Failure of Frequent Assessment: An Evaluation of India's Continuous and Comprehensive Evaluation Program

James Berry, Harini Kannan, Shobhini Mukherji, and Marc Shotland*

November 2018

Abstract

Continuous assessment programs, which emphasize frequent evaluation over infrequent high-stakes testing, have been implemented in many developing countries as a way to improve learning outcomes. In India, continuous assessment is mandated in all primary schools through the Continuous and Comprehensive Evaluation (CCE) Program. We conduct a randomized evaluation of CCE in the state of Haryana. We find that CCE had no significant effects on test scores, and we can rule out effects above 0.1 standard deviations. We present evidence that the lack of impacts was due to the program's focus on evaluation without linking evaluations to changes in teaching practices.

* Berry: University of Delaware, jimberry@udel.edu; Kannan: J-PAL, harini.kannan@ifmr.ac.in; Mukherji: J-PAL, shobhini.mukherji@ifmr.ac.in; Shotland: IDinsight, marc.shotland@idinsight.com. We thank Esther Duflo for insights throughout the study and Adrienne Lucas for helpful comments on the manuscript. We also are grateful to John Firth, Kanika Dua, Nina Harari, Shweta Rajwade, and Perwinder Singh for their superb work coordinating research activities in the field and to Harris Eppsteiner and Madeline Duhon for excellent research assistance. This research was funded by the Government of Haryana, by the Regional Centers for Learning on Evaluation and Results, and by the International Initiative for Impact Evaluation. The experiment is registered in the AEA RCT Registry, ID AEARCTR-0000008. All errors are our own.

1 Introduction

In many developing countries, primary school enrollment rates have risen to near-universal levels in recent years, yet learning levels have not matched this progress. In India, for example, a 2016 survey found that only 48 percent of fifth-graders could read a second-grade level text proficiently, even though 98 percent of children aged 7 to 10 were enrolled in school (ASER Centre 2017). Furthermore, as students progress through school, those who lag behind in early grades continue to fall further and further behind. Many do not possess even basic skills at the end of eight years of primary education.

One source of this problem is that in many developing countries, teachers progress through an overly demanding curriculum, regardless of whether students understand it or not. Students who start with a small gap quickly get lost, as teachers focus on the best performing students and move through the syllabus (Banerjee and Duflo 2011). A major need for reform of education systems in developing countries is to get teachers to focus more closely on what children in their classrooms know and what they can understand. When this is done, the learning gains are large (Banerjee et al. 2017).

One mechanism that may prevent teachers from focusing on students' actual learning levels is the salience of infrequent, high-stakes examinations based on the prescribed grade-level syllabus. With these exams as both their focus and the primary source of information about their students' abilities, teachers may lack adequate and timely information about students' educational achievement to address their individual needs. Thus, a common view among education scholars is that educational outcomes could improve if infrequent, high-stakes examinations were replaced by continuous assessment, in which teachers frequently evaluate students. The idea is that teachers would then be prompted to adapt teaching methods to focus on their students' specific learning

needs. Continuous assessment programs have been implemented in a number of developing countries, including Albania, Brazil, Ethiopia, and Morocco (UNESCO 2008). However, to date there has been no rigorous evaluation of such a system. Thus, we do not know either whether continuous assessment systems do in fact lead to either changes in teaching practices or to learning gains.

This paper fills this gap by conducting a large-scale randomized controlled trial in collaboration with one state government in India. Prior to 2009, the evaluation system in Indian primary schools—covering grades 1 through 8—followed the typical pattern of reliance on infrequent, high-stakes exams. In 2009, the Right to Education Act initiated a "no detention" policy, eliminating the use of high-stakes exams to determine promotion to the next grade. As a replacement for high-stakes exams as a means to evaluate students, the Right to Education Act mandated a system of "Continuous and Comprehensive Evaluation" (CCE; Government of India 2009). In the CCE framework, teachers are trained on how to frequently evaluate students using a variety of methods, along both academic and non-academic dimensions. A key component underlying the CCE's theory of change is that better tracking of children would allow (and lead) teachers to customize their teaching based on the current learning levels of individual students (Bhatia 2009).

We partnered with the Government of Haryana, a state in northwest India, to conduct a randomized controlled trial to evaluate the impact of CCE on student achievement in math and language before the statewide roll out of the program. Four hundred lower primary schools (grades 1 to 5) and 100 upper primary schools (grades 6 to 8) were assigned at random either to implement the CCE program or to be in a control group during the 2012-13 academic year. In lower primary schools, the CCE program was cross-randomized with a targeted instruction program which

combined simple assessment with ability grouping and targeted teaching activities, known as the "Teaching at the Right Level" Program, or TaRL. We report the main results of the TaRL program in Banerjee et al. (2017), in which we show that the TaRL program was effective in improving test scores. We use the results of the concurrent evaluation of TaRL with CCE in this paper to draw comparisons between the programs and better understand the functioning of CCE.

Our main finding is that CCE uniformly failed to improve test scores. Students in CCE schools did not perform significantly better at endline than students in control schools on either language or math tests, whether in lower or upper primary schools. The impacts are precisely estimated and allow us to rule out, with 95 percent confidence, effect sizes above 0.07 standard deviations in the lower primary sample and 0.11 standard deviations in the upper primary sample.

We then use our process evaluation data to understand potential sources of the lack of impacts of CCE. The vast majority of teachers attended CCE training, and during random school visits about two-thirds reported using CCE evaluation sheets and report cards, key activities of the program. However, the program did not lead to any differential use of evaluation data or changes in teaching practices recommended by the program. We argue that this was a result of the program's focus on the administrative components of evaluation, without a clear and explicit link to teaching practices. In effect, CCE became yet another burdensome administrative requirement, rather than a teaching aide. This contrasts with the TaRL program, which calls for simple assessment, but focuses on linking that assessment to level-appropriate teaching activities.

Our results have several policy implications. Most directly, the current implementation of India's flagship CCE policy may not be achieving its core objective of increasing learning outcomes of students. More generally, our results imply that the link between conducting

assessment and using assessment to improve learning is crucial for the success of continuous assessment programs.

The remainder of this paper is organized as follows. Section 2 describes the context and intervention; Section 3 provides an overview of our evaluation design; Section 4 discusses data sources and student testing; Section 5 presents impact results; Section 6 discusses potential mechanisms; and Section 7 concludes.

2 Background and Context

2.1 Continuous and Comprehensive Evaluation (CCE)

The CCE scheme was designed to provide teachers—as well as students and parents—with frequent and broad-based feedback on performance. The primary aim is to allow teachers to customize their teaching based on the current learning levels of individual students. To this end, CCE's mode of assessment is meant to be "continuous," in that teachers identify students' learning progress at regular intervals on small portions of content (such as a single module or lesson). This regular assessment can incorporate a variety of techniques, including unit tests, projects, and evaluation of class participation. In addition, CCE prescribes a more "comprehensive" assessment of student achievement than traditional testing: it assigns scores not only on the basis of scholastic performance, but also on the basis of co-scholastic activities (such as arts, music, or athletics) and personality development as reflected in life skills, attitudes, and values.

The frequent evaluation of CCE is designed to enable teachers to closely monitor student progress, better tailor their teaching to their students' needs, and to facilitate identification and remediation of learning deficits. Because the assessments are continuous and low-stakes, CCE is

also meant to reduce the stress of preparing for major exams, which could lead to increased retention in school.

CCE's continuous assessment methodology is based on previous work in education research emphasizing the importance of the use of *formative* assessments—those that provide frequent feedback to inform teachers' classroom practices, as well as parents and students (Black and Wiliam 2009). While continuous assessment programs have been promoted to improve learning in developing countries, (UNESCO 2008), evidence on effectiveness of these programs is scarce. Studies in Malawi (Kamangira 2003) and Zambia (Kapambwe 2010) provide suggestive evidence that such programs increase test scores using retrospective comparisons of program and non-program schools. To our knowledge, there are no rigorous large-scale evaluations of such programs in a developing country.[1,2] Our setting is also particularly policy relevant because CCE has been implemented as a national policy in India, the country with the largest population of primary-aged students in the world.

2.2 The CCE Program in Haryana

Although mandated by the Indian central government, details of the design and implementation of CCE were made the responsibility of state-level education departments. We evaluate the rollout of CCE in the state of Haryana. Haryana ranks third-highest among Indian states in per capita income (Reserve Bank of India 2011). While student learning levels in Haryana are similarly

[1] In the United States, evidence on the effectiveness of formative assessment is mixed, and studies also suffer from methodological limitations (Bennett 2011).

[2] Our study is related to the recent literature in development economics evaluating diagnostic feedback interventions, in which test scores are periodically provided to schools (Muralidharan and Sundararaman 2010; de Hoyos Navarro, Garcia-Moreno, and Patrinos 2017; de Hoyos, Ganimian, and Holland 2017). The CCE program differs from these interventions by utilizing continuous rather than one-off or infrequent assessment. It also assesses achievement along a wide range of dimensions, whereas diagnostic feedback interventions have focused primarily on math and language skills.

higher than the national average, they are still low relative to the prescribed syllabus: at the time of our study, 34 percent of students in grade 5 in Haryana could not read a grade-2 level text, compared with 52 percent of students in India overall (ASER Centre 2012).

