Are Child Centered Classrooms Effective? Impact of the CRI Program on Student Learning Outcomes in Pakistan*

Farooq Naseer† Manasa Patnam§ Reehana Raza‡

Abstract

The CRI program is a public-private partnership between the Government of Pakistan and Children Resources International (Pakistan), designed to improve teaching methodology and student performance through teacher training, interactive classroom equipment, and family literacy. It adopts an interactive teaching style with the objective of moving from a teacher-centered to a student-centered classroom environment. Since 2002, CRI has been working in 35 government schools in Islamabad. This paper uses Propensity Score Matching to evaluate the impact of the CRI program. It employs a two stage matching procedure to estimate differences in learning outcomes. We first match on school level covariates to select a comparison group of Non CRI schools and then match the children within these schools to measure the net effect of the CRI program. The methodology relies on comparing the scores of Grade four children on tests of numeric and reading ability. We find that CRI is effective in raising learning achievement. Our result stays unchanged when we account for the difficulty level of the questions in the test, using a three parameter item response model. The results are also robust to unobserved selection bias. The average gain represents an improvement of about 0.23 standard deviations.

Keywords: Public-Private Partnership; Impact evaluation; Education; Propensity Score Matching

JEL Classification: I20

*Thanks go to Mehnaz Aziz and her team, Irfan Ahmad, Rafiq Tahir at the FDE, and to the teachers, principals, and supervisors at the schools involved for their cooperation during the surveys. This research benefitted from the insightful direction and advice of Ali Cheema, Jishnu Das, Tahir Andrabi, and Asim Khwaja. The authors would also like to acknowledge Yasir Khan for his able research assistance.

†§‡ all authors: Lahore University of Management Sciences

1 Introduction

Classroom innovations are rare in public sector schools in Pakistan. An extensive system of formalized examinations and cultural norms that value hierarchy ensure that classroom teaching remains didactic. Classes are teacher-led, have limited teacher-child interaction, and are focused on maximizing skill transfer–often through rote learning, with little exploration outside the government determined curriculum. Limited resources mean that few initiatives are taken to visualize concepts or experiment with tools, providing little space for children to explore ideas outside the realm of their textbook (if they have them). Teachers have limited incentives to experiment in classrooms and often do not have the necessary skills. There are also broader challenges facing the school administers which make issues of instructional approach almost secondary. This includes lack of teachers, lack of physical infrastructure and in some cases, lack of children.

One recent exception has been the introduction of a child-centered instructional method in public sector primary schools in Islamabad. This intervention has taken place due to a Public Private Partnership (PPP) between the Federal Directorate of Education[1] (FDE) and Children's Resource International Pakistan (CRI), a non-profit training and education organization that is engaged in improving the quality of education in public sector schools in Pakistan. The focus of CRI is to democratize the learning environment, encourage children to take initiative and make active choices. The CRI intervention is essentially a package of mixed educational inputs which includes teacher training, family literacy, the development of school facilities as well as supplying certain school equipment. The package is introduced into government run primary schools under existing management structures.

Beginning 2002 CRI began working in 25 public schools in Islamabad as a pilot program. The cohort of kindergarten children entering school in 2002 have recently completed their primary schooling under the CRI approach. From 2006 the CRI program has been extended to 300 plus schools in Islamabad, essentially covering all public schools involved in teaching grades kindergarten to 8 in the country's capital. Islamabad is a model schooling district for Pakistan, and its experience will provide fruitful implications for the expansion in the rest of the country.

The intervention supports a number of the Government of Pakistan's (GOP) objectives outlined in the Education Sector Reforms 2001-2005 (ESR 2001-2005) including universal primary education, improving quality of education through teacher

[1] The Federal Directorate of Education is a department of the Ministry of Education responsible for implementing the education policy in Islamabad.

training and national literacy. The CRI intervention can be classified as a government contract for management and professional services[2]. In the CRI case the government has contracted out for a package of inputs that introduces a new instructional approach based on making classrooms more child-centered and teacher-facilitated rather than teacher-led. The intervention provides an opportunity to infuse new financial resources into the existing school system. It does not change the institutional environment nor set out new incentives.

Rigorous evaluations of education sector reforms, although increasingly the norm internationally, are quite few in Pakistan. Also the empirical literature on education sector PPPs and the different instructional approaches is quite sparse. This paper, by presenting an evaluation of the CRI intervention, contributes to all three areas. The starting point for most evaluations is to address the question of the effectiveness of the said intervention whether the output being measured is learning achievement, drop-out rates, retention, access, attendance or cost per unit. Conventionally this has meant starting with the education production function and thereby assessing and isolating the effect of different inputs on the chosen measures of outputs. Most impact evaluations, particularly of experimental design, have focused on discrete inputs and their impact, whether it is the introduction of vouchers, specific instructional material or aspects of teacher training. Other studies that have attempted to gauge school effectiveness have measured effectiveness against a range of explanatory variables, in a multiple regression framework with the usual endogeneity (concerns), impairing their usefulness.

A number of criticisms have been made of these empirical studies which are based on the education production function. Meta-analysis which examine the empirical literature on inputs and school effectiveness find that there is little consensus on which inputs actually have significant impacts on school effectiveness (Hanushek 1986, Fuller 1987, Harbison and Hanushek, 1992, Fuller and Clarke 1994, Glewwe 2002). Such contradictions in results have limited value for policymakers trying to optimize school performance[3] (Glewwe 2002)[4]. Second, positive analysis of actual policy and budgetary decision-making processes reveals that inputs into the education production function are not optimized ex-ante as predicted by theory, where marginal product per dollar are equalized across inputs. Rather input choice seems inordinately influenced by educators and their preferences (Filmer and Pritchett

[2]Patrinos (2006:2) outlines different types of contracting that exist under different PPPs in education. Governments contract out for either (i) management and professional services(input) (ii) operational services (input) (iii) education services (output) (iv) facility availability (input) (v) facility availability and education services (input and output bundle).

[3]Refer to Appendix A, for a formal treatment of this issue.

[4]For a contrary view see Krueger (2002).

1997). Last, the singular focus on inputs fails to capture the more complex process of education delivery where different local conditions determine how pedagogy affects achievement (Fuller and Clarke 1994). However, attempts to measure local context and its impact on achievement through pedagogy is extremely difficult. Few rigorous studies exist which capture the impact of the local context on educational processes.

The CRI intervention combines a mix of inputs for the purpose of delivering a new instructional approach which inevitably is affected by existing norms and practices within Pakistani schools. The focus of this evaluation primarily, is to determine the impact of the child-centered instructional approach on learning achievement in the initial pilot schools and their children. Given the limitations of the education production function approach and its inability to capture the local context, no attempt is made to measure the extent to which specific inputs of the CRI intervention improve learning achievement. The focus here is to capture the 'net effect' of the CRI intervention. The lack of a baseline was the driving force behind the research strategy used in this evaluation. By utilizing Propensity Score Matching (PSM), a comparison group of schools (and children), were selected in order to compare learning achievements of Grade IV children in both CRI and Non-CRI schools. The results are based on numeric as well as reading tests in English and Urdu. We find that the CRI intervention was effective in raising learning achievement and the average gain represents an improvement of 0.25 standard deviations.

2 Background

2.1 Related Issues

The challenges facing the education sector in Pakistan are great. Education achievement in Pakistan is extremely low. Adult literacy rates stand at 48.7 per cent which is the lowest in the region (South West Asia, with the exception of Bangladesh) and well below the region's average of 58.6per cent. Additionally, despite significant efforts to ensure that Pakistan meets its Education for All (EFA) targets by 2015, net enrollment rates at the primary level in 2002 stood at 59.1 per cent which is again well below the regional average of 82.5 per cent. Underlying these figures is the disparity in educational opportunities for men and women. The Gender Parity Index (GPI) is 0.57 which indicates the low participation of women in education. Similarly, rural and urban disparities are large with adult literacy in urban areas at 72 per cent compared to 44 per cent in the rural areas (UNESCO 2006).

Schools remain important, for improvements in learning achievement,. Schooling, whether public or private, continues to be the dominant factor in explaining learning achievements in developing countries. This is unlike developed countries where socio-economic characteristics outweigh school effects other in determining learning achievement (Fuller and Heyneman 1989; Heyneman and Loxley 1983). Recent evidence from rural Punjab presented by the Learning and Education Achievement in Punjab Schools (LEAPS)[5] project also indicates the importance of school effects on learning outcomes. The project indicates that, for these districts in Punjab, 50 per cent of the variation in learning achievement can be explained by the variation in schools. Adding district and village variables increases this to 60 per cent for certain subjects. While further addition of covariates for child and household characteristics only marginally improves the total explained variation to 68 per cent (Das, Pandey and Zajonc 2006: 20-21).

Public schools also remain important despite the recent growth in private schools. Pakistan has seen a very rapid rise of private schools since the late 1980s and 35 per cent of enrolment is in the private sector. Growth in private schools has been high in rural areas and amongst the poorest sections of the population (Andrabi, Das and Khawaja 2002(b)). Recent evidence on learning achievement in Pakistan also indicates that learning gaps between public and private schools are significant with government schools lagging behind the private sector (Das, Pandey and Zajonc 2006: 19-22; Alderman, Orazem and Paterno 1996). Private schools are clearly significant players in educational sector. Despite this, there are also limits to the roles they can play in meeting the educational challenges. First they are more successful in all areas of the economy. Andrabi, Das and Khawaja (2002(b): 18-20) show that private schools have been most successful in Punjab where they have spread more widely but in Balochistan and Sindh they have been less successful. The evidence suggests that despite the success of private schools market failures may continue to exist. Private and social returns may not equate and public schools may have to deliver where private schools fail to go. And although the evidence of learning achievements in rural Punjab indicates that private schools outperform public schools, there is great variation in public school performance with top government schools performing close to top private schools in mathematics. This indicates that there is a possibility of improvement in public schools (Das, Pandey and Zajonc 2006: 19-22). One means of doing this may be through PPPs which can offer new resources, pedagogical approaches, and if accompanied with new institutional

[5] The LEAPS project is funded by World Bank and is a multiyear study that examines impact of varies policies intervention in public and private schools in three districts in Punjab (Attock, Faisalabad and Rahim Yar Khan) though testing Urdu, English and Math Skills. The LEAPS testing instrument was also used for the CRI evaluation. See section 4

arrangements, different incentives for better performance.

Public Private Partnerships (PPPs) are a popular instrument through which governments across the world are seeking to address the ills of education delivery in the developing world By introducing non-government actors, through different contractual arrangements, to supply some (or all) educational inputs or outputs, the objective has been to increase quality, cost effectiveness and access to schools. Public Private Partnerships have also offered a means of infusing financial resources into otherwise fiscally constrained public sector schools. At the same time, PPPs have also offered the opportunity various new pedagogical approaches into public schools in the developing world. Despite the positives ascribed to PPPs, few rigorous evaluations about their impact on quality and learning achievement or otherwise on cost-effectiveness[6] .

