



USAID
FROM THE AMERICAN PEOPLE



Revitalizing, Innovating, Strengthening Education

Revitalizing, Innovating, Strengthening Education (RISE)

A PROJECT SPONSORED BY THE
UNITED STATES AGENCY FOR INTERNATIONAL DEVELOPMENT (USAID)-
PAKISTAN,
COOPERATIVE AGREEMENT NO. 391-A-00-06-01080-00
POST-EARTHQUAKE EDUCATION RECOVERY PROGRAM

Results of the January- February 2009 Post-Test of Teacher Classroom Observation and Student Assessment

Prepared by:

RISE

Contact: Dr. Naeem Sohail Butt
Tel: 0301-850-5719
e-mail: nsohailb@rise-pk.org
House 5, Street 10, F-8/3
Islamabad, Pakistan

A COLLABORATION OF
AMERICAN INSTITUTES FOR RESEARCH
INTERNATIONAL RESCUE COMMITTEE
SUNGI DEVELOPMENT FOUNDATION
NATIONAL RURAL SUPPORT PROGRAM
SARHAD RURAL SUPPORT PROGRAM

September 2009

Table of Contents

Executive Summary	iii
I. Introduction	1
II. Methods	1
Design.....	1
Sample Schools	2
III. Results	3
Teacher Classroom Observation.....	3
<i>District Comparison</i>	6
<i>Gender Comparison</i>	7
Student Assessment	8
<i>Grades 4 and 8 English, Mathematics, and Science Post-Test</i>	9
<i>Psychometric Quality</i>	9
<i>Student Test Performance</i>	11
<i>Total Raw Score Comparison</i>	12
<i>Districtwise Comparison</i>	12
<i>Gender Comparison</i>	13
<i>Performance Level Comparison</i>	14
IV. Conclusion	15

Executive Summary

The purposes of student assessment and teacher classroom observation in RISE are to (1) develop systems for measuring student learning outcomes and train Pakistani specialists, and (2) gauge impacts of project interventions, including improved teaching-learning processes in targeted schools. Through rigorous measurement and evaluation design, student achievement and teacher classroom observation data were collected annually, starting in 2008 and continuing through 2010. The 2008 study serves as the baseline, and two post-tests will be conducted in 2009 and 2010.

Although the first post-test was scheduled to be conducted in April 2009, it was administered in January - February 2009 to adapt to the recent change in the school year (April –March for both summer and winter zone schools) in Pakistan. Prior to the January-February test administration, the teachers who were trained in the Summer 2008 had about five months to provide classroom instruction. Teachers who were trained in Winter 2009 had not begun their classes at the time of the assessment. Therefore, teachers in the summer zone schools (who were trained in summer 2008) and their students in grade 4 and 8 were included in the interim test. Note that the baseline data were also collected from those same teachers and their students back in 2008.

A sample of 124 summer zone schools were visited in the post-test data collection, out of which 35 were in Bagh, 52 in Muzaffarabad, and 37 in Mansehra. A total of 132 teachers from the 124 schools were observed. In general, teachers in both grades 4 and 8 showed outstanding improvement in all six cluster variables. In some cases, teachers of grades 4 and 8 even obtained more than double their baseline scores in the interim test (e.g., *active learning teaching, lesson planning, presentation technique, and content knowledge*). In the post-teacher classroom observation, most teachers who were rated *unsatisfactory* or *satisfactory* in the baseline were rated either *satisfactory* or *excels* in all six cluster variables. When teachers' performance was compared by district, it was also revealed that teachers in Bagh improved most, followed by teachers in Muzaffarabad and Mansehra. In a gender-wise comparison, the grade 4 male teachers received consistently higher scores than the female teachers in all six cluster variables whereas for grade 8 they both received very similar scores.

With regard to the student assessment, although students did not have adequate opportunity to acquire all the expected grade level knowledge, skills, and abilities because of the recent school year change in Pakistan, students in grade 4 have even shown some improvement. Students in grade 4 had the highest improvement from the baseline to the interim test in English followed by mathematics and then science. In contrast, students in grade 8 did not perform well on the interim test. Grade 8 students' scores in the interim test slightly declined in English and science and showed about no improvement in the mathematics.