Haryana's state education system—consisting of about 15,000 schools, 90,000 teachers, and two million students (NUEPA, 2013)—is structured similarly to government school systems in other states in India. As with other states, policy and curricular decisions, such as syllabus development, the content of teacher training, and textbooks, are made at the state level.

Following the guidelines of the national CCE program, the Haryana program emphasized broad-based, frequent, and varied forms of evaluation. In turn, use of these evaluations was intended to allow teachers to more easily identify low-performing students and reduce students' stress from end-of-year exams. In the Haryana program, frequent assessments were operationalized through the use of evaluation sheets for recording evaluations of students. Evaluation sheets were to be completed every month (or quarterly for grades 6 to 8). An example evaluation sheet is provided in the Supplemental Appendix. Students were evaluated based on both academic topics, such as language and mathematics, as well as co-curricular topics such as creativity, sports, and personal qualities. Each topic contained a number of skills and sub-skills that students were expected to master. Rubrics were provided for each sub-skill to form the basis of assessment.

In addition to the preparation of evaluation sheets and report cards, Haryana's CCE guidelines called upon teachers to use results of evaluations to identify low performing students and address learning deficits. This could take the form of modifying lesson plans as a result of evaluations, changing in-class teaching practices, or more actively gauging students' understanding during class. If students did not understand a particular concept, teachers were encouraged to repeat the

concept, provide an alternative explanation, or use examples relevant to the students. While these practices were encouraged, the guidelines did not include strategies specific to particular topics or skills, nor did they emphasize particular areas of focus.

Twice per year, students' performance was to be summarized via report cards that were shared with parents. In grades 1 to 5, these report cards contained summary descriptive remarks, but no grades. In grades 6 to 8, letter grades were provided in addition to descriptive remarks.

Teacher training for the CCE program in Haryana was conducted over seven days by two education training companies partnering with the government. Following the state guidelines, much of the training time focused on conducting CCE evaluations and filling out evaluation sheets and report cards. There was also limited instruction on modifying teaching practices based on evaluation data. Schools in the CCE treatment arm were provided materials such as manuals, evaluation sheets, and report cards in order to implement the program.

Schools in our evaluation that were not selected to implement CCE followed the same evaluation practices that had been in place prior to the rollout of CCE. It should be noted, however, that Haryana eliminated the use of end-of-year exams for the purpose of grade promotion shortly after the passage of the Right to Education Act in 2009. Therefore, our evaluation estimates the effects of CCE relative a system without these end-of-year exams. Still, periodic testing was the most common method of student evaluation in the control schools, as presented in Section 6.1 and Table 8 below.

2.3 Cross-cutting Teaching at the Right Level Program

Alongside the CCE intervention, the research team evaluated a second intervention, the TaRL program, that was also implemented by teachers in the study schools. Designed by the nongovernmental organization Pratham, the TaRL program incorporated a simple assessment,

grouping of students by initial learning level, and teaching according to level-appropriate learning activities and materials, rather than the prescribed syllabus. To emphasize to teachers that the program represented a break from teaching the standard curriculum, an hour per day was set aside for teaching according to the TaRL program. We present the main evaluation of the TaRL program in Banerjee et al. (2017). Some impact results of TaRL are also presented in this paper to draw contrasts with those of CCE. We also evaluate potential complementarities between the two programs in Appendix A.

2.4 ABRC Monitoring

Since CCE was a new program that had not previously been proven effective, the research team stressed the importance of monitoring and management to the Government of Haryana. The Government in turn requested that the researchers help adjust the existing school-level monitoring structure to facilitate the monitoring of CCE. Working with the state administration, the researchers set up a mentoring and monitoring program using field-level supervisors, known as Associate Block Resource Coordinators (ABRCs). ABRCs were trained on general mentoring and monitoring of teachers in all schools in the evaluation. They were also trained to serve as resources for teachers in CCE and TaRL schools and to collect data on for the purposes of program administration. Each ABRC covered between 10 and 15 schools. As ABRCs were typically responsible for schools in both treatment and control groups, they were trained to supervise in each school according to its assigned group.

3 Evaluation Design

To estimate the impact of the CCE program, we incorporated a randomized-controlled-trial design. In consultation with the Government of Haryana, two districts were selected for the study:

the relatively more developed northern district of Kurukshetra and the less developed southern district of Mahendragarh. Within these two districts, four blocks (district subdivisions) were selected at random as our intervention sites. Across these four blocks, a total of 500 schools—400 lower primary schools (grades 1 to 5) and 100 upper primary schools (grades 6 to 8)—were randomly drawn from a list of all government schools.[3]

Following baseline testing conducted during the 2011-12 academic year, the 400 lower primary schools were randomly assigned in equal proportions to one of four groups: 1) CCE alone, 2) TaRL alone, 3) both CCE and TaRL, and 4) no additional intervention. The 100 upper primary schools were randomly assigned either to receive CCE alone or to a control group.

Appendix Figure 1 depicts the progression of the sample into treatment and control groups. In our study areas, schools of different levels or those serving different genders sometimes shared the same grounds or even the same building. Due to the possibility of spillovers between schools sharing a campus, the randomization was conducted on the level of the school campus: a group of schools at different levels in the same locality, usually occupying a single building or complex of buildings. Across the 500 study schools, there were 384 such campuses. Randomization was stratified by first matching campuses by block and whether the campus contained lower primary grades, upper primary grades, or both. Within these matches, campuses were sorted by average baseline test scores to create strata of four campuses each. Campuses in each stratum were then randomly assigned across the four treatment and control groups. Because TaRL was not conducted in upper primary schools, upper primary schools in campuses initially assigned to TaRL alone or CCE and TaRL were re-assigned to the control group or CCE alone, respectively.

[3] Selection of primary schools for the sample took place in two stages in late 2011 and early 2012. See Appendix B for additional detail.

4 Data

4.1 Data Sources

While CCE encompasses a variety of curricular and co-curricular elements, a core objective is improvement in basic academic skills. These skills were also identified by the Government of Haryana as their main outcomes of interest at the outset of the study. As a result, our primary source of data comes from a series of tests in Hindi and math. Baseline testing took place in the 2011-12 school year, before implementation of the programs, and endline testing took place at the end of the 2012-13 school year, following implementation in schools assigned to the treatment groups.

In the 400 lower primary schools, the sample of tested students consisted of students who were in grades 1 to 4 at baseline and were exposed to the program in grades 2 to 5. Tests were administered to 10 randomly selected students in each grade in each school at baseline, yielding a lower primary school sample of 12,576 students.[4] In the 100 upper primary schools, our sample consisted of all students in 7th grade in each school at baseline (exposed to the intervention as 8th graders), for a total of 3,261 students. All tests were administered by research staff during school hours within schools. For all students in the sample, we also collected basic demographic data, including gender and age, as well as records of recent school attendance from school registers in each round of testing.

In lower primary schools, the tests included both oral and written tests in Hindi and math. The oral tests were developed by the ASER Centre for use in its Annual Status of Education Report, a

[4] In a number of the 400 primary schools in our sample, there were fewer than 10 children in certain grades. In these cases, all of the children in the grade in question were sampled, yielding a total sample size of less than 10*4*400=16,000 students.

national assessment of basic learning levels across India. In reading, the ASER test covered skills ranging from letter recognition (a grade 1 skill, according to the Indian primary school curriculum) through reading a simple text fluently (a grade 2 skill). In math, the test began with number recognition and progressed to simple division (a grade 4 skill) (ASER Centre 2012).

Lower primary school students in grade 3 or 4 at baseline were also administered written Hindi and math assessments. The written tests, developed by the researchers and Pratham for a previous evaluation of Pratham's "Read India" program in the states of Bihar and Uttarakhand (Banerjee et al. 2016), assessed students on competencies which they should have been able to demonstrate by the end of grade 4, according to the official curriculum.

Students in upper primary schools were assessed using Hindi and math exams developed by the National Council of Educational Research and Training, a national-level education organization providing assistance to state and central education authorities. These exams covered competencies that students were expected to master by the end of grade 8.

Baseline testing took place between November 2011 and March 2012, in the academic year before the interventions took place.[5] Endline testing was conducted in February and March 2013, at the end of the following academic year. All students who had been tested in the 500 sample schools at baseline were targeted for endline testing, and schools were visited multiple times to minimize attrition due to student absences.

We also incorporated an extensive program of process monitoring into our study design. The primary process monitoring data comes from two surprise visits to each of the 500 sample schools by monitors employed by the research team between August 2012 and March 2013. During these

[5] Baseline testing took place over two separate rounds, held in November 2011 and February to March 2012. See Appendix B for additional detail.

visits, monitors administered a questionnaire that collected data on student attendance in each grade, CCE implementation, and performance of the ABRCs. Monitors also observed a randomly selected teacher for thirty minutes to collect data on teaching and evaluation practices in the classroom. During baseline and endline testing, we also conducted surveys of school headmasters that collected information on school composition, evaluation practices, and (at endline) implementation and opinions about the CCE program.

