	0.950 **	-0.0489	1.903
		(0.72)	(0.04)	(0.38)	(0.02)	(0.86)	(0.18)
	N	925	925	925	925	925	925

Note: :P-values in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Parental education refers to whether either/both parents have completed secondary education or not. 'Obs.' refers to number of observations. All covariates that were used for matching in the main results were employed as matching covariates in the estimation of ATEs.

Table A.6: Heterogeneity in treatment effects by gender of the student

		(1)	(2)	(3)	(4)	(5)	(6)
	Gender	Peabody	Math	Listening	Reading	Writing	Total Score
Books and Training	**Girls**	0.402*	0.636**	0.357	1.257***	0.707***	3.360***
		(0.06)	(0.02)	(0.12)	(0.00)	(0.00)	(0.00)
	N	1126	1126	1126	1126	1126	1126
	Boys	0.439**	0.671**	0.0933	0.654**	0.586***	2.443***
		(0.02)	(0.02)	(0.58)	(0.02)	(0.00)	(0.00)
	N	1207	1207	1207	1207	1207	1207
Training	**Girls**	0.160	-0.162	0.00960	-0.102	0.367**	0.273
		(0.58)	(0.52)	(0.82)	(0.70)	(0.04)	(0.72)
	N	1398	1398	1398	1398	1398	1398
	Boys	0.283	0.00146	-0.140	-0.306	0.325	0.164
		(0.30)	(1.00)	(0.48)	(0.28)	(0.24)	(0.96)
	N	1482	1482	1482	1482	1482	1482
Books	**Girls**	0.0202	0.393	0.146	0.957**	0.0858	1.601
		(1.00)	(0.22)	(0.50)	(.04)	(0.74)	(0.16)
	N	963	963	963	963	963	963
	Boys	-0.128	0.559**	0.198	0.930***	0.249	1.807*
		(0.68)	(0.02)	(0.38)	(0.00)	(0.52)	(0.08)
	N	1026	1026	1026	1026	1026	1026

Note: P-values in parentheses.

Table A.7: Estimated ATE of each intervention for different specifications

	Training and books				*Training*				*Books*			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Peabody	0.599***	0.495**	0.483**	0.481***	0.218	0.200	0.229	0.243	0.0632	0.00891	-0.0140	-0.105
	(0.00)	(0.02)	(0.02)	(0.00)	(0.46)	(0.52)	(0.40)	(0.38)	(0.60)	(0.78)	(1.00)	(0.78)
Math	0.674**	0.588**	0.584**	0.617***	-0.0690	-0.0899	-0.0744	-0.0563	0.663***	0.589***	0.540**	0.525**
	(0.02)	(0.02)	(0.02)	(0.00)	(0.76)	(0.72)	(0.74)	(0.84)	(0.00)	(0.00)	(0.02)	(0.02)
Listening	0.223	0.185	0.174	0.225	-0.0608	-0.0772	-0.0593	-0.0491	0.254	0.218	0.193	0.186
	(0.12)	(0.20)	(0..20)	(0.10)	(0.68)	(0.60)	(0.66)	(0.72)	(.12)	(0.20)	(0.20)	(0.20)
Reading	1.060***	0.944***	0.944***	0.989***	-0.282	-0.310	-0.264	-0.229	1.215***	1.100***	1.015***	0.982***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.44)	(0.32)	(0.44)	(0 .48)	(0.00)	(0.00)	(0.00)	(0.00)
Writing	0.473*	0.482**	0.487**	0.555**	0.288	0.271	0.293	0.321*	0.0836	0.0781	0.0795	0.124
	(0.06)	(0.04)	(0.04)	(0.02)	(0.18)	(0.20)	(0.16)	(0.08)	(0.84)	(0.78)	(0.76)	(0.72)
Total Score	3.029***	2.693***	2.673***	2.867***	0.0947	-0.00673	0.126	0.229	2.279**	1.994**	1.814*	1.712*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.96)	(1.00)	(0.92)	(0.88)	(0.02)	(0.04)	(0.06)	(0.08)
N	2424	2424	2424	2424	2968	2968	2968	2968	2111	2111	2111	2111

Note: P-values in parentheses: *** p<0.01, ** p<0.05, * p<0.1. The table presents ATEs with different groups of matching covariates: *Specification 1-4* match (treatment and control students) by characteristics of the students only; students and households; students, households and teachers; and students, households, teachers and schools, respectively.