Upcoming activities for student assessment and teacher classroom observation study are scheduled in November 2009 for the winter zone schools and in January-February 2010 for the summer zone schools. By that time, teachers who were trained in 2008 and 2009 will have had a full school year to use the teaching and learning processes (that they were trained on) in their classroom instruction. The additional months provides the time to acquire more robust information about the impact of RISE teacher training on student learning outcomes and teachers' classroom behavior.

Results of the January-February, 2009 Post-Test of Teacher Classroom Observation and Student Assessment

I. Introduction

In the USAID-Pakistan monitoring and evaluation framework, student learning outcomes are a high-level indicator for the RISE project. A proven approach to measuring learning outcomes in terms of validity, reliability, and practicality, is curriculum-based, criterion-referenced achievement testing.¹ A set of such criterion-referenced tests was administered in April-May, 2008 to set the baseline and January-February, 2009 (as the interim post test) for student achievement.

Concurrently with the student achievement testing, a teacher classroom observation study was also conducted. In the teacher observation study, teachers were observed in their classrooms twice, once before they received training from RISE (constituted the baseline measures) and then five months after the training (representing post-test measures). The sample teachers were observed and rated using a project-developed survey form while they were providing classroom instruction.

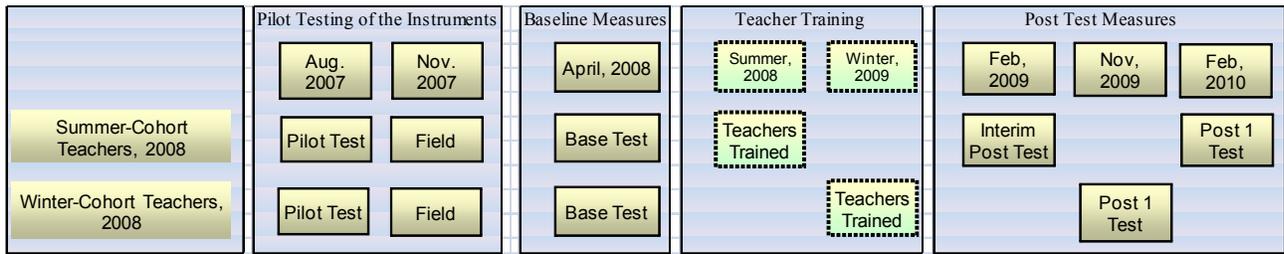
The baseline and interim student assessment and post teacher observation studies were conducted in three districts located in northwestern Pakistan: Mansehra district in the North West Frontier Province (NWFP) and Bagh and Muzaffarabad districts in Azad Jammu & Kashmir (AJK).

II. Methods

Design: The RISE project evaluation design uses multi-year achievement data for students in grades 4 and 8 in English, mathematics, and science to assess intervention effects. The evaluation features a cross-sectional design, with a baseline and two post-tests (i.e., three data collection points). Based on the initial design, the first post-test was supposed to be conducted in April 2009, however, it was administered in January-February 2009 (and should be considered interim for student assessment) to adapt to the recent change in school year in Pakistan (i.e., April–March for both summer and winter zone schools). The full scale post-test will be administered in November 2009 for winter zone schools (before they close for winter vacation in December-February) and January-February 2010 for summer zone schools.

¹ See Kellaghan, T., & Greaney, V. (2003). *Monitoring performance: Assessment and examinations in Africa*. Grand Baie, Mauritius: ADEA Biennial Meeting.

Figure 1: Research Design



Subsequent measurements in 2009 and 2010 on teachers and students will be taken from the same schools and classroom after the teachers are trained. The students will change each year, but teachers and grade levels will stay the same. This will provide a cross-sectional design to examine the effects of project-supported teacher training on student achievement and teaching-learning processes.