4.2 Summary Statistics, Baseline Balance, and Attrition

Consistent with statewide and national surveys (ASER Centre 2012), students' learning levels in both Hindi and math in our sample were generally poor. Figure 1 presents tabulations of oral test scores in the lower primary sample by competency. Over 25 percent of lower primary school students in our sample were unable to identify isolated letters, and almost 84 percent of students were unable to read a simple story (grade 2 level text). More than 55 percent of lower primary school students tested were unable to recognize two-digit numbers.

Table 1 displays summary statistics for test score and demographic variables and checks of balance across treatment groups in the lower primary sample. Columns 2 through 4 present differences in means for each variable between the control group and the three randomized treatment groups: CCE, TaRL, and CCE combined with TaRL. Differences are computed by regressing the baseline value of the variable on three dummies for treatment status, controlling for stratum dummies. We do not observe significant differences in any of the variables examined. Columns 5 and 6 test for baseline differences across treatments by regressing the baseline value of the variable on indicators for receiving CCE or TaRL, with no separate indicator for the combined CCE and TaRL group. In this case, we do observe a small but statistically significant imbalance in oral test scores between receiving CCE and the control group: the CCE group had Hindi scores

that were 0.053 standard deviations lower (significant at the 5 percent level) and math scores that were 0.037 points lower (significant at the 10 percent level). In our analysis of test score impacts in Section 5.3 below, our preferred regression model controls for all baseline test scores. We also present several specifications for each outcome to examine whether the impacts are robust to the inclusion of controls. As we show, the estimates do not change appreciably across specifications.

Table 2 presents summary statistics and balance across treatment groups in the upper primary sample. As shown in the table, there is no evidence of imbalance by test scores. We do observe a small but significant difference in age between students in the CCE group and the control group (-0.17 years, significant at the 1 percent level).

In both lower and upper primary schools, attrition was low: at endline we were able to reach 95 percent of lower primary students and 92 percent of upper primary students tested at baseline. As shown in Tables 1 and 2, attrition does not significantly differ by treatment assignment: differences between treatment and control groups are less than 1 percentage point in all cases, and no difference is statistically significant.

5 Results

5.1 Program Implementation

Table 3 displays the percentage of teachers trained by treatment arm and school type. Overall, 89 percent of teachers in the CCE arms were trained. This fraction was higher in lower primary schools than in upper primary schools (93 percent vs. 75 percent). The table also shows very low contamination of the non-CCE treatment arm: only 1.7 percent of teachers in non-CCE lower primary schools and 0.6 percent of teachers in non-CCE upper primary schools reported being trained.

The most basic indicators of implementation of CCE in the study schools are the presence of CCE materials in schools and the use of the CCE evaluation sheets and report cards. Presence and use of these materials are summarized in Table 4. Eighty-two percent of teachers in CCE schools reported having their CCE manuals, and 44 percent were able to show their manual to the surveyor during the process evaluation visits. Likewise, 66 percent of teachers in CCE schools reported using evaluation sheets and 45 percent reported using report cards, while 35 percent and 24 percent were able to show the surveyor a completed evaluation sheet or a completed report card, respectively, for one of their students. Verified presence of these materials was higher in lower primary schools than in upper primary schools, both with evaluation sheets (39 percent vs. 21 percent) and report cards (25 percent vs. 17 percent).

5.2 Regression Specification

As described in Section 4.1, our main outcomes of interest are students' test scores. These scores are normalized using the mean and standard deviation of the control group's scores in each testing round (baseline or endline). We estimate the following regression using ordinary least squares:[6]

$$Y_{1is} = \beta_0 + \beta_1 T_{Cs} + \beta_2 T_{Ts} + \delta \boldsymbol{Y}_{0i} + \pi \mathbf{C}_{\text{is}} + \gamma \mathbf{S_s} + \varepsilon_{is}$$

In this regression, Y_{1is} represents the endline test score for student i in school s. T_{Cs} indicates assignment of school s to the CCE treatment, and T_{Ts} indicates assignment to the TaRL treatment (for lower primary schools). $\boldsymbol{Y}_{0i}$ represents a set of controls for oral and written baseline test scores, $\mathbf{C_{is}}$ represents a set of student characteristics (age, grade, and gender), and $\mathbf{S_s}$ are fixed

[6] Before the data were analyzed, a pre-analysis plan including the main regression specifications, as well as the weights used in aggregating test scores, were uploaded to the American Economic Association's Randomized-Controlled-Trial Registry, https://www.socialscienceregistry.org/trials/8.

effects for stratum. We include indicators for missing covariates and replace missing values with zero to avoid respondents' dropping out of our analysis due to nonresponse for particular variables. The error-term ε_{is} is clustered at the school campus level, the unit of randomization.

5.3 Test Scores

We present impact results on test scores separately for lower and upper primary students. The regression specified in our pre-analysis plan includes baseline test scores and additional demographic variables as controls, and we present this as our preferred specification. To examine robustness to the inclusion of control variables, we also present results with no controls and with only baseline test scores as controls.

Table 5 presents our main impact results for lower primary school students. As shown in the table, there is no evidence that the CCE program had significant impacts on oral or written test scores in either Hindi or math. These results are robust across specifications. The estimates are precise enough to rule out relatively small impacts: using our preferred specification in Columns 3 and 6, the upper bounds on the 95 percent confidence intervals range from 0.036 standard deviations on the oral math test to 0.069 standard deviations on the written Hindi test.

We find similar null effects on Hindi and math test scores of students in the upper primary sample, as shown in Table 6. Using our preferred specification in Column 3, the estimated impacts are 0.023 standard deviations for Hindi (s.e. = 0.045) and -0.041 standard deviations for math (s.e. = 0.056). Confidence intervals allow us to rule out effects of above 0.11 in Hindi and 0.07 in math.[7]

[7] We also find no consistent evidence for heterogeneity in treatment effects by initial test score. Results are displayed in Appendix Tables 1 and 2.

5.4 School Attendance

As described in Section 2, by providing continuous feedback to students rather than relying primarily on exams, one of the goals of the CCE program was to decrease stress of students and reduce dropout. In India, end-of-year exams were eliminated nationwide with the passage of the RTE in 2009 and were therefore not in place in Haryana during the time of our study. However, control schools in our study still relied mainly on periodic testing for evaluation, and we can evaluate the stated aim of CCE to reduce stress and dropout by decreasing this reliance on exams.

In practice, students in Indian schools often do not drop out but rather stop attending school regularly. Reduced stress could also lead to reduced absence for kids who remain enrolled. Table 7 presents impacts of CCE on school attendance. We use two measures for attendance: the head count as measured by enumerators during the process monitoring visits (measured at the school level) and the number of days missed in the two months before the endline exams according to school attendance records (measured at the student level). In neither case do we find evidence that CCE increased school attendance, in either the lower or upper primary samples. In fact, the impact on the number of days missed in the lower primary sample is *positive* and significant, representing an increase in days missed from 5.7 days in the control group to 6.4 in the CCE group. However, the impact on the student head count in the lower primary sample is a fairly precise null, indicating an increase of 1.3 students, with a standard error of 1.0, relative to the control group mean of 52 students.[8]

[8] We also do not find evidence of impacts of CCE on whether the student was present in school during the initial endline testing visit (results not shown).

6 Mechanisms

In this section we utilize the results on program implementation and process evaluation to discuss potential explanations for the lack of impacts of the CCE program found in the previous section.

6.1 Analysis of Process Data

As described in Section 5.1, basic implementation of CCE—as measured by teacher training and completion of CCE evaluation sheets—occurred in many schools in our sample. Still, implementation was imperfect: about one third of teachers could show the field officers a completed evaluation sheet, one third claimed to generally complete evaluation sheets but were not able to produce the sheet, and one third admitted not completing them. However, our impact estimates are precise enough to suggest that lack of implementation at this level does not drive the null impacts we observe. Appendix Tables 3 and 4 provide Treatment-on-the-Treated (ToT) estimates of the impacts of the CCE program. We use a conservative measure of take-up, coding a school as implementing the program only if completed CCE evaluation sheets and report cards were observed by the monitors during at least one process evaluation visit. By this measure 41 percent of primary schools and 21 percent of upper primary schools assigned to CCE took up the program. We then instrument this take-up measure with random assignment of the school to CCE. As shown in Appendix Table 3, in the lower primary school sample we are able to rule out ToT impacts of between 0.08 and 0.16 standard deviations, suggesting that the lack of effects is not only due to a low overall use of CCE materials. Because the upper primary sample was smaller and had lower take-up, confidence intervals on the ToT estimates for this group are considerably larger, as shown in Appendix Table 4.

In addition to leading teachers to fill out the evaluation sheets, the CCE program led to increases in the breadth of evaluation methods used by teachers. Table 8 displays impacts of CCE on whether a teacher reported using a particular method of evaluation. The CCE program encouraged a variety of evaluation methods, including periodic tests, assignments, projects, and in-class interactions. We find significant impacts of the program on the use of assignments, projects, and other activities for evaluation. There is also a positive but statistically insignificant impact on the use of in-class interactions. While the impact on periodic tests is not significant, this was a relatively common method of evaluation in all schools in the sample.