Pakistan has been no exception. The Government of Pakistan (GOP) has been a forerunner in the experimenting with PPPs in the educational sector. A new initiative was launched in 2001 under the ESR 2001 giving special emphasis to PPPs. The GOP recognized the government's limitation in managing the education sector and advocated a much larger role for PPPs for the purpose of both (i) mobilizing financial resources and (ii) designing, executing and monitoring education activities (GOP 2001). The GOP's goal as stated in the ESR was to sign on 26000 partnerships by 2005. As of 2003, the government has already had 6,240 schools upgraded through PPPs in Punjab and NWFP (GOP 2004, 6). Yet given this breadth of experimentation with PPPs, few evaluations exist of these interventions. The exception is the evaluation of Urban Girls Fellowship Program in Quetta (Kim, Alderman and Orazem, 1998). Through the program, the GOP randomly allocated a government subsidy to private schools to encourage increased female enrolment. The evaluation found that the program increased female enrolment by 33 per cent and male enrolment by 27.5 per cent.

Little is known about the adoption and quality of non-didactic learning approaches in Pakistan. Adoption of child-centered and teacher facilitated instructional approaches are likely to be limited to the private sector schools, with the exception of CRI intervention in public schools. Even internationally, the empirical literature is limited on outcomes under non-didactic approaches and mostly

[6] Most evaluations of PPPs are of individual cases given the project based characteristics of most PPPs. The exception is Woessman (2006) which offers an international cross-country comparison of PPPs against PISA Test. On project based evaluations most of the recent literature is from Latin America. See Allcott and Ortega (2007) on the Fe y Alegra private school system in Venezuela, Angrist et al. (2002) and Angrist et al (2006) on the PACES Vouchers Program in Columbia, and Barrera (2007) on the Concession School Program in Bogot Columbia. Others

draws on the US experience. The overall evidence from a number of longitudinal studies in the US favors child-centered instructional approaches for children of lower income groups in specific subject areas (Schweinhart and Weikart 1988, Miller and Bizzell, 1983; and Stipek and Stalling 1986). Other studies have found that socio-economic status matters in determining whether child-centered programs affect learning achievement. Studies that examine the introduction of child-centered approach in middle-class children see no significant difference in learning achievement between these two approaches (Hirsh-Pasek et al 1990; Hirsh-Pasek 1991). A more recent study which compares children in child-centered pre-school programs against those in non child-centered programs, found that children in latter fared better in mathematics tests but did not do so well in tests that tested for letters and reading. However, on others measures of motivational achievement children in child-centered programs fared much better. In this study there were no differences in the treatment effect between different socio-economic groups (Stipek et al, 1995).

The literature on learning achievements in non-didactic approaches in developing countries is even more sparse and actually contradictory to the literature from the US. This may partly be explained by the fact that most assessments of child-centered and teacher facilitated approaches assess children of a higher age group. Lockheed and Zhao (1993) for instance highlight that approaches that are more child-centered correlate negatively with science and math achievement in the Philippines (Fuller and Clarke, 1994). Other studies that examine issues of a more participatory nature, find that children do better on achievement tests in classrooms where more didactic approaches are followed.(Lockheed and Komenan 1989).

2.2 Intervention

The CRI intervention was adopted in 2002 in Islamabad Capital Territory (ICT) schools as part of a larger strategy of the Government of Pakistan (GOP) to establish ICT as the model school region in Pakistan. With USAID funding, and the aid of its parent organization, Children Resource International Washington, CRI was established in 2002 to implement the CRI program in a total of 118 pilot primary schools throughout Pakistan of which 35 schools were in Islamabad. Twenty five schools began with the program in 2002 and an additional ten schools joined in 2003. The CRI program was introduced in yearly steps, starting with Kindergarten/Grade I teachers in 2002, later with teachers of each subsequent grade being introduced to it every year. By September 2006, teachers in all grades between I - V in the 35 schools had been trained in the child-centered instructional approaches. Also by August 2007, the cohort of Kindergarten/Grade I children were the first to have completed their whole primary education (KG/Grade I-V) under the child-centered instructional approaches.

At an operational level the program works at improving the teaching techniques and improving the classroom environment. These aspects of the program have fit well with the national level initiatives taken by the GOP and directly address three of the seven key focus areas of the ESR: (i) National Literacy (ii) Universal Primary Education and (iii) Improving the Quality of Education through Teacher Training.

In concrete terms the CRI program has a number of components which essentially focuses on training teachers on interactive learning techniques and supporting them with technical support and learning materials. Twice a year, the teachers in program schools undergo rigorous five day training sessions which impart techniques, curriculum guidance and hands-on experience with using learning material. During the training sessions, each teacher's progress is individually monitored by master teacher trainers (MTTs) who provide guidance and offer help. Perhaps the most important component of the CRI program is the continual technical assistance offered to teachers and their schools after they have left the training workshops. Subsequent to the training, trained teachers are regularly monitored by MTTs who make periodic visits to the schools. This allows taught techniques to be reinforced besides allowing teachers to engage with MTTs throughout the year, providing them with an opportunity to share experiences and address concerns as the academic year progresses. In addition the CRI program has introduced different activity centers within the classroom to facilitate interactive learning. Centers usual set up include centers for Mathematics, Science, Language and Arts and Drama. CRI provides the teaching and learning materials for their activity centers. Overall between 2002-2006, 788 teachers were trained including 173 head and deputy head teachers (see Table 1).

The CRI program or intervention is essentially an interactive teaching and learning program that focuses on moving from a teacher-directed learning process to a teacher facilitated learning process. In addition to the focus on teaching training, there is a strong emphasize on parent involvement and parent literacy in the program. The philosophy behind the program is that children develop best when they are actively involved in their own learning process. Hence the goal is to create a class room environment, where children feel confident to drive the learning process, which is open-ended and is overall a fun place to learn. Pedagogically, the objective also is to develop non-cognitive elements of the child's personality on the belief that non-cognitive development stimulates future success. This latter aspect of the program awaits further research for complete assessment.

3 Methodology

The basic idea, in assessing program impact, is to evaluate the causal effect of treatment on some outcome of interest. Since selection of schools for the CRI program was not random, it would be incorrect to infer causality through standard regression analysis. This is because program impact could be overestimated due to selection bias. Moreover, there was no baseline survey conducted at the time of the initiation of the CRI intervention. We therefore did not have any information on pre-treatment characteristics and outcomes.

In view of this, we make use of Propensity Score Matching (PSM) techniques in order to estimate the true impact of the CRI intervention on various outcomes. PSM employs a predicted probability of group membership -e.g., treatment vs. comparison group– based on observed predictors, usually obtained from logistic regression to create a counterfactual group[7]. In this case, we are interested in understanding how an exposure to the CRI program impacts on student performance. In a population of individuals, there will be some who have received the treatment and some who have not. Following Sianesi (2001), let us denote: Y_i as the actual observed outcome of any individual i; Y_{1i} as the outcome of unit i if i were exposed to treatment; Y_{0i} as the outcome if i were not exposed to the treatment; $D_i \in [0, 1]$, indicates the treatment actually received by unit i i.e. the causal effect. Thus we have,

$$E(Y_{1i} - Y_{0i}|D = 1) = E(Y_{1i}|D = 1) - E(Y_{0i}|D = 1) \tag{1}$$

In the absence of information on how the outcome of an individual would have varied if he had not received treatment (a hypothetical case), the empirical challenge is to extract, for comparison, the outcome of interest from the population of those untreated individuals who are most similar in characteristics to the treated i.e $E(Y_{0i}|D = 0)$. This will imply matching on certain observed characteristics two individuals, one who has received the treatment and one who has not. The difference in their outcomes can then be attributed to the effect of the treatment. However, if the dimension of the vector of covariates, X, is high as is usually the case, it will be very difficult to undertake categorical matching and find exact matches for every treated unit. It is therefore necessary to base our matching criteria on some function of these observed covariates, $b(X)$. This has been termed by Rosenbaum and Rubin (1983), as the balancing scores. This means that the conditional distribution of X given $b(X)$ is independent of assignment into treatment. Propensity Scores is one such function of X, a type of balancing score. For brevity, it is the the predicted probability of participation into a program conditional on the observed characteristics X, or $p(X)$.

[7] See Rosenbaum and Rubin (1983) for their pioneering work on the introduction of this method.

In our case, we are interested in extracting from the population of untreated Islamabad schools, those schools which would have been as likely to receive the CRI program, but in fact did not. It has to be noted that while the selection for the CRI program was restricted to schools, our objective is to compare individual test score differences. Thus our identification strategy hinges on a two-stage matching procedure. We first construct a comparison group of schools who have not received the CRI program (Non-CRI schools, henceforth) but are most similar in characteristics to the set of schools who have received the CRI program (CRI schools, henceforth). We then pool the student level information from both the treatment and control schools to subsequently match the children across these schools in order to compare their performance. The comparison here is between CRI children and Non-CRI children.

Our parameter of interest is the 'average treatment effect on the treated' which is (Caliendo, Kopeinig, 2005):

$$\tau_{ATT} = E(Y_{1i}|D = 1) - E(Y_{0i}|D = 0) \tag{2}$$

In order for the true parameter to be estimated we require that:

$$E(Y_{0i}|D = 1) - E(Y_{0i}|D = 0) = 0 \tag{3}$$

This will ensure that the τ_{ATT} is free from self selection bias. In non-experimental studies, we make use of the following assumptions in order to fulfill the above criteria.

Conditional Independence Assumption All the relevant differences between the two groups are captured by their observables X.

$$Y_{0i} \perp D|X \tag{4}$$

Common Support Further, to ensure that every unit in the population has a positive probability of being both participants and non-participants (therefore ruling out perfect predictability) we make use of the common support or the overlap condition,

$$0 < Pr(D = 1|X) < 1 \tag{5}$$

Both these conditions together form the '*strong ignorability of treatment*' assumption which allows us to apply the PSM estimator as follows:

$$\tau_{ATT}^{PSM} = E_{P(X)|D=1}[E(Y_{1i}|D = 1, P(X)) - E(Y_{0i}|D = 0, P(X))] \tag{6}$$

3.1 Identification Strategy

The paper employs a two-stage matching procedure to estimate the ATT estimator. In the first stage we undertake school level matching and in the second stage, we match children in these sample schools[8].

3.1.1 School Level Matching

Twenty five schools were selected for the CRI program in 2002 by the Federal Directorate of Education (FDE). These comprise our treatment group for estimation. The comparison group of schools was thus constructed from the remaining population of Islamabad schools. In order to choose a suitable comparison group of Non-CRI schools, we estimated a logistic regression where the dependent variable (y_s) was whether the school was selected for the CRI program by the FDE. This enabled us to predict each school's probability of being selected by the FDE for the CRI program. The FDE had provided us with a fair idea of the school attributes on which they had based their selection decision. An important criteria was the location of schools. Islamabad is divided into various sectors and sub-sectors. Thus our matching strategy relied in large part on finding a good match for CRI schools within the same sector. In order to define the geographical boundaries evenly, we defined a cluster variable based on the divisions of sectors in Islamabad. Further, we included only those variables that were unaffected by participation or its anticipation. We took precautions as not to include too many extraneous variables as this would have increased the variance of our estimates. Also including some variables which predicted perfect participation was avoided as it tended to exacerbate our common support condition. After incorporating these considerations, we arrived at a fairly parsimonious specification for our school selection. Thus,

$$Pr(y_s = 1) = \Lambda(\alpha_1 + \beta_1 K_s + \psi) \quad (7)$$

where K_s is the vector of school quality covariates and ψ is the cluster specific effects. Λ is the cumulative distribution function for a standard logistic random variable.