Sample Schools: The school year in Pakistan was changed in 2008. Now, the summer zone and winter zone schools have the same school year (April to March). Although the first post-test was supposed to have been conducted in April 2009, it was administered in January-February 2009 to adapt to this change in school year. The teachers who were trained in summer 2008 had about five months to practice the teaching processes (that they had been trained on) in their classroom instruction; the teachers who were trained in Winter 2009 had no time to use the new techniques in their classroom instruction before the January-February 2009 test administration. Therefore, only the teachers in the summer zone schools (who were trained in the last summer) and their students in grades 4 and 8 were included in the interim test. Note that the baseline data was also collected from those same teachers and their students back in 2008.

Since the teachers did not have adequate time (only five months) to complete the whole syllabus and the interim test might have been included items on certain topics that were not taught in class, students' performance on the interim test was underestimated. Thus, the student assessment results should be considered as interim, and more focus should be given to the teacher classroom observation study results, which are assumed to be more robust. Moreover, since the interim test data included only those summer zone schools from which teacher were trained in Summer 2008, readers should also be cautious while generalizing the results at the district, province or at the national level.

A total of 124 summer zone schools were visited in the post-test data collection. Of the 124 schools, 35 schools (including 38 teachers) were in Bagh, 52 (including 55 teachers) in Muzaffarabad and 37 (including 39 teachers) in Mansehra (Table 1). Note that these 132 teachers were trained in June-August 2008 and their baseline classroom observation data were collected in April 2008.

Table 1: Summary of Teacher Classroom Observation and Interim Student Assessment Sample Data

Data	District	No. of Schools	Grade 4	Grade 8	Total
Teacher Observation	Bagh	35	19	19	38
	Muzaffarabad	52	35	20	55
	Mansehra	37	25	14	39
	Total	124	79	53	132
Student Assessment	Bagh	35	115	221	336
	Muzaffarabad	52	213	261	474
	Mansehra	37	188	242	430
	Total	124	516	724	1240

For the student assessment, a total of 1240 students were tested out of which, 336 students were in Bagh (115 in grade 4 and 221 in grade 8). In Muzaffarabad, 213 of the 474 students were in grade 4 and 261 in grade 8. In Mansehra, a total of 430 students in 37 schools were assessed, with 188 in grade 4 and 242 in grade 8 (see Table 1). The interim tests were conducted in three subject areas (English, mathematics and science) for grade 4 and 8 students in January-February 2009. Each test form was comprised of 35 items.

III. Results

Teacher Classroom Observations

Teacher classroom observation data were analyzed to learn what teachers are doing in the classroom with respect to teaching processes (content and pedagogy). A total of six cluster variables were created using both quantitative and qualitative analyses of the baseline data. A quantitative advanced statistical analysis (using factor analysis) was conducted followed by a qualitative analysis to form the clusters. The variables were labeled as (i) *Active Learning Teaching*, (ii) *Lesson Planning*, (iii) *Use of Presentation Technique*, (iv) *Content Knowledge*, (v) *Teacher-Student Relationship*, and (vi) *Positive Teaching Behavior*. Note that questions within each cluster variables were rated on a four point Likert type scale: 3 for *Excels*, 2 for *Satisfactory*, 1 for *Unsatisfactory*, and 0 for *No Evidence*.

- i. *Active Learning Teaching*: This cluster variable includes six questions that are related to (1) active learning teaching techniques, (2) involving students in classroom activities, and (3) encouraging students to ask questions, and (4) encouraging student interaction, (5) listening to student responses, and (6) using teaching aids. This cluster variable was rated in a scale of a minimum of 0 (i.e., *no evidence* in all six questions) to a maximum of 18 (i.e., *excels* in all six questions). The percent scale equivalent to 0-18 scale is; 0% represents *No Evidence* (0 on the 0-18 scale); 1%-33% represents *Unsatisfactory* (greater than 0 and up to 6 on the 0-18 scale); 34%-66% represents *Satisfactory* (greater than 6 and up to 12 on the 0-18 scale); and 67% -100% represents *Excels* (greater than 12 and up to 18 on the 0-18 scale).
- ii. *Lesson Planning*: This cluster variable includes four questions that are related to how the teacher (1) introduces the lesson clearly, (2) allocates time effectively, (3) delivers the

lesson logically and coherently, and (4) assesses student understanding. This cluster variable was rated in a scale of a minimum of 0 (i.e., *no evidence* in all four questions) to a maximum of 12 (i.e., *excels* in all four questions). Thus, on average, 0% would represent *no evidence*, 1%-33% *unsatisfactory*, 34%-66% *satisfactory*, and 67%-100% *excels*.