Although CCE did lead to changes in teachers' evaluation practices, there appears to have been little impact on the use of evaluation data or on teaching methods. Table 9 displays differences between the CCE and non-CCE schools in how teachers reported using evaluation data. There are few differences between CCE and non CCE schools. Of particular note is that the additional data gathered under CCE were not used to identify low-performing students, a key benefit attributed to CCE (Bhatia 2009). The only case where we observe a significant difference in use of evaluation data is the provision of feedback to parents (43 percent vs. 35 percent, significant at the 1 percent level). Although we do not have precise data on the type of feedback provided, this impact likely arose from teachers filling out and sending home CCE report cards, rather than having discussions with parents about their children's performance.

In Table 10 we present estimates of the impact of CCE on teaching practices, using data from classroom observations. We focus on several pedagogical techniques emphasized as part of teacher training for CCE, including using examples from everyday life, using local information to explain concepts, repeating concepts based on interactions with students, and changing explanations based on interactions with students. On the whole, teachers in CCE schools did not use the CCE-

recommended techniques any more than teachers in control schools. In the two cases where we observe significantly different use of a technique, CCE teachers in primary schools repeat concepts and simplify their explanations *less* frequently than teachers in control schools. The negative results, although only for two out of four variables, are cause for concern: it appears as though completing the CCE forms may have served as a substitute to whatever flexibility the teachers were willing to exercise before.

6.2 Assessment for Assessment's Sake

Based on the analysis of our process data, we speculate that the limited changes to teaching practices as the result of CCE—and the subsequent lack of impacts on test scores—were likely due to its focus on filling out lengthy evaluations rather than on using evaluation data to improve basic learning outcomes. As described in Section 2, the CCE program required teachers to complete frequent evaluations in a large number of skills and sub-skills. Returning to the sample evaluation sheet shown in the Supplemental Appendix, students in grades 1 and 2 were to be evaluated monthly in English, Hindi, math, co-curricular performance, and personal and social qualities. Under each of these topics, there were three to four skills, and within each skill there were several sub-skills along which students were expected to be evaluated. In all, there were 20 skills and 41 sub-skills along which a student received numeric scores each month. Teachers could also leave descriptive remarks related to each of the 20 skills.

Additional process data from headmasters and teachers corroborate the argument that program was viewed as an administrative burden. When asked whether CCE-related paperwork affected time spent on teaching, 35 percent of teachers trained in CCE indicated that it adversely affected teaching, while only 9 percent indicated that it positively affected teaching. When school headmasters were asked whether they had issues or problems with the CCE program, the two most

commonly cited issues were feeling overburdened by the additional requirements imposed by CCE and feeling that the program requirements were too time-consuming.

In effect, CCE turned into another paperwork obligation for teachers, which would likely have consumed a large amount of their time if they conducted the evaluations carefully. At the same time, there limited guidance or encouragement for them to adjust the pace of the curriculum or modify teaching practices. If anything, for diligent teachers, it may have reduced the time they had to adjust teaching to the pace of learning of their students (which may be the reason why we observe some negative impacts on activities that take extra time, such as repeating concepts). For less diligent teachers, it probably just turned into another set of forms to fill out.

6.3 *Can* Teaching Practices Be Changed?

In light of the lack of impacts of CCE, it is natural to ask whether existing government school teachers *can* modify teaching practices in a manner that improves learning levels of students. As we show in Table 5 and discuss in detail in Banerjee et al. (2017), the TaRL program—a program that was implemented by the same population of government teachers and provided a similar amount of training—was successful in improving learning in the lower primary schools in our sample. In Banerjee et al. (2017), we argue that a key to the success of this program was shifting the focus of teaching from covering the syllabus to improving basic Hindi and math skills. Unlike the CCE program, the TaRL program was centered on the *outcome* that was sought (imparting

basic skills to children) rather than a step in the process (conducting evaluations and filling out evaluation sheets).[9]

6.4 External Validity within India

As noted in Section 2.2, the overall guidelines of CCE were set at the national level, but implementation was left to the individual states. Our evaluation covers the rollout of CCE within the state of Haryana only. However, the broad directives at the national level to implement CCE have led many states to similarly design cumbersome evaluation programs with little or no focus on learning outcomes. Documenting the implementation experience across the country, a 2014 report by the National Council of Educational Research and Training concluded that that CCE assessments were "highly rigid and cumbersome for both teachers and children….The over emphasis on quantification of the achievement through marks/percentage/grades does not rule out the labeling of children as claimed by almost all states. In addition it limits the crucial role and contribution of the qualitative component of assessment towards improving and enhancing children's learning" (Sharma 2014). A 2017 analysis of India's CCE program argued, "The amount of work related to the recording and reporting of students' progress is stupendous. The requirement of following certain formats has come to dominate schooling and teachers fulfil the needs of CBSE's [Central Board of Secondary Education's] CCE regime in a mechanical manner, without much reflection and analysis (Srinivasan, 2015)" (Yagnamurthy 2017).

[9] Indeed, the linkage from assessment to pedagogy has been cited as a critical component of formative assessment systems, upon which the CCE model was based. In summarizing the experience with formative assessment in the United States, Bennett (2011) recommends that "teachers will need useful classroom materials that model the integration of pedagogical, domain, and measurement knowledge (e.g., developmentally sequenced tasks that can help them make inferences about what students know with respect to key domain competencies, and about what next to target for instruction)."

We note, however, that our evaluation included an additional monitoring component that that was not included in the national CCE guidelines. As described in Section 2.4, the ABRC monitoring program was designed in partnership by the researchers and the Government of Haryana in order to strengthen implementation of CCE. Process monitoring data indicate that ABRCs did indeed assist in implementation: 79 percent of CCE schools reported a visit from an ABRC in the 30 days prior to the process monitoring visit. Sixty-two percent of teachers in CCE schools reported asking ABRCs questions about CCE implementation, with 92 percent of those who asked questions reporting that the ABRC's answer was helpful. Thus, because this type of monitoring is not a general design feature of the CCE program, other states in India may face challenges in basic implementation of the program beyond what was experienced in Haryana.

7 Conclusion

Across a number of developing countries, systems of continuous evaluation have been proposed and implemented as a means to improve learning outcomes by providing teachers, students, and parents with more feedback on students' progress. This paper presents the results of a randomized evaluation of India's CCE program, the country's current flagship reform in primary schools that has been implemented across the country. Using a randomized evaluation in the state of Haryana, we estimate the program's impacts on language and math test scores in 500 primary schools. We find that the program failed to improve learning outcomes: there are no statistically significant impacts on test scores, and the estimates are precise enough to allow us to rule out relatively small effect sizes.

Using our process data, we show that while many teachers in the CCE program did complete the CCE evaluation sheets and broadened the methods they used to evaluate students, this did not result in changes in teaching practices. We argue that this likely resulted from CCE's failure to

provide a clear link between diagnosis and further action. In this case, the “comprehensive” aspect of CCE may actually inhibit such changes by diverting the focus from basic learning outcomes.

Although this study was conducted based on one state’s implementation model, we argue that the focus of CCE on evaluation without linking it to teaching practices makes it likely to be ineffective in improving learning levels in other states, and evidence on the implementation experience in other states supports this claim. Along the same lines, studies have documented similar implementation challenges with continuous assessment programs in other countries, particularly the time involved in completing assessments and a lack of clear understanding by teachers of how to use them to improve learning outcomes (Gule 1999; UNESCO 2008; Kapambwe 2010).

While our study shows that such evaluation systems may not be effective, the results of the TaRL evaluation show that teachers can change teaching practices and improve learning when the focus is on teaching basic skills at the current level of the students (Banerjee et al, 2017). The basic premise of continuous assessment—to provide frequent feedback to teachers—may indeed be effective if it is linked to targeted instruction and if such instruction is made a core responsibility of teachers. This has the potential to improve learning outcomes for millions of students, without substantial changes in school resources.

Bibliography

ASER Centre. 2012. “Annual Status of Education Report (Rural) 2011.” New Delhi: Pratham.

———. 2017. “Annual Status of Education Report (Rural) 2016.” New Delhi: Pratham.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. “Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of ‘Teaching at the Right Level’ in India.” *NBER Working Paper 22746*.

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

Banerjee, Abhijit, and Esther Duflo. 2011. *Poor Economics*. New York: PublicAffairs.

Bennett, Randy Elliot. 2011. “Formative Assessment: A Critical Review.” *Assessment in Education: Principles, Policy and Practice* 18 (1): 5–25.

Bhatia, Sangeeta. 2009. “CCE - Paradigm Shift from 'Teaching to the Test’ to 'Holistic Education’.” *Quarterly Bulletin of the Central Board of Secondary Education* 48 (4): 24–28.

Black, Paul, and Dylan Wiliam. 2009. “Developing the Theory of Formative Assessment.” *Educational Assessment, Evaluation and Accountability (Formerly: Journal of Personnel Evaluation in Education)* 21 (1): 5.

Government of India. 2009. “The Right of Children to Free and Compulsory Education Act 2009.” *Gazette of India* 39 (August).