We then assigned to each school its predicted probability (p-score). Since the ATT is best defined over the region of common support, we examine the minima and maxima of estimated p-scores in each block to determine the region of overlap between the treatment and comparison groups. This ensures that any combination of characteristics observed in the treatment group will also be observed in the control

[8]Intuitively this implies that the probability of a child being picked for CRI intervention would be given by the joint probability of the school being selected for the CRI program and the child going to that school i.e. $P(D = 1) = P(D_s = 1) * P(D_c = 1)$.

group, discarding away those observations which fail to meet this criteria. Thus, observations whose propensity score is smaller then the minimum and larger than the maximum in the opposite group, fall outside of the region of common support and are deleted. Based on this criterion, we were able to trim our sample of Islamabad schools from 100 to 67. The remaining schools were then stratified into various blocks -which we will term as school level blocks- based upon their p-scores. Finally, we check whether our matching procedure is able to balance the distribution of the relevant variables in both treatment and comparison groups. Since our matching strategy is primarily based on stratification we test for the following two balancing properties to ensure that our matching exercise is valid.

I. The average estimated p-score of the treated group is not statistically different from the average estimated p-score of the comparison group in each block.

II. Within each block the means of each characteristic are not statistically different between the treated and comparison groups

Ensuring that the balancing properties have been satisfied will imply that the covariates are balanced and that the assignment to treatment can be considered random within each block.

3.1.2 Child Level Matching

We use a similar procedure to match CRI children to non CRI children. However in this case we estimate the p-scores employing a logistic regression of participation *within each school level block*, which was defined previously by school level matching. Thus we have,

$$Pr(y_i = 1) = \Lambda(\alpha_2 + \beta_2 C_i) \tag{8}$$

where y_i is whether the child has been recipient to the CRI program (being in a CRI school at the time of testing) or not; C_i is a vector of all child covariates. Each child is then assigned a p-score based on his probability of recieving the CRI treatment i.e. his/her probability of attending a CRI school. Based upon this, children are further stratified into blocks. As stated earlier, we restrict our observations within the common support region as well as test for the two balancing properties. It has to be noted that each school level block is further stratified into various child level blocks. The final blocks, the child level blocks, that have been created are balanced not only on child level covariates but also on school level covariates. That being, the case assignment of both schools and children within each block to treatment group can be considered random.

3.1.3 Estimation

Within each block, in which both CRI and non-CRI children are present, we will compute the difference between the average outcome of the treated units and the control units. Following Becker and Ichino (2002), the ATT estimator is given by,

$$\tau_q = \frac{\sum_{i \in I(q)} Y_i^T}{N_q^T} - \frac{\sum_{j \in I(q)} Y_i^C}{N_q^C} \tag{9}$$

where $I(q)$ is the set of units in block q while N_q^T and N_q^C are number of treated and control units in block q. This is then averaged across blocks to give us the ATT,

$$\tau_{ATT} = \sum_{q=1}^{Q} \tau_q \frac{\sum_{i \in I(q)} D_i}{\sum_{\forall i} D_i} \tag{10}$$

This is a weighted average of the estimators computed within blocks, where the weight for each block is given by the corresponding fraction of treated units and Q is the number of blocks.

3.2 Self-selection and Sensitivity Analysis

children in Grade IV who undertook the test also included those who were not in their present school since KG or Grade I. They are lateral entrants into either CRI or Non CRI schools. Their inclusion into our analysis presents a particularly confounding effect whereby it could be argued that they have self selected into the sample. In this case, in fact, we do not have a strong prior on the direction of the resulting bias. It could be that interested parents choose CRI schools as they are aware of its value for their child's education. But it may also be the case that parents are apprehensive, like users of most new innovations, of the new CRI methodology and tend to prefer the old means of instruction that have been tried and tested. Hence, the bias could go either way. Be that as it may, there is an additional advantage to focusing on the restricted sample which arises from the fact that it only comprises of children with (at least) four years of CRI exposure. If we believe that learning is persistent and cumulative, then including children with different levels of exposure to the program would dampen the overall estimate of the program's effect due to the heterogeneity of individual-level treatment effects. Thus, net of any selection bias, we would expect higher estimates of the CRI's learning effect from the restricted sample.

We also perform a range of analysis to test for the robustness of our estimates of the average treatment effect. In the first instance, we consider alternative estimators for the treatment effects by comparing the results from stratified matching to those obtained by using various other propensity score based matching techniques viz. nearest-neighbor matching, radius matching and kernel-based matching. Secondly, we consider the robustness of our estimates to unobserved bias as explained below.

Our matching technique balances the assignment of treatment and control units conditional only on observed covariates. This is founded on the conditional independence assumption stated earlier. However, if there are any unobserved variables that additionally influence the assignment into treatment, then our results will be biased. This influence of unobservables causes a hidden bias to emerge to which our matching estimator is not robust (Rosenbaum, 2002). Although it will not be possible to correct for this hidden bias (if such a selection bias exists) due to the use of non-experimental data, we will however try and draw insights as per the specification and validity of our matching analysis. To do so, we will make use of the bounding approach as proposed by Rosenbaum (2002). This approach provides us with a 'worst-case' scenario by allowing us to determine how strongly an unobserved variable can affect selection thereby undermining our conclusions about the causal effects of the program. Briefly, the approach involves estimating the following (Rosenbaum 2002): If there are certain unobservables that simultaneously affect the selection process then the probability of treatment is given by,

$$P = Pr(y_i = 1|X, u) = \beta X + \gamma u \tag{11}$$

where X is a vector of all our observed covariates (school and child level) whilst u represents the unobserved covariates affecting school selection into the program. The odds that the individual receives a treatment is given by, $P/(1-P)$. This means that for a matched pair of individuals i and j, the odds ratio is,

$$\frac{P_i/(1-P_i)}{P_j/(1-P_j)} = \frac{exp(\beta X_i + \gamma u_i)}{exp(\beta X_j + \gamma u_j)} \tag{12}$$

Further since $x_i = x_j$, the odds ratio reduces to $exp[\gamma(u_i - u_j)]$. The bounds on the odds ratio that either of the two individuals receive the treatment is thus,

$$\frac{1}{e^{\gamma}} \leq \frac{P_i/(1-P_i)}{P_j/(1-P_j)} \leq e^{\gamma} \tag{13}$$

For the case where $e^{\gamma} = 1$, there will be no hidden bias as either $\gamma = 1$ or $u_i = u_j$. The bounds on the odds ratio start to vary as e^{γ} takes on values greater than one i.e. as we allow for the presence of unobserved variables to influence the assignment into treatment. The Rosenbaum Bounds calculates at each value of e^{γ}, the significance level for the null hypothesis that treatment effect is equal to zero.

By examining this, one can identify and examine the point at which the hidden bias causes us to question our findings. This cut-off point should be large enough to render our estimates robust against the presence of unobservable selection bias.

4 Data

The paper uses two primary sources of data: the FDE database on school-level attributes and information collected through tests and surveys administered to individual Grade IV children and their teachers in selected sample schools. The FDE database contained detailed information on all Islamabad schools from a census conducted in 2003. This data was used to select matching non-CRI schools on the school attributes deemed important in the selection of CRI schools.[9]

We examine the impact of the CRI program on five outcomes of interest: the total score obtained by a child on the test as well as the score in English, Urdu and Mathematics along with his rate of class attendance. The child testing instrument, which had three sub-sections on Urdu, English and Mathematics comprised of 48, 51 and 44 questions respectively. The language sections tested children, among other things, on their reading comprehension and simple sentence creation abilities and the Mathematics test included arithmetic operations like multiplication, division, LCM etc. The questions ranged in their difficulty level and were appropriate for the curriculum being taught in the schools for grades 3-5. To maintain reliability of the test outcomes, the testing instrument was kept confidential from the school authorities in both CRI and non-CRI schools.[10] In addition, we obtained school attendance records of children for the months of January, February and March 2007. The survey also provided us with detailed information on children and their parents' occupation, education as well as household asset and demographics using a child questionnaire which was administered on the same day to all children who took the test. Relevant information was also collected on class teachers in the sample schools.

A simple aggregate of correct responses to a test instrument is often used as a measure of a student's learning. Our first results table will indeed use the aggregate score on the test. However, such an approach has a serious shortcoming in that it

[9]This was discussed extensively with the Director Training of FDE, Professor Rafique Tahir, and our specification was based on the criteria highlighted in that meeting. See table A2 in the appendix.

[10]The testing instrument used in this study was borrowed from another large-scale school research project, LEAPS. For more details on the testing instrument, refer to Andrabi, Das, and Khwaja (2002a).

does not account for the differences between individual questions on the test. The probability of giving a correct response to some question on a test depends not just on the student's ability but several features of the question itself such as its level of difficulty and discriminating power between high and low ability individuals. Thus two individuals with the same aggregate score on the achievement test might still differ in their latent cognitive ability as pointed out by an example in Das, Pandey and Zajonc (2006).

Therefore, following the standard IRT literature, we specify a three parameter logistic model for the probability of a correct response by individual i to test question (or item) j, P_{ij}:

$$P_{ij}(\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{exp(a_i(\theta_j - b_i))}{1 + exp(a_i(\theta_j - b_i))} \tag{14}$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, and c_i is the lower asymptote for the probability function denoting random guessing on an item. Given data on individual responses to items on a test, it is possible to estimate the true latent ability θ for each examinee using the method of maximum likelihood. For our sample, we estimate the latent cognitive ability for each student (along with the 143×3 item parameters) and re-scale θ to obtain the expected score on each section of the test as our outcome variable. The latter score is essentially a sum of the probabilities of providing a correct answer to each question on the test.

The controls used in the participation model consist thus of school attributes as well as teacher quality, child and household level variables. At the school level we use geographical sector specific dummies (clusters) to capture location specific heterogeneity of schools. School attributes such as type of school, medium of instruction and gender stratification of schools are included in the usual fashion. At the child level matching, gender of the child, mother's education and father's occupation are entered as categorical variables in the model. Further, we include age of the child, number of siblings, number of elder siblings and child anthropometries (child weight and child height). Finally we account for the wealth of the household by including an asset index. The child questionnaire, administered to each student test taker, recorded the number of each asset (about 20 different types) present in the household of the child. We were able to combine this information into an asset index by the method of principal components (refer to Table 11. Briefly, this procedure enabled us to extract from the large matrix of assets, an eigenvector which best captured the common information. This gave us the weights for each of the assets with the help of which we defined the composite asset index. For household

i, the asset index A_i is given by (Filmer and Pritchett, 1998):

$$A_i = f_1 * \frac{(a_{i1} - a_1)}{s_1} + \ldots\ldots + f_N * \frac{(a_{iN} - a_N)}{s_N} \tag{15}$$

where f_1 is the individual weight of the first asset, a_{i1} is the household's value of the asset, and a_1, s_1 are the mean and standard deviation of the first asset.