- iii. *Use of Presentation Techniques*: This cluster variable includes only one question that asked about how teachers use presentation techniques in the classroom. This was also rated using the same Likert scale mentioned above.
- iv. *Content Knowledge*: This cluster variable only includes only one question that asked about the teacher's command of the subject matter. This was also rated using the same Likert scale mentioned above.
- v. *Teacher-Student Relationship*: This cluster variable includes questions that are related to whether the teacher (1) makes effective seating arrangement, (2) addresses student by name, (3) does not call a student by a negative nickname, and (4) does not shout in class. As this variable comprises four questions (and each was rated 0-3 scale), it was rated in a scale with a minimum of 0 (when *no evidence* in all four questions) to a maximum of 12 (when *excels* in all four questions). Thus, on average, 0% would represent *no evidence*, 1%-33% *unsatisfactory*, 34%-66% *satisfactory*, and 67% -100% *excels*.
- vi. *Positive Teaching Behavior*: This cluster variable includes questions that are related to whether the teacher (1) makes eye contact with students, (2) connects lessons to students' experience, (3) moves around the class to help students, (4) praises student work, and (5) uses positive behavior management. This cluster variable was scored in a scale with a minimum of 0 (when *no evidence* in all five questions) to a maximum of 15 (when *excels* in all five questions). Thus, on average, 0% would represent *no evidence*, 1%-33% *unsatisfactory*, 34%-66% *satisfactory*, and 67% -100% *excels*.

In general grades 4 and 8 teachers in the baseline did not perform satisfactorily in *active learning teaching*, *lesson planning*, and *use of presentation techniques*. No teacher got an average score of 30% in the respective cluster variables; overall teachers were rated *unsatisfactory*. On the other hand, they (both grades 4 and 8 teachers) obtained over 60% score in *teacher-student relationship* and over 37% score in *positive teaching behavior* and thus were rated *satisfactory*. With regard to teacher's content knowledge, overall teachers in grades 4 (with 26% scores) and 8 (with 35% scores) were rated *unsatisfactory* and *satisfactory*, respectively (Table 2).

In the post-test, both grades 4 and 8 teachers in summer zone schools obtained statistically significantly higher scores than their scores in the baseline. In most cases, they rated either *satisfactorily* or *excels* in the post-test as opposed to *unsatisfactory* or *satisfactory* in the baseline, respectively (Table 2). A one-to-one comparison has been made among teachers in the summer zone schools in the baseline and post-test. Teachers in grade 4 achieved 36% score (i.e., *satisfactory*) in the post-test as compared to 24% (i.e., *unsatisfactory*) in the baseline in *Active Learning Teaching*, 48% (i.e., *satisfactory*) compared to 20% (i.e., *unsatisfactory*) in *Lesson Planning*, 62% (i.e., *satisfactory*) compared to 25% (i.e., *unsatisfactory*) in *Presentation Technique*, 55% (i.e., *satisfactory*) compared to 29% (i.e., *unsatisfactory*) in *Content Knowledge*,

71% (i.e., *Excels*) compared to 62% (i.e., *satisfactory*) in *Teacher Student Relationship*, and 52% (i.e., *satisfactory*) compared to 40% (i.e., *satisfactory*) in *Positive Teaching Behavior*. The same pattern of results was also observed for grade 8 teachers in the post-test (Table 2).