Gule, Elvis D. 1999. “Problems Experienced by Classroom Primary Teachers, Headteachers and Pupils in Implementing the National Continuous Assessment Programme in Schools in the Manzini Region, Swaziland.” University of the Witwatersrand.

Hoyos Navarro, Rafael de, Vincent A. Garcia-Moreno, and Harry Anthony Patrinos. 2017. “The Impact of and Accountability Intervention with Diagnostic Feedback: Evidence from Mexico.” *Economics of Education Review* 58: 123–40.

Hoyos, Rafael de, Alejandro J. Ganimian, and Peter Holland. 2017. “Teaching with the Test: Experimental Evidence on Diagnostic Feedback and Capacity Building for Public Schools in Argentina.” *Mimeo, New York University*.

Kamangira, Y. T. 2003. “Feasibility of a Large Scale Implementation of Continuous Assessment as a Stimulus for Teacher Development in Malawi.” *An Improvement of Education Quality (IEQ) Project Report in Collaboration with American Institute for Research in USA*.

Kapambwe, William M. 2010. “The Implementation of School Based Continuous Assessment (CA) in Zambia.” *Educational Research and Reviews* 5 (3): 99–107.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2010. “The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India.” *Economic Journal* 120 (August): F187–203.

National University of Educational Planning and Administration (India). 2013. “Elementary Education in India: Where Do We Stand? State Report Cards, 2013-2014.” New Delhi: NUEPA.

Reserve Bank of India. 2011. “Handbook of Statistics on Indian Economy 2010-11.” New Delhi: Reserve Bank of India.

Sharma, Kavita. 2014. “CCE Programme/Scheme of States and UTs.” New Delhi: Department of Elementary Education, National Council of Educational Research and Training.

UNESCO. 2008. “EFA Global Monitoring Report.” Paris: United Nations Educational, Scientific and Cultural Organization.

Yagnamurthy, Sreekanth. 2017. “Continuous and Comprehensive Evaluation (CCE): Policy and Practice at the National Level.” *The Curriculum Journal* 28 (3): 421–41.

Figure 1. Baseline Oral Test Levels, Lower Primary Schools

Figure 1A. Hindi

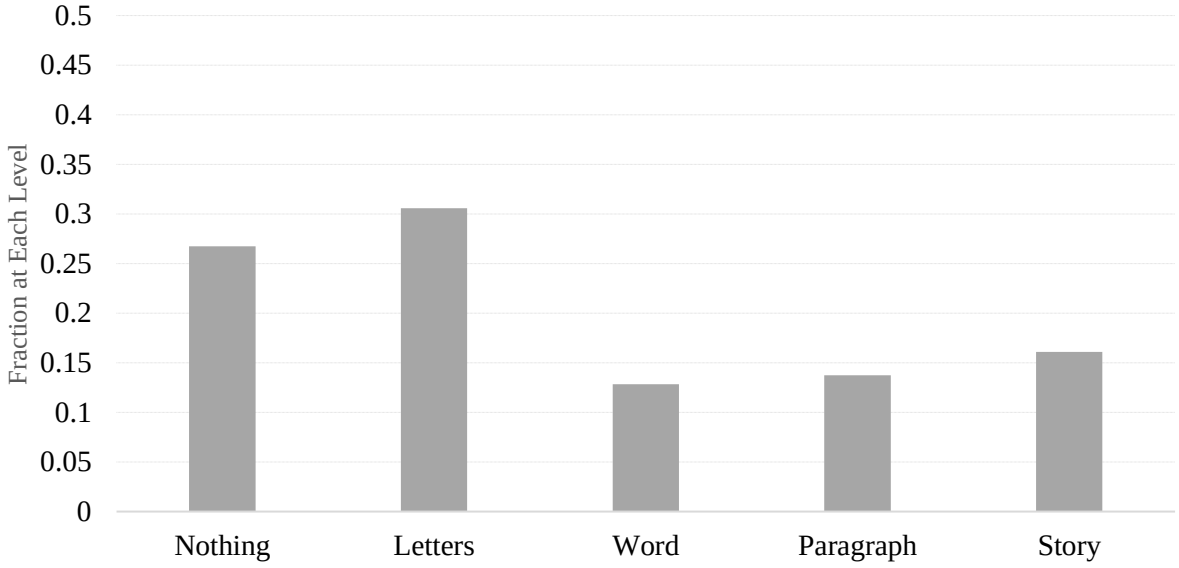


Figure 1B. Math

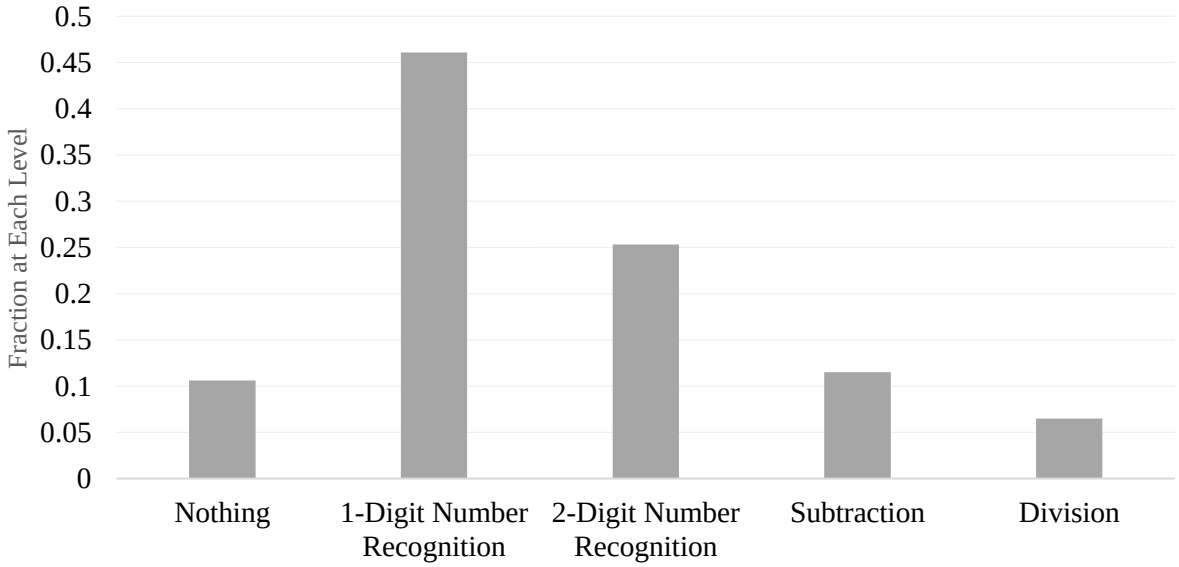


Table 1: Descriptive Statistics and Balance Check of Randomization, Lower Primary Schools

		Relative to Control					
		Individual Treatment Groups			Without Interaction		
	Control Mean	CCE Only	TaRL Only	CCE & TaRL	Any CCE	Any TaRL	Obs
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Demographic characteristics							
Female	0.506	0.00656	0.00184	0.0154	0.0102	0.00540	12576
	[0.500]	(0.0172)	(0.0151)	(0.0149)	(0.0106)	(0.0114)	
Age (years)	9.04	-0.00134	0.0491	0.0147	-0.0183	0.0323	12555
	[1.568]	(0.0518)	(0.0524)	(0.0494)	(0.0372)	(0.0371)	
Grade in 2011-12 school year	2.57	-0.0280	-0.0100	-0.0195	-0.0185	-0.000588	12576
	[1.114]	(0.0174)	(0.0176)	(0.0188)	(0.0118)	(0.0118)	
B. Normalized baseline test scores							
Oral Hindi	0.000	-0.0368	0.0329	-0.0364	-0.0534**	0.0164	12472
	[1.000]	(0.0354)	(0.0331)	(0.0360)	(0.0236)	(0.0241)	
Oral math	0.000	-0.0483	-0.0184	-0.0450	-0.0372*	-0.00736	12393
	[1.000]	(0.0297)	(0.0284)	(0.0308)	(0.0205)	(0.0206)	
Written Hindi	0.000	-0.0430	-0.0438	-0.0251	-0.0115	-0.0125	6208
	[1.000]	(0.0505)	(0.0464)	(0.0496)	(0.0342)	(0.0339)	
Written math	0.000	-0.00410	-0.0660	-0.0597	0.00120	-0.0607*	6204
	[1.000]	(0.0473)	(0.0433)	(0.0448)	(0.0314)	(0.0314)	
C. Attrition							
Not reached at endline	0.0527	-0.00603	-0.00859	0.00656	0.00484	0.00218	12576
	[0.223]	(0.00714)	(0.00744)	(0.00779)	(0.00490)	(0.00534)	

Notes: Standard deviations in square brackets, standard errors in parentheses, clustered at the school campus level. Column 1 displays the mean of the variable in the control group. Columns 2, 3, and 4 display the differences between the CCE only, TaRL only, and combined CCE and TaRL treatment groups, respectively, and the control group. Column 5 displays the difference between the CCE treatment groups and the control group, controlling for TaRL treatment status. Column 6 displays the difference between the TaRL treatment groups and the control group, controlling for CCE treatment status. Differences are computed by regression, controlling for stratum. Written Hindi and math tests were conducted for students in grades 3 and 4, leading to fewer observations. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table 2: Descriptive Statistics and Balance Check of Randomization, Upper Primary Schools