Table 2 reports descriptive statistics on all those Grade IV children who took the test and compares the means of different variables across treatment and comparison groups. Several variables have a significant difference across the two school groups. But that is to be expected since the table is comparing the average children in CRI and non-CRI schools after pooling together all the data from different matched blocks. A better comparison is to look at CRI and matching non-CRI children within a block defined on *child* attributes. Figures 4 and 5 do such a comparison by depicting the mean age and assets of children within blocks. Additionally, Figure 6 presents this comparison for the distribution of age. Overall, the children in the treatment and control group are well-matched on age and wealth within these blocks.

5 Results

The full sample of matched children contains data on 1984 children from 63 schools. After having matched children in the CRI schools with similar children in the Non-CRI schools, we proceed to obtain estimates of the average CRI effect on a student's exam scores. Before obtaining non-parametric estimates of the treatment effect, we plot the distribution of exam scores for the children in the two groups. Figure 3 shows that the overall distribution of scores for CRI lies to the right of the scores obtained by non-CRI children. Thus, without controlling for individual heterogeneity, there is a marked difference in scores in favor of the CRI children.

Table 3 reports the average treatment effect on the treated obtained after stratifying the sample by blocks of matched children (treated and untreated). First we compute within-block mean treatment effects which are then combined in a weighted average to compute the overall effect. The table reports the treatment effect for each of the three subject components of the test as well as the overall test score and average rate of attendance during the period Jan-Mar 2007. It also reports the bootstrapped t-statistics for the estimated coefficients as well as the 95% confidence intervals.

As reported in Table 3, the estimated treatment effect is positive for all of the outcomes, i.e. CRI children perform better than similar children who have not received exposure to the CRI program and tend to attend school more often. The difference is larger in magnitude for the two languages as opposed to mathematics. Hence, the CRI program seems to have the greatest effect on a child's reading and writing skills as opposed to simple numeric abilities. The effect on math scores is, in fact, statistically insignificant even at higher significance levels whereas the ATT estimate is statistically significant at 5% for urdu and 1% for english.

The effect of the CRI program on a student's overall performance on the test is to increase the average score by 4.893 points. That is, a typical CRI-exposed student will correctly answer about 5 more questions as compared to an identical student in a non-CRI school. Since the exam contained 143 questions in total, this constitutes an improvement in test scores of 3.4 percentage points.[11] The table also reports the treatment effect on attendance. The attendance rate over the last three months has been greater by 4.7 percentage points for CRI children as compared to non-CRI children, which is statistically significant at 5%.

However, the above table reports the effect on a simple, un-weighted aggregate score obtained by each student which may not be a good measure of a test-taker's (unobserved) ability in the presence of random guessing by examinees and the inclusion of questions with varying difficulty levels in the test. Thus, as mentioned in the above section, we fit a 3-parameter logistic item response model to the test data in order to compute a better measure of child ability. Specifically, we use our estimates from the item response model to construct a *true score* for each child defined as the expected test score on each section of the test. The following analysis uses these IRT-adjusted test scores to compute treatment effects.

In addition, Table 3 also includes some CRI schools where teachers without CRI training were found teaching class IV as well as a few non-CRI schools having CRI-trained teachers interacting with the children in our sample. The assignment of teachers to government schools in Islamabad is done by the Federal Directorate of Education (FDE). Transfers across schools also happen at the discretion of FDE, which assured CRI that there will be no shifting of their trained teaching staff to other non-CRI schools in the pilot phase of the project, 2002-06. This re-assignment is thus an outcome of the CRI's expansion beyond the pilot phase to all Islamabad schools starting in August 2006. Given that we find these teachers towards the end of school year after CRI has already organized its initial and follow-up training for that grade, it is likely that this contamination of our treatment and comparison

[11] Notice that this is a simple comparison of the total number of correct responses. It does not take into account the relative difficulty of the different questions included in the exam, an analysis which we do below according to the item response theory.

groups will introduce a downward bias in the estimates. We drop such observations (less than 18% of the full sample) from subsequent analysis.

Table 4 reports the effect of CRI on IRT-adjusted test scores after dropping all those students with re-assigned teachers.[12] The exam results all show a positive and highly significant effect of the CRI program, with the effect being largest in magnitude for english and urdu. Thus the true expected score of students, which is a measure of their latent cognitive skill, continues to show an improvement in CRI schools relative to non-CRI schools. Moreover, the attendance rate continues to be significantly higher in the CRI schools with an estimated treatment effect that is similar to the one reported in Table 3.

Next we examine the robustness of these results by using alternative estimators. We employ three alternative non-parametric estimators of the treatment effect based on the propensity score matching method. All of these estimators use the estimated propensity score of each child in the CRI school to define the matching non-CRI children. The first estimator **attnd** matches CRI-exposed individuals with the nearest control in terms of the propensity score, computes the individual-level treatment effect for all such matched pairs and averages over all treated individuals to compute the average treatment effect. The second estimator **attr** is similar in design although it might match each treated subject to several control subjects having propensity scores within a radius of 0.1 units. The last estimator **attk** matches the treated with a weighted average of *all* un-treated children with weights that are inversely proportional to the distance between their propensity scores. Thus, the last two estimators might create a many-to-one match for each treated individual on the sample thereby increasing the number of matched observations but potentially reducing the quality of the match.

Table 5 reports the average treatment effects obtained using these alternate estimation methods. For all the outcome variables, the sign of the treatment effect stays positive although the magnitude of the point estimate is generally smaller than the stratified estimate reported earlier. Moreover, eleven out of the twelve estimates for exam scores are statistically significant. Of the three subjects, the effect of CRI exposure on English skills is the strongest across all columns. Regardless of the estimator used, the overall effect on combined scores is statistically significant at 5% and lies in the range of 2.35-2.65 increase in the expected score. As found earlier, the difference in attendance rates across the two groups is positive and, except one

[12]The ATT estimates for simple scores were higher in magnitude, as expected, and significant (not reported).

instance, statistically significant. Combining this with the earlier estimate, it seems that the CRI program increased child attendance by about 3.0-4.2%.

An effect of this magnitude on student test scores is quite big. One has to keep in mind that while the actual size of the learning effect varies from individual to individual, the above estimates capture the *average* CRI effect on student's learning. A difference in expected IRT-adjusted score of 3.0 (say) implies an increase in the student's rank by about 5 percentiles at the mean of student score distribution.[13]

However the large estimates above may reflect a positive bias arising from self-selection of high- achieving children into CRI schools in the years after the start of the program in 2002 (or negative bias vice versa). Even though we matched children on their observed attributes prior to obtaining these results, we do not explicitly control for variables like parental motivation and involvement in schooling, which might affect school selection conditional on various child attributes especially for all those children who were admitted to our schools a year or two after the start of the CRI program. In addition, as mentioned in the methodology section, the problem of heterogenous treatment effects is also likely to be more problematic in the presence of these lateral entrants with less years of exposure to the program. We attempt a more careful analysis of the exam scores by estimating the treatment effect on a restricted sample consisting of only those children who have stayed in the same schools since Grade KG or I and thus had at least four years of CRI exposure.[14] While this helps address the potential issues of selection bias and heterogeneity in the above results, it also limits the number of observations available for analysis.

Table 6 reports the average treatment effect obtained when this restricted sample was stratified into blocks of similar CRI and non-CRI children using child and teacher quality variables, as before.[15] Notice that the difference in all subject test scores is now significantly positive, that is, CRI children with a minimum four years of exposure perform better than similar non-CRI children in each component of the test. The point estimates are lower than the estimates obtained using the full sample in Table 4, which implies that the earlier results were biased upwards perhaps due

[13]To measure the lower bound of CRI effect, we look at the 2.35 estimate which implies a rank improvement of 4 percentiles at the mean.

[14]The number of lateral entrants into the non-CRI schools represent 32 percent of all non-CRI children in the sample, while they were 28% of all CRI children in the sample. However, lacking detailed information on children who leave the sample schools, we are unable to attribute this child mobility to self-selection.

[15]See Table 11 for the variables used and the estimates of the probit selection equation in the restricted sample.

to self-selection by more able children into CRI schools (net of the lower cumulative effect of CRI on these late entrants).

Next we do a similar sensitivity analysis of these results using alternative non-parametric estimators of the treatment effect. The practical as well as statistical significance of the results stays largely unchanged in Table 7. There is a positive effect of CRI program on all the five outcome variables. However, the results are less robust than the full sample estimates with only five out of the twelve exam results being statistically significant. Moreover, the magnitude of the effect is similar for the first two estimators, *atts* and *attnd*, and higher than the bandwidth estimators (*attr*, *attk*). Also noteworthy is the fact that, apart from the attr estimate for total score and attk estimate for attendance, all other coefficients are insignificant for these two estimators in contrast to the estimates obtained using *atts* and *attnd*.

Thus it is safe to conclude that there is a positive and significant effect on the overall learning of children with at least four years of exposure in CRI schools. The effect ranges in value from 2.26-3.83 and is significant for three out of the four estimators used. In addition, we observe an increase in the IRT-adjusted score of between 1.32-1.38 for Urdu, 0.95-0.99 for Mathematics and 1.41-1.46 for English. These subject scores are significant in only two of the four cases but are broadly consistent with the results reported earlier and denote a cumulative effect of the CRI program on those children who stayed in the same school since Grade I. Once again, a positive effect on average attendance rates is observed in the restricted sample.

At this point, it is natural to wonder whether this effect of an active learning environment on child learning interacts with factors such as child gender and economic background. That is, whether girls or boys (similarly, rich or poor children) are more receptive to such a pedagogical innovation and show additional gains in CRI schools. The motivation for the first disaggregation comes from a curiosity about how cultural gender stereotypes of chatty, social girls and 'analytical' boys plays into a changed classroom with greater opportunities for discussion and student-teacher interaction. The latter analysis on wealth is interesting from the perspective of an education production function approach where the home environment and parental resources are likely to influence the outcomes from classroom learning.

We decompose our data from the uncontaminated "restricted" sample by gender and wealth to obtain separate ATT estimates after defining new child-level blocks with balanced covariates.[16] Tables 9 and 10 report the estimated CRI effect in each

[16] The variables used for matching are the same as the ones used earlier for the full restricted sample. In some cases, a school block had to be dropped due to lack of sufficient observations in treatment and comparison groups.

sub-group. Interestingly, the point estimates for the program's impact on boys and girls are very close for the learning outcomes and, if anything, slightly higher for boys. However, there might be something to the cultural stereotypes as shown by a reduced attendance rate for boys in CRI schools as opposed to an increase in girls' attendance. So if the boys don't find CRI classrooms 'fun', what motivates them to do better on achievement tests than their non-CRI counterparts? In the concluding section of this paper, we return to a discussion of different channels through which the CRI intervention might have a positive effect on learning, a topic whose fuller treatment must await further research.