Table 2: Percentage of Score Obtained on Teacher Observation Indicators

Gr	Indicator	Baseline				Baseline (Summer)				Post-Test (Summer)			
		Total	Bagh	Muz	Mans	Total	Bagh	Muz	Mans	Total	Bagh	Muz	Mans
4	Active Learning Teaching	25%	25%	27%	22%	24%	23%	26%	23%	36%*▲	46%	38%	25%
4	Lesson Planning	20%	18%	26%	17%	20%	15%	26%	16%	48%▲	52%	47%	44%
4	Presentation Tech.	23%	18%	26%	24%	25%	15%	28%	25%	62%▲	66%	61%	60%
4	Content Knowledge	26%	20%	37%	20%	29%	15%	39%	22%	55%▲	58%	53%	58%
4	Teacher-Student Relationship	62%	65%	68%	52%	62%	66%	70%	52%	71%▲	74%	70%	71%
4	Positive Teaching Behavior	39%	39%	44%	33%	40%	39%	46%	32%	52%▲	53%	55%	47%
8	Active Learning Teaching	24%	22%	27%	24%	23%	21%	27%	22%	47%*▲	60%	45%	31%
8	Lesson Planning	23%	20%	27%	21%	23%	17%	27%	25%	59%▲	66%	56%	51%
8	Presentation Tech.	28%	20%	38%	27%	30%	20%	38%	32%	66%▲	67%	61%	70%
8	Content Knowledge	35%	30%	43%	31%	37%	27%	43%	44%	68%▲	76%	62%	64%
8	Teacher-Student Relationship	60%	64%	61%	54%	61%	63%	61%	58%	72%▲	74%	70%	70%
8	Positive Teaching Behavior	37%	36%	40%	34%	38%	37%	40%	36%	58%▲	62%	59%	50%

Note: 0% = No Evidence, 1%-33% = Unsatisfactory, 34%-66%= Satisfactory, 67%-100% = Excels; * denotes statistically significant difference at $p < 0.05$; ▲ represents improvement.

When the frequency of the teachers in each rating category in the baseline and post-test were compared, it is evident that teachers have made tremendous improvement in their classroom behavior (Table 3). Please note that the same teachers were observed both in the baseline and post-test and their performance was evaluated using the same rating scale, so any decline in the percentage in the lower rating categories would represent improvement. In grade 4, about 78% of the sample baseline teachers (in the summer zone schools only) were rated *unsatisfactory* (74%) and *no evidence* (4%) categories in *active learning teaching* whereas in the post-test only one-half of the 78% of teachers in the baseline were rated in those categories (7% in *no evidence* 32% *unsatisfactory*). The result was more interesting for the *content knowledge* cluster variable; about 50% of the grade 4 baseline teachers were rated *no evidence* (42%) and *unsatisfactory* (8%)

categories, no teachers in the post-test were in those categories. A similar pattern of results was also observed for grade 4 teachers in *presentation technique* (Table 3).

Table 3: Percentage of Teachers Rated in the Baseline and Post-test

Gr.	Indicator	Baseline (Summer Zone)				Post-test (Summer Zone)			
		NE	US	S	E	NE	US	S	E
4	Active Learning Teaching	4%	74%	22%		7%	32%	50%	11%
4	Lesson Planning	13%	65%	18%	4%		14%	70%	16%
4	Presentation Tech.	50%	8%	20%	22%			28%	72%
4	Content Knowledge	42%	8%	18%	32%			51%	49%
4	Teacher-Student Relationship		8%	33%	59%			27%	73%
4	Positive Teaching Behavior		40%	50%	10%		13%	61%	26%
8	Active Learning Teaching	2%	82%	16%		4%	19%	60%	17%
8	Lesson Planning	6%	62%	28%	4%	2%	2%	64%	32%
8	Presentation Tech.	40%	4%	28%	28%	2%		15%	83%
8	Content Knowledge	32%	4%	22%	42%	2%		23%	76%
8	Teacher-Student Relationship		6%	44%	50%		2%	21%	77%
8	Positive Teaching Behavior		38%	58%	4%	2%	4%	60%	34%

Note: NE – No Evidence, US – Unsatisfactory, S – Satisfactory, E – Excels.