	Control Mean	Relative to Control CCE	Obs
	(1)	(2)	(3)
A. Demographic characteristics			
Female	0.496	0.0172	3261
	[0.500]	(0.0428)	
Age (years)	13.9	-0.174***	3255
	[1.182]	(0.0516)	
B. Normalized baseline test scores			
Hindi	0.000	-0.0724	2610
	[1.000]	(0.0513)	
Math	0.000	0.000205	2602
	[1.000]	(0.0633)	
C. Attrition			
Not reached at endline	0.0791	0.00457	3261
	[0.270]	(0.0115)	

Notes: Standard deviations in square brackets, standard errors in parentheses, clustered at the school campus level. Column 1 displays the mean of the variable in the control group. Column 2 displays the difference between the CCE treatment and the control group. Differences are computed by regression, controlling for stratum. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table 3. Teacher Training

	All Schools		Lower Primary Schools		Upper Primary Schools	
	Non-CCE	CCE	Non-CCE	CCE	Non-CCE	CCE
	(1)	(2)	(3)	(4)	(5)	(6)
Trained	0.0146	0.892	0.0167	0.926	0.00641	0.747
Not Trained	0.985	0.108	0.983	0.0736	0.994	0.253
Number of Schools	248	245	196	198	52	47

Notes: This table presents the fraction of teachers who attended CCE training, by school type and treatment arm, measured during process evaluation visits. The fraction in upper primary schools includes only teachers of grade 8. The unit of observation is the school. Because most schools have two measures (corresponding to the two process evaluation visits), averages are taken over visits within each school to yield school-level measures.

Table 4. Presence and Completion of CCE Evaluation Materials in CCE Schools

	All CCE Schools	Lower Primary CCE Schools	Upper Primary CCE Schools
	(1)	(2)	(3)
A. Presence of CCE Manual			
Teacher has CCE manual (shown to monitor)	0.442	0.517	0.128
Teacher claims to have CCE manual (not shown to monitor)	0.381	0.39	0.346
Teacher does not have CCE manual	0.179	0.0968	0.527
B. Completion of CCE Evaluation Sheets			
Completes evaluation sheets (shown to monitor)	0.356	0.391	0.213
Claims to completes evaluation sheets (not shown to monitor)	0.305	0.276	0.426
Does not complete evaluation sheets	0.329	0.327	0.34
C. Completion of CCE Report Cards			
Completes report cards (shown to monitor)	0.235	0.25	0.17
Claims to complete report cards (not shown to monitor)	0.218	0.191	0.33
Does not complete report cards	0.538	0.552	0.479
Number of Schools	245	198	47

Notes: This table presents the proportion of schools in which CCE evaluation materials were reported or observed to be present. Multiple observations within a school are averaged to yield school-level measures. In Panel A, the measures were taken for all teachers over two process evaluation visits. In Panels B and C, the measures were taken for one randomly selected teacher in each of the two visits.

Table 5: Test Results, Lower Primary Schools

	Oral			Written		
	(1)	(2)	(3)	(4)	(5)	(6)
A. Hindi						
CCE	-0.0366	0.00370	0.00230	0.000575	0.0294	0.0283
	(0.0240)	(0.0172)	(0.0173)	(0.0299)	(0.0204)	(0.0207)
TaRL	0.159***	0.151***	0.152***	0.128***	0.135***	0.135***
	(0.0238)	(0.0168)	(0.0167)	(0.0294)	(0.0206)	(0.0208)
Observations	11963	11963	11963	9204	9204	9204
R-squared	0.0749	0.633	0.637	0.0838	0.646	0.651
B. Math						
CCE	-0.0287	0.00465	0.00557	-0.0103	0.0139	0.0138
	(0.0207)	(0.0154)	(0.0153)	(0.0288)	(0.0213)	(0.0213)
TaRL	-0.0128	-0.00605	-0.00581	0.00302	0.0224	0.0232
	(0.0214)	(0.0155)	(0.0154)	(0.0297)	(0.0222)	(0.0222)
Observations	11950	11950	11950	9204	9204	9204
R-squared	0.0639	0.649	0.652	0.102	0.666	0.666
Baseline Scores?	NO	YES	YES	NO	YES	YES
Other Controls?	NO	NO	YES	NO	NO	YES

Notes: This table presents impact estimates of the CCE and TaRL programs on normalized test scores. Scores are normalized using the mean and standard deviation in the control group in each round of testing. All regressions control for stratum dummies. "Baseline Scores" are the all baseline test scores listed in Table 1, and "Other Controls" are the other demographic variables listed in Table 1. Missing values of control variables are coded as 0, with additional dummy variables to indicate missing values. The TaRL program covered only Hindi skills. Standard errors in parentheses, clustered at the school campus level. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table 6: Test Results, Upper Primary Schools

	(1)	(2)	(3)
A. Hindi			
CCE	-0.00304	0.0460	0.0222
	(0.0506)	(0.0459)	(0.0447)
Observations	2999	2999	2999
R-squared	0.0911	0.402	0.420
B. Math			
CCE	-0.0556	-0.0301	-0.0404
	(0.0577)	(0.0554)	(0.0556)
Observations	3000	3000	3000
R-squared	0.0637	0.152	0.155
Baseline Scores?	NO	YES	YES
Other Controls?	NO	NO	YES

Notes: This table presents impact estimates of CCE program on normalized test scores. Scores are normalized using the mean and standard deviation in the control group in each round of testing. All regressions control for stratum dummies. "Baseline Scores" are the all baseline test scores listed in Table 2, and "Other Controls" are the other demographic variables listed in Table 2. Missing values of control variables are coded as 0, with additional dummy variables to indicate missing values. Standard errors in parentheses, clustered at the school campus level. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table 7: Attendance Results

	Dependent Variable	
	Head Count	Days Missed
	(1)	(2)
A. Lower Primary		
CCE	1.306	0.712***
	(1.009)	(0.275)
TaRL	-0.316	-0.0851
	(1.081)	(0.270)
Control Group Mean	52.36	5.731
Number of Observations	394	7888
R-squared	0.953	0.115
B. Upper Primary		
CCE	-2.522	1.139
	(2.301)	(0.818)
Control Group Mean	26.32	6.398
Number of Observations	100	2326
R-squared	0.769	0.0920

Notes: This table presents impact estimates of CCE program on attendance. Column 1 uses a measure of the number of students present during process monitoring visits. Panel A is the total head count for students in grades 2 to 5. Panel B is the head count for students in grade 8. Averages are taken by school over the two process evaluation visits. Column 2 uses a student-level measure of days missed in the two months prior to the exam, for all students who took the endline test. Regressions control for stratum dummies and total number of students on the roster (Column 1) or days missed at baseline (Column 2). Missing values of control variables are coded as 0, with additional dummy variables to indicate missing values. Standard errors in parentheses, clustered at the school campus level. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table 8: Teachers' Evaluation Practices

	All Schools			Lower Primary Schools			Upper Primary Schools		
	Non-CCE	CCE	Difference	Non-CCE	CCE	Difference	Non-CCE	CCE	Difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Periodic/unit tests	0.764	0.747	-0.0173	0.721	0.717	-0.00558	0.925	0.876	-0.0485
			(0.0245)			(0.0313)			(0.0382)
Workbook / homework assignments	0.390	0.456	0.0674**	0.387	0.465	0.0826**	0.399	0.418	-0.00137
			(0.0328)			(0.0378)			(0.0631)
Projects	0.0241	0.0497	0.0235**	0.0102	0.0311	0.0195**	0.0755	0.128	0.0444
			(0.0116)			(0.00972)			(0.0440)
In-class interactions	0.617	0.663	0.0438	0.609	0.661	0.0450	0.648	0.674	0.0396
			(0.0311)			(0.0371)			(0.0686)
Other activities / games / attendance	0.0388	0.0735	0.0332**	0.0298	0.0657	0.0342**	0.0723	0.106	0.0239
			(0.0147)			(0.0163)			(0.0373)
TaRL assessments / activities	0.132	0.117	-0.0186	0.168	0.145	-0.0238	0.000	0.000	0.000
			(0.0182)			(0.0302)			---

Notes: This table presents responses to the open-ended question "How do you evaluate students?" asked during process evaluation visits to one randomly-selected teacher per school. Averages are taken by school over the two process evaluation visits. Differences are computed by regression, controlling for TaRL treatment and stratum dummies. Standard errors in parentheses, clustered at the school campus level. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table 9: Teachers' Use of Evaluation Data