Table 10 reports the estimates for children from rich and poor households. The rich kids' sample consists of children whose household asset index belongs in the top 35% of the wealth distribution in the sample and vice versa for the poor kids. These wealth thresholds were necessitated by the need to have sufficient data for analysis while leaving out a sizable median segment. The table shows a positive impact of the CRI program on the achievement of both rich and poor children with a relatively higher impact on richer children. Given the point estimates, there is a difference of about 1.67 units in CRI's impact on the overall IRT-adjusted score for rich and poor children, which indicates considerable heterogeneity of the program's effect across students. In addition, the ATT estimates in panel B are not statistically significant owing perhaps to the relative imprecision from using a small estimation sample. The same difference in statistical significance can be seen between panels A and B of Table 9.

5.1 Sensitivity Analysis

Finally, we would like to examine the various assumptions underlying our non-experimental estimators in the spirit of Heckman and Hotz (1989). First we present evidence that the individual covariates are well balanced across the treatment and comparison groups within each block. Table 12 reports two regressions. The regression in column (a) attempts to determine if, within a block of comparable CRI and non-CRI children, any of the child attributes is significantly correlated with going to a CRI school. We find that the individual variables are jointly insignificant whereas the block variables are jointly significant. This implies that while there are between-block differences, there are no differences within the same block in the probability of going to a CRI school as a function of these observed child and teacher attributes. Column (b) attempts to explain the selection of a school for CRI intervention within each *school block*. We obtain a similar result for the joint significance of school features, including geographical location and type, within school-level blocks which indicates that schools within a block are similar in all these features except their assignment to the treatment and comparison group.

From the above regressions, we conclude that propensity score matching works well in matching CRI and Non-CRI subjects post-treatment implying that the key attributes of children and their schools are not significantly different across the treatment and control groups. Invoking the *Ignorability of Treatment* assumption standard in the matching literature, we have assumed for our results above that the attributes used in the propensity score matching are sufficient determinants of an individual student's selection into the treatment group. That is, any unobserved determinants of selection into treatment and control groups are adequately accounted for once we control for these factors.

5.1.1 Selection on Unobservables

To what extent are our results robust to a potential imbalance in the unobservable factors across matching blocks of observations? We attempt to account for the unobserved bias by computing the Rosenbaum bounds for the estimated average treatment effects. Table 8 reports the different levels of unobservable imbalance across the two groups captured by the gamma values. For each gamma value, columns 2-5 report the significance levels (or p-values) for the null hypothesis of no treatment effect in test scores based on the stratified matching estimator. Each column can thus be used in a Rosenbaum bounds analysis for the ATT estimates as it helps identify the gamma values just before the effect becomes insignificant at 10%.

Using the restricted sample, we compute the bounds for the *atts* estimator and find that the treatment effect for Urdu continues to be significant at 10% level for gamma values up to 2.40 and it stops being significant in between 2.40 and 2.45 indicating that the cutoff gamma value lies in this interval.[17] This means that the unobserved effect would have to increase a child's odds of receiving CRI treatment by a factor of 2.40 or more than 140% before it would alter our conclusion that exposure to CRI treatment improves a student's Urdu. For Mathematics, the critical gamma value lies in the range of 2.60 to 2.65, between 2.40 and 2.45 for English as well as the total test score (see Table 8).

To discuss what these values suggest about the sensitivity of results to hidden bias, observe that a gamma value of 1 corresponds to the assumption that the study is free of any hidden bias. If the *cut-off* gamma value is close to 1, the study would be considered as sensitive to hidden bias. There is not yet clear consensus as to

[17]The 'Gamma' reported in Table 8 is the same as the e^{γ} discussed in the methodology section.

what qualifies as a near enough cut-off point that would undermine the results of estimation. Watson (2005) notes that while in health sciences, the cut-off point for the odds ratio are found to be quite large (greater than 6), the same estimates are expected to be much lower in social science (as low as 1.1). In that sense, our results can be termed as quite robust to hidden bias.

Furthermore, in our case, the probability of exposure to CRI depends on school choice. And prior to obtaining estimated CRI effects we have already matched children on a range of household attributes that are known to influence the school choice decision such as household wealth, parental education, occupation as well as child age, gender, and number of siblings. In addition, as noted elsewhere, Rosenbaum bounds present a worse-case scenario in the sense that it tells us just how large the unobserved effect has to be in order to undermine our results without actually estimating the actual extent of such bias. Hence a cutoff value of 2.40 can be considered fairly large in our case as it indicates a 140% change in odds ratio of attending a CRI school due to unobservables. Therefore, we conclude that our results are reasonably robust to hidden bias arising from unobserved effects driving school selection.

6 Conclusion

The CRI program has been working for the last five years to improve the learning environment in Islamabad government schools by making classrooms child-centered. The Program has adopted a multi-pronged strategy where it trains the teachers in interactive teaching styles besides provide additional teaching aids and instructional material for the treated classrooms. We have found that there is a significant positive learning impact of the CRI program on all the curricular components of elementary education which may be explained by its effect on the classroom environment and increased student attendance in CRI schools. Grade IV children in CRI schools performed better, on average, than comparable children in non-CRI schools. This higher performance was seen in all three subject areas: English, Urdu and Mathematics, when comparing children who have had a greater exposure to the program. Based on our estimates and the empirical distribution of test scores in the sample, exposure to the CRI program improved the average student's ranking by 4-11 percentiles above his current standing vis-a-vis other students in the cohort.

References

[1] Alderman,H., Orazem F., and Paterno E. (1996) "School Quality, School Cost, and the Public/Private School Choices of Low-Income Households in Pak-

istan", *World Bank Working Paper Series on Impact Evaluation of Education Reforms*, No. 2.

[2] Allcott, H., and Ortega, D.E. (2007) "The Performance of Decentralized School Systems: Evidence from Fe Y Alegri in Venezuela", *The World Bank mimeo*, Washington D.C.

[3] Andrabi, T., Das, J., and Khawaja, A. (2002a) "Test Feasibility Survey Pakistan: Education Sector", *The World Bank mimeo*, Washington D.C.

[4] Andrabi, T., Das, J., and Khawaja, A. (2002b) "The Rise of Private Schooling in Pakistan: Catering to the Urban Elite or Educating the Rural Poor", *The World Bank mimeo*, Washington D.C.

[5] Angrist, J., Bettinger, E., and Kremer, M. (2002) "Vouchers For Private Schooling In Colombia: Evidence from a Randomized Natural Experiment", *American Economic Review* , 92 :5.

[6] Becker, S.O. and Ichino, O. (2002) "Estimation of Average Treatment Effects Based on Propensity Scores", *The Stata Journal*, 2(4).

[7] Caliendo,M., and Kopeinig, S. (2005) "Some Practical Guidance for the Implementation of Propensity Score Matching", *Dicussion Paper* , No. 1588 Bonn Germany.

[8] Das, J., Pandey, P., and Zajonc, T (2006) "Learning Levels and Gaps in Pakistan", *World Bank Policy Research Working Paper*, 4067 Washington D.C.The World Bank.

[9] Di Prete, T., and Gangl, Markus. (2004) "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments", Working Paper, IZB.

[10] Filmer D. and Pritchett L.(1997) "What Production Functions Show: A Positive Theory of Education Spending", *A World Bank Policy Research Working Paper*, 1975. Washington D.C.

[11] Filmer, Deon and Pritchett, Lant. (1998) "Estimating Wealth Effects without Expenditure Data – or Tears: An Application to Educational Enrollments in States of India", *World Bank Policy Researh Working Paper*, No. 1994

[12] Fuller, B. (1987) "What Factors Raise Achievement in the Third World?", *Review of Educational Research*, No. 57.

[13] Fuller, B., and Heyneman, S. (1989) "Third World School Quality: Current Collapse, Future Potential", *Educational Researcher*, 18:2.

[14] Fuller, B., and Clarke, P.(2002) “Raising School Effects While Ignoring Culture? Local Conditions and the Influence of Classroom, Tools, Rules and Pedagogy”, *Review of Educational Research*, 64:1.

[15] Glewwe, P. (2002) “Schools and Skills in Developing Countries: Education Policies and Socio-Economics Outcomes”, *Journal of Economic Literature* , 40:2.

[16] Hanushek, E. (1986) “The Economics of Schooling: Production and Efficiency in Public Schools”, *Journal of Economic Literature* , Vol. 24.

[17] Harbison, R., and Hanushek, E.(1992)*Education performance of the poor*, London, Oxford University Press.

[18] Heckman, J. and V. J. Hotz. (1989) “Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training.” *Journal of the American Statistical Association* 84(408): 862-874.

[19] Heyneman, S., and Loxley, W.(1983) “The Effect of Primary School Quality on Academic Achievement across Twenty-Nine High-and-Low Income Countries”, *The American Journal of Sociology*, 88:6.

[20] Hirsh-Pasek, K. (1991) “Pressure or Challenge in Pre-School? How Academic Environments Affect Children”, *New Directions for Child Development*, 53.

[21] Hirsh-Pasek, K., Hyson M., and Rescorla L.(1990) “Academic Environments in Preschool: Do they Pressure or Challenge Young Children”, *Education and Development*, 1.

[22] Kim, J., Alderman, H., and Orazem, P.(1998)“Can Private School Subsidies Increase Schooling for the Poor: The Quetta Urban Fellowship Program”, *World Bank Working Paper Series on Impact Evaluation of Education Reforms* , No 11.

[23] Lockheed, M., and Komenan, A. (1989) “Teaching Quality and Student Achievement in Africa: The Case of Nigeria and Swaziland. ”, *Teaching and Teacher Education* , 5:2.

[24] Lockheed, M., and Zhao, Q(1993) “The Empty Opportunity: Local control of Secondary Schools and Student Achievement in the Philippines”, *International Journal of Educational Development*, 13:1.

[25] Miller, L., and Bizzell, R.(2002) *The Louisville Experiment: A comparison of four programs. As the twig is bent: Lasting effects of preschool programs*, New Jersey: Erlbaum.

[26] Ministry of Education, Government of Pakistan (2001) *Education Sector Reform Action Plan 2001-2005.* , Islamabad: GOP.

[27] Ministry of Education, Government of Pakistan (2004) *Public Private Partnerships in the Education Sector: ESR Action Plan 2001-2005* , Islamabad: GOP.

[28] Patrinos, H. (2006) "Public-Private Partnership: Contracting Education in Latin America", *The World Bank mimeo*,Washington D.C.

[29] Rosenbaum, P., and Rubin, D.(1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 10:1.

[30] Rosenbaum, P.(2002) *Observational Studies*, New York, Springer.

[31] Schweinhart, L., and Weikart, D. (1988) "Education for young children living in poverty: Children Initiated Learning or Teacher-Directed Instruction?", *Elementary School Journal*, 89.

[32] Sianesi, B. (2001) "Implementing Propensity Score Matching Estimates with STATA", , Prepared for UK STATA User Group, UK Meeting, London.

[33] Stalling J., and Stipek, D.(1986) "Research on Early Childhood and Elementary School Teaching Programs", in: M.Wittrock (ed.)*Handbook of Research on Teaching* ,New York: Macmillan.

[34] Stipek, D., Feiler R., Daniels D., and Milburn S. (1995) "Effects of Different Instructional Approaches on Young Children's Achievement and Motivation", *Child Development*, 66.

[35] UNESCO (2006) *EFA Monitoring Report 2007: Strong Foundations* ,Paris: UNESCO.

[36] Watson, I.(2005) "The Earnings of Causal Employers: The Problem of Unobservables", *p*resented at the 2005 HILDA Survey Research Conference, University of Melbourne.

APPENDIX

A: Policy change and Student Achievement

The relationship between school quality and student achievement has been extensively analyzed in the education literature. Based upon this policy makers often draw insights and implement useful school quality improving changes. Now we ask the question: Do such kinds of school quality improving policy changes actually impact student achievement and how does this vary under situations with different degrees of policy responsiveness. The model builds in parental effort at home as an important determinant of child achievement along with school quality choice at the household level. We will try and investigate how an exogenous policy shock in favor of raising school quality can alter achievement and parental effort at the optimum.

We consider a simple two period world where parents work in the first period and in the second period receive remittances from their children. Schooling is a cost that parents have to incur in the first period, with expectations of earning returns in the second period. The household thus maximize their utility from consumption in the two periods (δ is the utility weights for each period):

$$U = (1-\delta)f(C_1) + (\delta)g(C_2) \tag{16}$$

Household consumption in the first period is constrained by income earned in the first period (Y_{1p}) and consumption on other goods (O),

$$C_1 = Y_{1p} - pO \tag{17}$$

Income in the first period is derived from: $Y_{1p} = (T-E)w_1$ where T is the total time endowment available for working; E, our variable of interest represents parental effort i.e. the time spent by the parents on child schooling and w_1 is the exogenous wage rate that prevails in period 1. Consumption of other goods is given by: $pO = p_sN + p_g\beta$. This includes schooling where the price of schooling is given by p_s and N is the number of children. Additionally the parents derive satisfaction from consuming a prestige good, β which is defined as the learning efficiency of the child. This in turn is an implicit function of E, parental effort and S, school quality. Finally, household consumption cannot fall below a subsistence level of consumption ($\bar{C}$).

The family's consumption in period 2 is given by the extent of transfers(τ) from their children's income Y_{2c},

$$C_2 = \tau(Y_{2c})N \tag{18}$$

The child's income as an adult depends on his achievement in school in the first period in the following manner:

$$Y_{2c} = rAw_2 \tag{19}$$

where A is the average achievement of children in period 1; r is the rate of return to schooling and w_2 is the exogenous wage rate that prevails in period 2. Achievement is derived from learning efficiency, β; School Quality, S and other fixed inputs, L.

$$A = \beta S^{1-\alpha} L^{\alpha} \tag{20}$$

A final constraint is on school quality *choice*.

$$S = h(P, I) \qquad S'_P > 0 \quad and \quad S'_I < 0 \tag{21}$$

Parents maximize over school quality S in the sense that they can choose to send their children to higher or lower quality schools. Their choice is determined by the relevant education policy which regulates public (sometimes private) school adminisration, function etc. This is given by a policy index P which represents a composite of policies according to their impact on school quality. Thus a high score on the policy index would imply the presence of an aggregation of policies best suited for raising school quality. Additionally, the household are restricted in the choice of schools by their own individual constraints which are weighted through a similar index I. Thus, $S = S(P, I)$ where; $S'_P > 0$ and $S'_I < 0$.
This presents the following optimization problem for the households:

$$max \qquad U = (1-\delta)f(C_1) + (\delta)g(\tau N * rAw_2) \tag{22}$$

$$s.t. \qquad C_1 = (T - E)w_1 - p_s N - p_g \beta \tag{23}$$

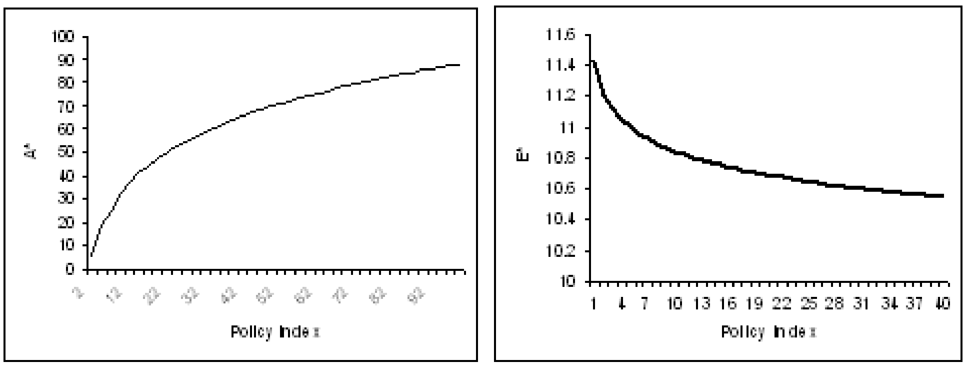
$$C_1 > \bar{C} \tag{24}$$

$$A = \beta S^{1-\alpha} L^{\alpha} \tag{25}$$

$$S = h(P, I) \tag{26}$$

This gives us the optimum levels of achievement, parental effort, and school quality, A^*, E^*, and S^* respectively. We are interested in understanding the behavior of these optimum values in response to an exogenous change in the policy index. Since the policy change is such that it leaves individual constraints unaffected, we can assume (for simplicity) individual constraints to be zero. We calibrate the model for a representative household[18] by varying the values of P ranging from 1 to 100 to examine how achievement and parental effort vary at optimum as a response to this exogenous policy intervention. Figure 1. depicts the numerical simulations.

Figure 1: Effects of Policy Changes on A^* and E^*

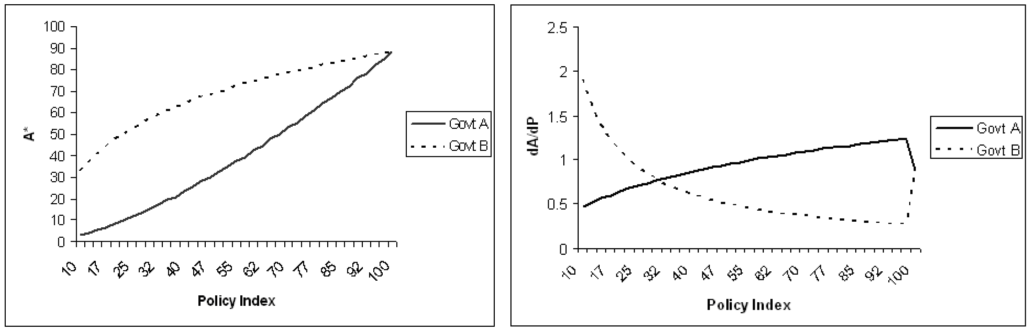


It can be seen that as the policy index increase i.e governments choose to invest in policies which have the most favorable impact on school quality, achievement increases and parental effort decreases. Thus a favorable policy has increased achievement at the optimum precisely because it has allowed for an increase in school quality (choice) which raises the learning efficiency of the child independent of parental effort. In fact it is seen, that as a response parents can afford to relax their efforts.

We are now in a position to analyze the behavior of a policy intervention with respect to two scenarios. One, where school quality is extremely responsive to a policy change impacting it (Govt. A), and the second where school quality is only moderately responsive (Govt B.). We show below that achievement gains are more rapidly achieved where school quality is most responsive to policy change or where the government enact the policy change in an environment where initial school quality is found to be quite low. Figure 2., shows the optimal levels of achievements vis-a-vis policy changes with respect to the two scenarios, as well as its corresponding derivatives.

[18]choosing the following parameter values: $\delta = 0.5$, $\tau = 0.4$, $N = 4$, $r = 0.7$, $w_2 = 300$, $T = 18$, $w_1 = 100$, $p_s = 10$, $p_g = 100$, $\bar{C} = 600$, $L = 70$, $\alpha = 0.5$

Figure 2: Policy Responsiveness vis-a-vis Achievement and corresponding derivatives



In this sense our empirical results are justifiable viz. the small gains in achievement, given the context of Islamabad, where we find that school quality has been typically high both over the past and relative to other provinces. This implies also that any policy change that seeks to impact school quality in an environment where it has been generally low, should bring about achievement gains in children at an increasing rate.