	All Schools			Lower Primary Schools			Upper Primary Schools		
	Non-CCE	CCE	Difference	Non-CCE	CCE	Difference	Non-CCE	CCE	Difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Provides feedback to students	0.434	0.465	0.0221	0.417	0.462	0.0361	0.500	0.479	-0.0146
			(0.0338)			(0.0393)			(0.0773)
Provides feedback to parents	0.345	0.427	0.0808***	0.352	0.434	0.0797**	0.318	0.394	0.0781
			(0.0284)			(0.0344)			(0.0688)
Identifies low-performing students	0.169	0.163	0.00219	0.177	0.158	-0.00791	0.138	0.181	0.0396
			(0.0226)			(0.0264)			(0.0494)
Changes teaching practices if most students are not performing well	0.253	0.205	-0.0495	0.276	0.206	-0.0726**	0.167	0.202	0.0246
			(0.0307)			(0.0356)			(0.0637)
Reports information to headmaster	0.0890	0.0660	-0.0219	0.0706	0.0539	-0.0175	0.157	0.117	-0.0469
			(0.0166)			(0.0171)			(0.0558)
Does not use evaluation data	0.0475	0.0490	0.00180	0.0502	0.0530	0.00464	0.0377	0.0319	-0.0102
			(0.0141)			(0.0160)			(0.0300)

Notes: This table presents responses to the open-ended question "How do you use the data on student evaluations?" asked during process evaluation visits to one randomly-selected teacher per school. Averages are taken by school over two process evaluation visits. Differences are computed by regression, controlling for TaRL treatment and stratum dummies. Standard errors in parentheses, clustered at the school campus level. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table 10: Use of CCE-encouraged teaching practices

	All Schools			Lower Primary Schools			Upper Primary Schools		
	Non-CCE	CCE	Difference	Non-CCE	CCE	Difference	Non-CCE	CCE	Difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Teacher use examples from everyday life	0.423	0.414	-0.00974	0.402	0.424	0.0127	0.500	0.372	-0.131
			(0.0328)			(0.0362)			(0.0860)
Teacher uses local information to explain concepts	0.156	0.156	0.000	0.126	0.141	0.00355	0.264	0.220	-0.0394
			(0.0224)			(0.0247)			(0.0704)
Teacher repeats concept based on student answers	0.0489	0.0271	-0.0201*	0.0451	0.0208	-0.0233*	0.0629	0.0532	-0.0150
			(0.0121)			(0.0126)			(0.0319)
Teacher simplifies explanation based on student answers	0.363	0.303	-0.0628*	0.351	0.301	-0.0549	0.409	0.312	-0.0945
			(0.0322)			(0.0366)			(0.0838)

Notes: This table presents enumerator observations of selected teaching practices during process evaluation visits to one randomly-selected class in each school. Averages were taken by school over the two process evaluation visits. Differences computed by regression, controlling for TaRL treatment and stratum dummies. Standard errors in parentheses, clustered at the school campus level. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Apendix A. Complementarities Between CCE and TaRL

In this appendix we present estimates of complementarity between CCE and TaRL in the lower primary schools in our sample. We initially hypothesized that the CCE and TaRL programs could be complementary if the continuous evaluation under CCE helped teachers better identify students' learning levels, while the TaRL program provided a specific means to target teaching to students. While the TaRL program did incorporate a simple assessment to form and modify learning groups, it is possible that CCE could provide additional information for teachers, particularly if they were not using TaRL assessments frequently or effectively.

We evaluate complementarities by augmenting our main estimating equation with an interaction term of CCE and TaRL:

$$Y_{1is} = \beta_0 + \beta_1 T_{Cs} + \beta_2 T_{Ts} + \beta_3 T_{Cs} * T_{Ts} + \delta \boldsymbol{Y}_{0i} + \pi \mathbf{C}_{\text{is}} + \gamma \mathbf{S_s} + \varepsilon_{is}$$

In this equation, the coefficient on the interaction term $T_{Cs} * T_{Ts}$ indicates an additional impact of both CCE and TaRL, over and above the impacts of the two programs individually.

Appendix Table 5 displays the results. None of the interaction terms between the CCE and TaRL dummies is statistically significant, indicating no evidence for complementarities between the two programs. The point estimates range from 0.011 standard deviations in oral Hindi to 0.056 standard deviations in written math. As with the main impact results of CCE, the estimates are relatively precise: confidence intervals allow us to rule out effects as large as 0.078 standard deviations in oral Hindi to 0.14 standard deviations in written math.

These null results admit several interpretations. Teachers may not have obtained the information they needed for TaRL from CCE evaluations, or they may have been unable to effectively link the CCE evaluation data to activities in the TaRL program. However, process monitoring data suggest that the lack of effects was likely a result of teachers' effective use of the

TaRL assessments, obviating the need for additional information from CCE. The vast majority (89 percent) of TaRL schools were testing students using the Pratham instruments, implying that these instruments were the primary sources of information used to form and modify the TaRL tracking groups. Use of these assessments was nearly identical between the TaRL schools with CCE (89 percent) and without CCE (90 percent), suggesting that CCE assessments were not used in place of TaRL assessment in schools with both CCE and TaRL.

Apendix B. Additional Discussion of Study Design and Implementation

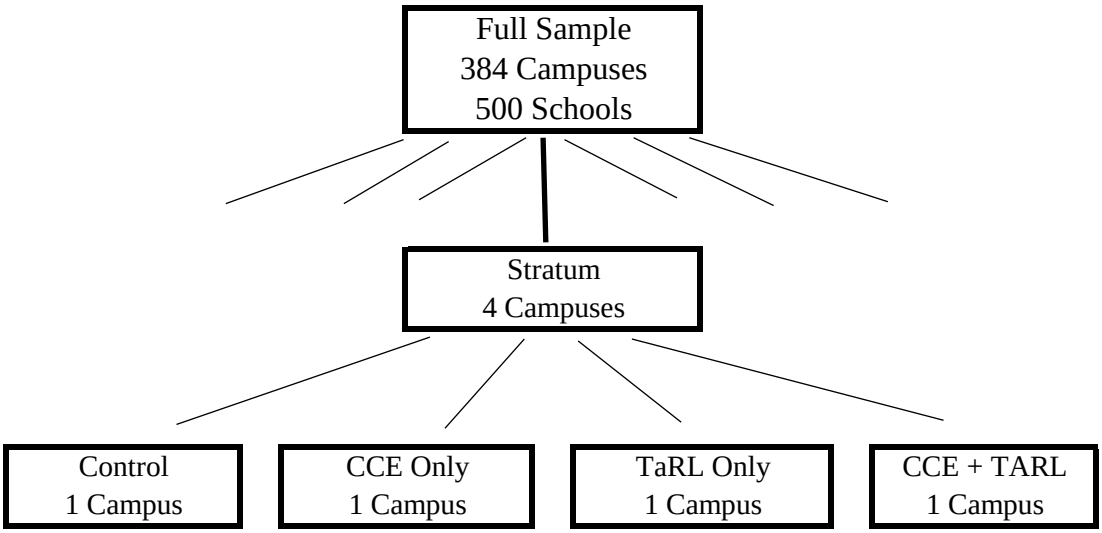
The sample for the study was constructed in two stages. In November 2011, 200 lower primary schools and 100 upper primary schools were selected for the sample. These schools were then randomly assigned to one of the four treatment arms of our study (lower primary) or one of two treatment arms (upper primary). However, due to considerations of statistical power, we decided to increase the number of lower primary schools in the sample to 400. The additional lower primary schools were selected and randomly assigned in February and March 2012.

Baseline testing in lower primary schools followed the two phases of sample selection. In November 2011, we conducted baseline testing of all children in grade 7 in the 100 upper primary schools as well as oral testing of up to seven randomly selected students in each of grades 1 to 4 in the 200 lower primary schools in the original sample. In February and March 2012, up to ten randomly selected students from each of grades 1 to 4 in the 200 “new” schools were tested using the oral tests. In the 200 “old” schools, up to three additional students in each grade were given the oral tests to bring the maximum number of children tested in each grade to 10. In all lower primary schools, selected students in grades 3 and 4 were also administered the written tests at that time.

The first training sessions for CCE took place in November 2011 for teachers in the 150 schools initially selected for inclusion in the study that had been assigned to receive CCE. However, actual implementation by the teachers did not begin until April 2012, just after the training for teachers in the 100 "new" schools that had been assigned to CCE.[1]

[1] There were a variety of reasons cited by teachers for the delay of implementation in the first set of schools, among them a lack of CCE materials for students (such as report cards and evaluation sheets), uncertainty regarding government guidelines, and unwillingness to introduce a major change in evaluation in the middle of an ongoing academic year.

Appendix Figure 1: Treatment Assignment



Notes: To create strata, campuses were first matched based on block and whether the campus contained lower primary grades, upper primary grades, or both. Within these matches, groups of four campuses were created based on average baseline test scores. Upper primary schools in campuses assigned to TaRL Only or CCE + TaRL were re-assigned to Control and CCE Only, respectively.