B: Tables and Figures

Figure 3: Kernel Density Distribution of Scores - CRI and Non CRI

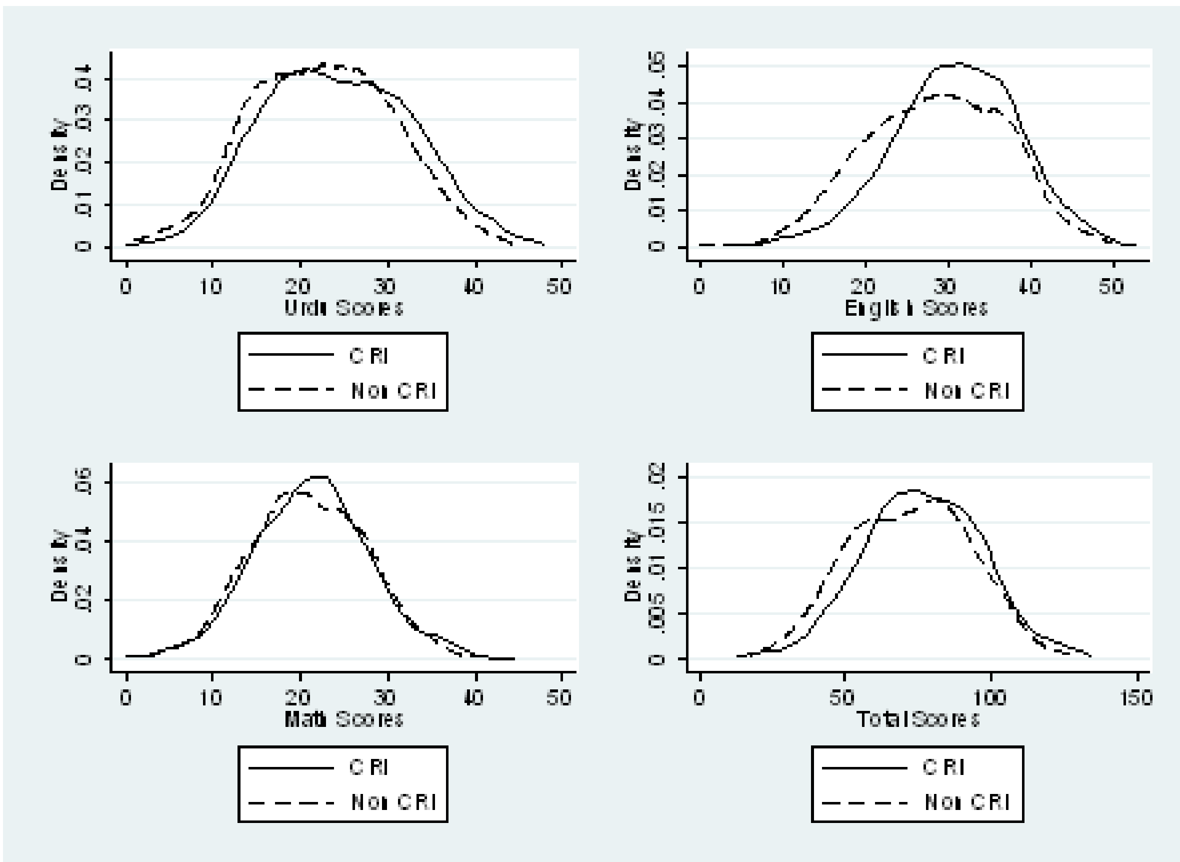


Figure 4: Mean Age of Children by Blocks - CRI and Non CRI

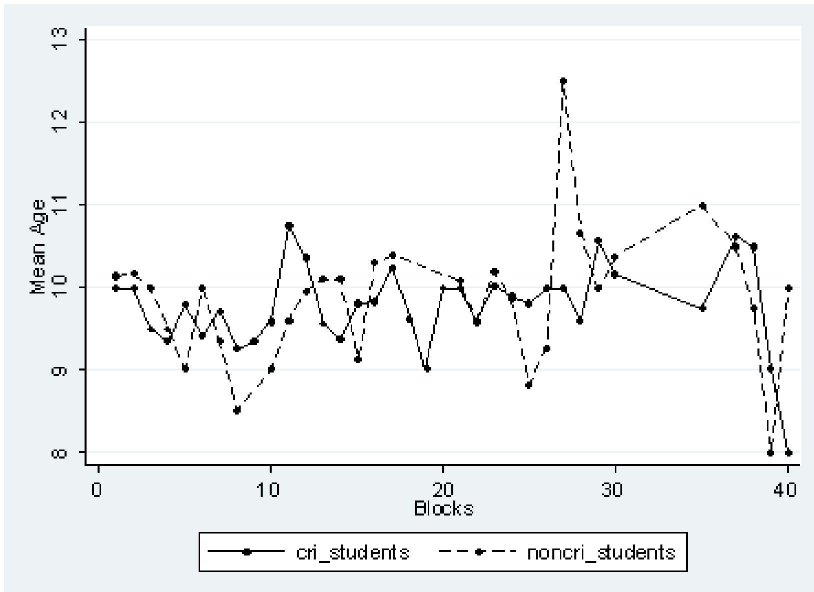


Figure 5: Mean Assets of Children by Blocks - CRI and Non CRI

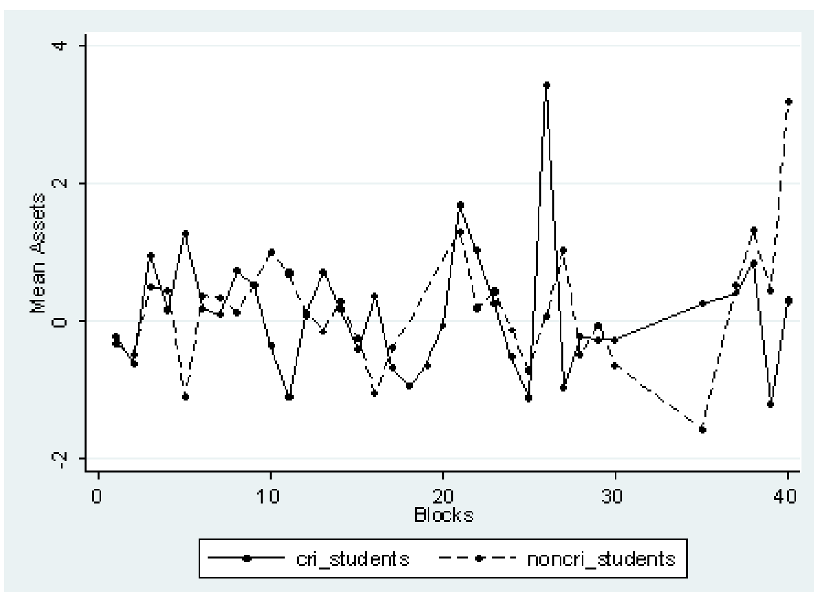


Figure 6: Box Plots of Age by Blocks - CRI and Non CRI

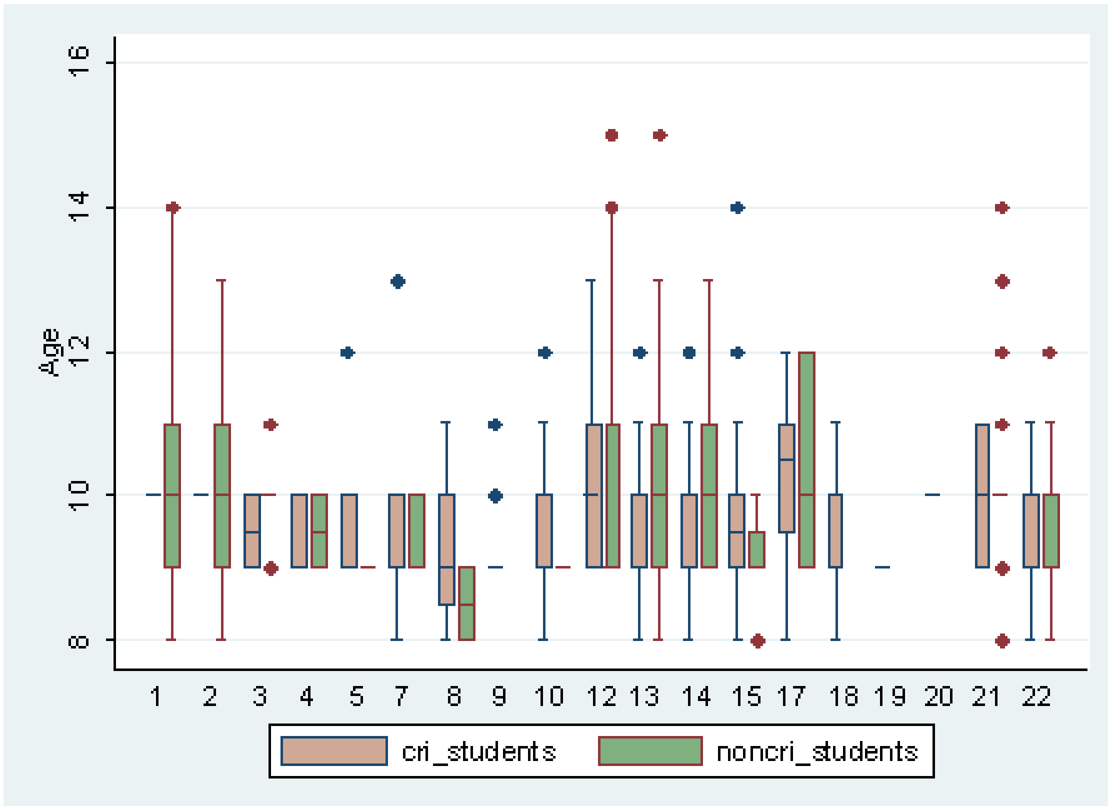


Table 1: Number of Teachers Trained under CRI in Islamabad (2002-2006)

Teachers trained (2002-2006)	**Grade**	**Numbers of Trained**
Teachers	KG to III	304
-do-	IV	142
-do-	V	169
School Heads and Deputy Heads	All school	173
Total Number of Teachers Trained		**788**

Table 2: Descriptive Statistics

Variable	CRI Mean	CRI SD	Non-CRI Mean	Non-CRI SD	Diff Mean	Diff SE
Child						
Male†	0.53	0.50	0.56	0.50	-0.03	0.02
Age in years	9.91	1.26	10.46	1.61	-0.56**	0.07
Mother went to school†	0.23	0.42	0.24	0.42	-0.10	0.02
Father is professional†	0.61	0.49	0.53	0.50	0.07**	0.02
Father is entrepreneur†	0.15	0.36	0.17	0.38	-0.02	0.02
Number of siblings	3.42	1.79	3.81	1.95	-0.39**	0.09
Number of older siblings	1.74	1.59	1.86	1.72	-0.13	0.08
Asset index	0.25	2.19	-0.14	2.22	0.40**	0.10
Child height	137.74	8.73	139.09	10.02	-1.35*	0.44
Child weight	63.46	16.48	65.88	17.80	-2.42*	0.80
Teacher						
Teacher academic qual.† (FA/FSc or above)	0.97	0.17	0.87	0.33	0.09**	0.01
Teacher experience (yrs)	8.62	7.28	8.02	7.28	0.60	0.33
School						
English medium†	0.87	0.33	0.57	0.49	0.30**	0.02
Isl. Model College†	0.24	0.43	0.22	0.42	0.02	0.02
F.G. Model School†	0.63	0.48	0.49	0.50	0.14**	0.02
Boys school†	0.10	0.30	0.22	0.41	-0.12**	0.02
Co-educational†	0.50	0.50	0.49	0.50	0.01	0.02
Urdu score	24.26	8.14	21.48	8.23	2.78*	0.38
Math score	21.36	6.42	19.99	6.81	1.37**	0.31
English score	30.81	7.49	27.44	8.53	3.37**	0.38
Total score	76.43	19.65	68.91	21.58	7.53*	0.97
Urdu score (IRT)	26.55	5.68	24.24	6.00	2.31*	0.28
Math score (IRT)	24.21	4.01	22.55	4.28	1.65**	0.19
English score (IRT)	31.39	5.97	28.96	6.31	2.43**	0.29
Total score (IRT)	82.14	15.62	75.76	16.59	6.39*	0.75
Average attendance	0.86	0.09	0.88	0.08	-0.02**	0.00
Number of Observations	741		1243			

Note: This simple comparison of means is done after pooling observations from all the different blocks together. Thus, as expected, CRI and non-CRI children are not evenly matched along several dimensions as indicated by significant differences in the mean values of covariates.

† indicates dummy variable; ** indicates significance at 1% level.

Table 3: Average Treatment Effect: Full Sample (contaminated)

	Treatment Effect (atts)	T-stat	95% Confidence Interval	
Urdu	1.817*	2.436	0.318	3.317
Maths	0.441	1.094	-0.369	1.252
English	2.634**	4.821	1.536	3.732
Total score	**4.893****	**4.622**	**2.765**	**7.020**
Avg. Attendance	0.047*	2.496	0.009	0.084
	Treatment	Control	Total	
No. of observations	609	579	1188	

Notes:
1. The last two columns report confidence intervals obtained using bootstrapped standard errors.
2. The total number of observations used in computing a non-parametric estimator of the treatment effects, **atts** above as compared to other estimators reported in the next table, might vary depending on the number of individuals that match on observable attributes.

** indicates significance at 1% level
* indicates significance at 5% level
$^{+}$ indicates significance at 10% level

Table 4: Average Treatment Effect: Full Sample IRT scores (uncontaminated)

	Treatment Effect (atts)	T-stat	95% Confidence Interval	
Urdu	2.00**	3.957	0.985	3.016
Maths	1.421**	4.623	.803	2.040
English	2.131**	4.528	1.185	3.076
Total score	**5.553****	**4.277**	**2.944**	**8.161**
Avg. Attendance	0.042*	2.189	.003	.081
	Treatment	Control	Total	
No. of observations	490	298	788	

Notes:
1. The last two columns report confidence intervals obtained using bootstrapped standard errors.
2. The total number of observations used in computing a non-parametric estimator of the treatment effects, **atts** above as compared to other estimators reported in the next table, might vary depending on the number of individuals that match on observable attributes.