Appendix Table 1: Heterogeneity by Baseline Test Score, Lower Primary Schools

	Hindi		Math	
	Oral Hindi	Written Hindi	Oral Math	Written Math
	(1)	(2)	(3)	(4)
CCE	0.00365	0.0235	0.00931	0.00315
	(0.0173)	(0.0207)	(0.0153)	(0.0220)
Baseline test score	0.622***	0.383***	0.434***	0.351***
	(0.0172)	(0.0194)	(0.0189)	(0.0200)
CCE*Test score	0.0123	-0.0317*	0.0183	-0.0221
	(0.0146)	(0.0177)	(0.0153)	(0.0186)
Observations	11876	5991	11789	5988
R-squared	0.639	0.696	0.656	0.695

Notes: This table presents regressions of the normalized test score indicated on the CCE treatment group, the baseline score of that test, and the interaction of the treatment group and test score. Regressions control for all variables listed in Table 1, assignment to the TaRL treatment, the interaction between TaRL and test score, and stratum dummies. Missing values of control variables are coded as 0, with additional dummy variables to indicate missing values. Standard errors in parentheses, clustered at the school campus level. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Appendix Table 2: Heterogeneity by Baseline Test Score, Upper Primary Schools

	Hindi	Math
	(1)	(2)
CCE	0.0221	-0.0248
	(0.0442)	(0.0536)
Baseline test score	0.630***	0.127***
	(0.0404)	(0.0319)
CCE*Test score	-0.00395	-0.0590
	(0.0471)	(0.0457)
Observations	2480	2475
R-squared	0.469	0.165

Notes: This table presents regressions of the normalized test score indicated on the CCE treatment group, the baseline score of that test, and the interaction of the treatment group and test score. Regressions control for all variables listed in Table 2 and stratum dummies. Missing values of control variables are coded as 0, with additional dummy variables to indicate missing values. Standard errors in parentheses, clustered at the school campus level. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Appendix Table 3: Lower Primary Test Score Results
Treatment-on-the-Treated Estimates

	Hindi		Math	
	Oral Hindi	Written Hindi	Oral Math	Written Math
	(1)	(2)	(3)	(4)
CCE Implemented	0.00423	0.0662	0.00865	0.0301
	(0.0402)	(0.0488)	(0.0357)	(0.0503)
TaRL	0.154***	0.137***	-0.00418	0.0248
	(0.0167)	(0.0207)	(0.0154)	(0.0221)
Observations	11909	9156	11896	9156
R-squared	0.637	0.651	0.653	0.666

Notes: This table presents instrumental-variables estimates of CCE implementation on normalized test scores. "CCE Implemented" is an indicator for whether a randomly-selected teacher in the CCE treatment showed the enumerator the a completed evaluation sheet and a completed report card on at least one visit. Forty-one percent of schools implemented CCE by this definition. This indicator is instrumented with CCE treatment status. Regressions control for all variables listed in Table 1 and stratum dummies. Missing values of control variables are coded as 0, with additional dummy variables to indicate missing values. Standard errors in parentheses, clustered at the school campus level. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Appendix Table 4: Upper Primary Test Score Results
Treatment-on-the-Treated Estimates

	Hindi	Math
	(1)	(2)
CCE Implemented	0.166	-0.271
	(0.321)	(0.398)
Observations	2986	2987
R-squared	0.417	0.151

Notes: This table presents instrumental-variables estimates of CCE implementation on normalized test scores. "CCE Implemented" is an indicator for whether a randomly-selected teacher in the CCE treatment showed the enumerator the a completed evaluation sheet and a completed report card on at least one visit. Twenty-one percent of CCE schools implemented CCE by this definition. This indicator is instrumented with CCE treatment status. Regressions control for all variables listed in Table 2 and stratum dummies. Missing values of control variables are coded as 0, with additional dummy variables to indicate missing values. Standard errors in parentheses, clustered at the school campus level. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Appendix Table 5: Complementarity Between CCE and TaRL

	Hindi		Math	
	Oral	Written	Oral	Written
	(1)	(2)	(3)	(4)
CCE	-0.00358	0.00237	-0.00759	-0.0152
	(0.0234)	(0.0274)	(0.0233)	(0.0327)
TaRL	0.146***	0.109***	-0.0189	-0.00540
	(0.0223)	(0.0285)	(0.0221)	(0.0295)
CCE * TaRL	0.0115	0.0504	0.0256	0.0564
	(0.0337)	(0.0415)	(0.0307)	(0.0444)
Observations	11963	9204	11950	9204
R-squared	0.637	0.651	0.653	0.666

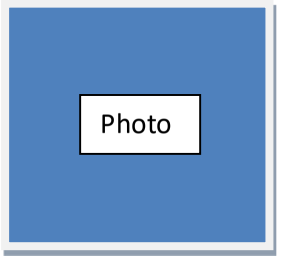
Notes: This table presents estimates of complementarities between the CCE and TaRL programs in lower primary schools. Regressions follow the specification used in Columns 3 and 6 of Table 5, including an additional indicator for the combined CCE and TaRL treatment group. Regressions include stratum dummies and all variables listed in Table 1. Missing values of control variables are coded as 0, with additional dummy variables to indicate missing values. Standard errors in parentheses, clustered at the school campus level. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Name: _____

Class: _____

EVALUATION SHEET

(Classes 1st to 2nd)



ENGLISH

1. Listening Skill:

S. No.	Sub Skill	April				May				July				August				Sept.				October				Nov.				Dec.				Jan.				Feb.				March			
		4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1
1.	Comprehension																																											
2.	Interest																																											
Descriptive Remarks																																											

2. Verbal Skill:

S. No.	Sub Skill	April			May			July			August			Sept.			October			Nov.			Dec.			Jan.			Feb.			March		
		3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
1.	Conversation																																	
2.	Recitation																																	
Descriptive Remarks																																		

3. Reading Skill:

S. No.	Sub Skill	April			May			July			August			Sept.			October			Nov.			Dec.			Jan.			Feb.			March		
		3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
1.	Fluency																																	
2.	Pronunciation																																	
Descriptive Remarks																																		

4. Writing Skill:

S. No.	Sub Skill	April			May			July			August			Sept.			October			Nov.			Dec.			Jan.			Feb.			March		
		3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
1.	Vocabulary																																	
2.	Dictation																																	
3.	Spelling																																	
4.	Handwriting																																	
Descriptive Remarks																																		

HINDI

1. Listening Skill:

S. No.	Sub Skill	April				May				July				August				Sept.				October				Nov.				Dec.				Jan.				Feb.				March			
		4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1
1.	Comprehension																																											
2.	Interest																																											
Descriptive Remarks																																												

2. Verbal Skill:

S. No.	Sub Skill	April			May			July			August			Sept.			October			Nov.			Dec.			Jan.			Feb.			March		
		3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
1.	Conversation																																	
2.	Recitation																																	
Descriptive Remarks																																		

3. Reading Skill:

S. No.	Sub Skill	April			May			July			August			Sept.			October			Nov.			Dec.			Jan.			Feb.			March		
		3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
1.	Fluency																																	
2.	Pronunciation																																	
Descriptive Remarks																																		

4. Writing Skill:

S. No.	Sub Skill	April			May			July			August			Sept.			October			Nov.			Dec.			Jan.			Feb.			March		
		3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
1.	Vocabulary																																	
2.	Dictation																																	
3.	Spelling																																	
4.	Handwriting																																	
Descriptive Remarks																																		

MATHEMATICS

1. Counting & Tables:

S. No.	Sub Skill	April				May				July				August				Sept.				October				Nov.				Dec.				Jan.				Feb.				March			
		4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1
1.	Counting (Numbers)																																											
2.	Tables																																											
Descriptive Remarks																																												

2. Mathematical Concepts:

S. No.	Sub Skill	April				May				July				August				Sept.				October				Nov.				Dec.				Jan.				Feb.				March			
		4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1
1.	Mathematical Concepts																																											
Descriptive Remarks																																												

3. Mathematical Shapes:

S. No.	Sub Skill	April				May				July				August				Sept.				October				Nov.				Dec.				Jan.				Feb.				March			
		4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1
1.	Mathematical Shapes																																											
Descriptive Remarks																																												

CO-CURRICULAR ASPECTS

1. Literary

S. No.	Sub Skill	April			May			July			August			Sept.			October			Nov.			Dec.			Jan.			Feb.			March		
		3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
1.	Poem Recitation																																	
2	Story telling																																	
3	Calligraphy																																	
Descriptive Remarks																																		

2. Cultural

S. No.	Sub Skill	April			May			July			August			Sept.			October			Nov.			Dec.			Jan.			Feb.			March		
		3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
1.	Singing																																	
2.	Dance																																	
3.	Acting																																	
Descriptive Remarks																																		

3. Creativity

S. No.	Sub Skill	April			May			July			August			Sept.			October			Nov.			Dec.			Jan.			Feb.			March		
		3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
1.	Interest																																	
2.	Innovative																																	
Descriptive Remarks																																		

4. Games & Sports

S. No.	Sub Skill	April			May			July			August			Sept.			October			Nov.			Dec.			Jan.			Feb.			March		
		3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
1	Interest																																	
2	Enthusiasm																																	
3	Discipline																																	
4	Team Spirit																																	
Descriptive Remarks																																		

PERSONAL AND SOCIAL QUALITIES

S. No.	Traits	April			May			July			August			Sept.			October			Nov.			Dec.			Jan.			Feb.			March		
		3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
1.	Regularity & Punctuality																																	
	Descriptive Remarks																																	
2.	Cleanliness																																	
	Descriptive Remarks																																	
3.	Initiative																																	
	Descriptive Remarks																																	
4.	Co-operative																																	
	Descriptive Remarks																																	
5.	Confidence																																	
	Descriptive Remarks																																	