** indicates significance at 1% level
* indicates significance at 5% level
+ indicates significance at 10% level

Table 5: Sensitivity Analysis

	attnd	attr	attk
Urdu	0.899^{+}	0.842*	0.950**
	(1.739)	(2.079)	(2.749)
Maths	0.666^{+}	0.626^{+}	0.702*
	(1.969)	(1.990)	(2.676)
English	0.942	0.878^{+}	0.994*
	(1.645)	(1.791)	(2.622)
Total score	**2.508***	**2.346***	**2.647***
	(2.095)	(2.352)	(2.452)
Avg. attendance	0.033**	0.006	0.030**
	(3.418)	(1.232)	(4.042)
# Treatment	499	499	499
# Control	168	326	326

Notes: 1. **attnd** denotes the Nearest neighbor matching estimator based on the estimated propensity scores; **attr** the radius matching with the acceptable radius set at 0.1 and **attk** denotes the kernel matching estimator based on a weighting scheme that uses Gaussian kernel.
2. Bootstrapped t-stats reported for all the results.

** indicates significance at 1% level
* indicates significance at 5% level
$^{+}$ indicates significance at 10% level

Table 6: Average Treatment Effect: Restricted Sample IRT scores (uncontaminated)

	Treatment Effect (atts)	T-stat	95% Confidence Interval	
Urdu	1.376*	2.584	0.305	2.446
Maths	0.987*	2.393	0.158	1.816
English	1.462*	2.472	0.273	2.650
Total score	**3.826***	**2.485**	**0.732**	**6.920**
Avg. Attendance	0.011	1.031	-0.010	0.032
	Treatment	Control	Total	
No. of observations	340	160	500	

Notes:
1. The last two columns report confidence intervals obtained using bootstrapped standard errors.
2. The total number of observations used in computing a non-parametric estimator of the treatment effects, **atts** above as compared to other estimators reported in the next table, might vary depending on the number of individuals that match on observable attributes.

** indicates significance at 1% level
* indicates significance at 5% level
$^{+}$ indicates significance at 10% level

Table 7: Sensitivity Analysis: Restricted Sample

	attnd	attr	attk
Urdu	1.323*	0.813	0.619
	(2.269)	(1.546)	(1.333)
Maths	0.950*	0.598	0.467
	(2.356)	(1.531)	(1.370)
English	1.411*	0.849	0.644
	(2.325)	(1.430)	(1.073)
Total score	**3.684**$^+$	**2.260**$^+$	**1.730**
	(1.856)	(1.695)	(1.071)
Avg. attendance	0.019$^+$	0.011	0.022**
	(1.973)	(1.225)	(3.735)
# Treatment	340	340	340
# Control	98	177	177

Notes: 1. **attnd** denotes the Nearest neighbor matching estimator based on the estimated propensity scores; **attr** the radius matching with the acceptable radius set at 0.1 and **attk** denotes the kernel matching estimator based on a weighting scheme that uses Gaussian kernel.
2. Bootstrapped t-stats reported for all the results.

** indicates significance at 1% level
* indicates significance at 5% level
$^+$ indicates significance at 10% level

Table 8: Rosenbaum Bounds

Gamma	Urdu	Maths	English	Total score
1	2.20E-15	1.10E-16	3.80E-15	2.20E-15
1.05	4.60E-14	1.60E-15	7.40E-14	4.60E-14
1.1	6.90E-13	2.70E-14	1.10E-12	6.90E-13
1.15	7.90E-12	3.60E-13	1.20E-11	7.90E-12
1.2	7.20E-11	3.80E-12	1.10E-10	7.20E-11
1.25	5.30E-10	3.20E-11	8.00E-10	5.30E-10
1.3	3.30E-09	2.20E-10	4.80E-09	3.30E-09
1.35	1.70E-08	1.30E-09	2.50E-08	1.70E-08
1.4	7.80E-08	6.60E-09	1.10E-07	7.80E-08
1.45	3.10E-07	2.90E-08	4.30E-07	3.10E-07
1.5	1.10E-06	1.10E-07	1.50E-06	1.10E-06
1.55	3.50E-06	4.00E-07	4.80E-06	3.50E-06
1.6	0.00001	1.30E-06	0.000014	0.00001
1.65	0.000027	3.70E-06	0.000036	0.000027
1.7	0.000067	9.80E-06	0.000087	0.000067
1.75	0.000153	0.000024	0.000197	0.000153
1.8	0.000327	0.000057	0.000418	0.000327
1.85	0.000661	0.000124	0.000833	0.000661
1.9	0.001262	0.000254	0.001573	0.001262
1.95	0.00229	0.000496	0.002824	0.00229
2	0.003965	0.000921	0.004839	0.003965
2.05	0.006572	0.001635	0.007942	0.006572
2.1	0.010468	0.002782	0.012526	0.010468
2.15	0.016064	0.004553	0.019041	0.016064
2.2	0.023818	0.007187	0.027972	0.023818
2.25	0.034207	0.010968	0.039814	0.034207
2.3	0.047692	0.016224	0.055029	0.047692
2.35	0.06469	0.023307	0.074014	0.06469
2.4	**0.08553**	0.032587	**0.09706**	**0.08553**
2.45	0.110429	0.044419	0.124326	0.110429
2.5	0.139461	0.059132	0.155812	0.139461
2.55	0.172547	0.076997	0.191354	0.172547
2.6	0.20945	**0.098211**	0.230623	0.20945
2.65	0.249781	0.122877	0.273142	0.249781
2.7	0.293024	0.150992	0.318312	0.293024
2.75	0.338556	0.182441	0.365441	0.338556

Table 9: Average Treatment Effect by Gender (uncontaminated)

A. Girls

	Treatment Effect (atts)	T-stat	90% Confidence Interval	
Urdu	2.253	1.494	-0.276	4.782
Maths	1.616^{+}	1.696	0.185	3.213
English	2.395^{+}	1.951	0.336	4.453
Total score	**6.264**$^{+}$	**1.841**	**0.558**	**11.96**
Avg. Attendance	0.022	1.661	-0.000	0.043
	Treatment	Control	Total	
No. of observations	180	82	262	

B. Boys

	Treatment Effect (atts)	T-stat	90% Confidence Interval	
Urdu	2.410**	3.633	1.297	3.521
Maths	1.716**	3.346	0.855	2.575
English	2.554**	3.133	1.187	3.919
Total score	**6.679****	**3.365**	**3.351**	**10.006**
Avg. Attendance	-0.027^{+}	-1.689	-0.054	-0.000
	Treatment	Control	Total	
No. of observations	129	223	352	

Notes:
1. The last two columns report the confidence intervals obtained using bootstrapped standard errors, reported above.
2. The total number of observations used in computing a non-parametric estimator of the treatment effects, **atts** above as compared to other estimators reported in the next table, might vary depending on the number of individuals that match on observable attributes.

** indicates significance at 1% level
* indicates significance at 5% level
$^{+}$ indicates significance at 10% level

Table 10: Average Treatment Effect by Wealth (uncontaminated)

A. Top 35% of the Wealth Distribution

	Treatment Effect (atts)	T-stat	90% Confidence Interval	
Urdu	1.977**	2.912	0.83	3.115
Maths	1.434**	3.678	0.780	2.088
English	2.090**	2.312	0.575	3.605
Total score	**5.501****	**3.024**	**2.451**	**8.551**
Avg. Attendance	0.005	0.309	-0.023	0.033
	Treatment	Control	Total	
No. of observations	149	123	272	

A. Bottom 35% of the Wealth Distribution

	Treatment Effect (atts)	T-stat	90% Confidence Interval	
Urdu	1.383	1.107	-0.712	3.479
Maths	0.948	1.105	-0.491	2.387
English	1.501	0.950	-1.148	4.150
Total score	**3.832**	**1.126**	**-1.875**	**9.540**
Avg. Attendance	-0.023	-0.759	-0.075	0.028
	Treatment	Control	Total	
No. of observations	81	72	153	

Notes:
1. The last two columns report the confidence intervals obtained using bootstrapped standard errors, reported above.
2. The total number of observations used in computing a non-parametric estimator of the treatment effects, **atts** above as compared to other estimators reported in the next table, might vary depending on the number of individuals that match on observable attributes.

** indicates significance at 1% level
* indicates significance at 5% level
$^{+}$ indicates significance at 10% level

Table 11: School and Child Selection Probits

Dependent variable:	CRI	
	(a) School Level (Cluster F.E.)	(b) Child Level School Block F.E.
F.G. Model School†	-0.65	–
	(0.96)	
Isl. Model College†	-1.50*	–
	(0.86)	
English medium†	2.61*	–
	(1.08)	
Co-educational†	-1.52*	–
	(0.70)	
Boys school†	-0.43	–
	(0.78)	
Male†	–	0.09+
		(0.09)
Age in years	–	-0.16*
		(0.045)
Mother went to school†	–	-0.05
		(0.12)
Teacher academic qual. †	–	1.64**
		(0.20)
Teacher experience (yrs)	–	0.09**
		(0.01)
Father is professional†	–	0.04
		(0.11)
Father is entrepreneur†	–	0.06
		(0.15)
Number of siblings	–	-0.07*
		(0.03)
Number of older siblings	–	0.05**
		(0.04)
Asset index	–	0.00**
		(0.014)
Child height	–	0.025
		(0.01)
Child weight	–	-0.014
		(0.003)
Observations	**77**	**990**

Notes: † indicates dummy variable; z-statistics in parentheses; ** indicates significance at 1% level * indicates significance at 5% level; + indicates significance at 10% level

Table 12: Verifying the Balancing Property

Dependent variable:	CRI	
	(a) Within school blocks	(b) Within child blocks
F.G. Model	0.054	–
	(0.12)	
Isl Model College	0.121	–
	(0.32)	
Whether boys school?	0.082	–
	(0.24)	
Whether Co-ed school?	-0.083	–
	(0.21)	
Male	–	0.022
		(0.47)
Age of child (in years)	–	-0.003
		(0.15)
Mother went to school?	–	0.014
		(0.31)
Teacher experience	–	0.006^{+}
		(1.69)
Whether father is a professional?	–	-0.008
		(0.18)
Whether father is entrepreneur?	–	-0.004
		(0.07)
Number of siblings	–	0.001
		(0.10)
Number of elder siblings	–	-0.002
		(0.14)
Asset index	–	0.000
		(0.03)
Child height	–	0.002
		(0.66)
Child weight	–	-0.002
		(0.92)
Observations	46	511
Number of matching school blocks	4	–
Number of matching child blocks	–	20
F-test for x-variables: p-value	0.99	0.98

Notes: Absolute value of t statistics in parentheses. $^{+}$ indicates significance at 10%; * at 5%; ** at 1%. Regression (a) includes cluster i.e. geographical controls for different areas within Islamabad.