Over 90% of grade 8 baseline sample teachers in the post-test were rated either *Satisfactory* or *Excels* in all six cluster variables, except for *Active Learning Teaching*. In the baseline, about 84% of the baseline sample teachers were rated either **No Evidence** (2%) or **Unsatisfactory** (82%) in *Active Learning Teaching*. In contrast, only 23% of them were rated either **No Evidence** (4%) or **Unsatisfactory** (19%) in the post-test (Table 3).

District Comparison: The classroom performance of grades 4 and 8 teachers varied substantially both in the baseline (with varying baseline estimates in Table 2) and post-test; it is difficult to make any inference about which district teachers--whether in Bagh, Muzaffarabad, or Mansehra--have improved significantly in teaching and learning process due to the project-supported teacher training, without bringing all three districts' teacher classroom performance (ratings) in the baseline on the six cluster variables at the same starting point. An analysis of covariance (ANCOVA) was utilized to make a district-wise comparison among the teachers (Table 4). In this statistical method, the teachers' baseline rating score was used as a covariate for the post-test. The covariates make the districts statistically equivalent on the baseline rating score so that the districts can be evaluated on an equal basis on the post-tests; it is similar to making sure that a race is fair by having two runners begin at the same starting line, and not in front or behind the other runner.

It was revealed from the ANCOVA that teachers in grade 4 (in all three districts) obtained an equivalent estimated score of 25% (i.e., *unsatisfactory*) on *Active Learning Teaching* in the baseline, which is considered to be the reference point for fair comparisons. In contrast, teachers in Bagh scored double (50%, *satisfactory*) in the post-test compared to their score in the baseline (25%, *unsatisfactory*); teachers in Muzaffarabad and Mansehra obtained scores of 39% (barely *satisfactory*) and 24% (*unsatisfactory*), respectively (Table 4). This is the only cluster variable (i.e., *Active Learning Teaching*) in which teachers in Mansehra did not show improvement. In the remaining five cluster variables teachers in all three districts have improved substantially;

they were all rated higher in the post-test compared to their baseline (e.g., *unsatisfactory* to *satisfactory* or *satisfactory* to *excels*). The highest growth was observed for teachers in Bagh, followed by Muzaffarabad and then Mansehra in all cluster variables, except for **Content Knowledge** and **Teacher-Student Relationship**. In these two cluster variables, teachers in Mansehra outperformed the teachers in Muzaffarabad.

Table 4: Teacher Classroom Behavior: District wise Comparison

Grade	Cluster Variable	Estimated Baseline Scores (summer zone)	Estimated Post-Test Score (Summer Zone)			
			Total	Bagh	Muz	Mans
4	Active Learning Teaching	25%	38%*	50%▲	39%▲	24%▼
4	Lesson Planning	21%	49%	55%▲	48%▲	44%▲
4	Presentation Tech.	24%	62%	66%▲	61%▲	60%▲
4	Content Knowledge	30%	56%	60%▲	53%▲	56%▲
4	Teacher-Student Relationship	65%	72%	73%▲	71%▲	72%▲
4	Positive Teaching Behavior	40%	53%	53%▲	55%▲	49%▲
4	Overall	34%	55%	60%▲	55%▲	51%▲
8	Active Learning Teaching	23%	46%*	59%▲	45%▲	33%▲
8	Lesson Planning	23%	58%*	65%▲	57%▲	52%▲
8	Presentation Tech.	31%	67%	68%▲	61%▲	70%▲
8	Content Knowledge	39%	67%	75%▲	61%▲	64%▲
8	Teacher-Student Relationship	61%	72%	74%▲	72%▲	72%▲
8	Positive Teaching Behavior	39%	57%	60%▲	60%▲	50%▲
8	Overall	36%	61%	67%▲	59%▲	57%▲

Note: 0% = No Evidence, 1%-33% = Unsatisfactory, 34%-66%= Satisfactory, 67%-100% = Excels; * denotes statistically significant difference at $p < 0.05$ among the districts; ▲ represents improvement and ▼ represents decline.

For the grade 8 post-test, teachers in three districts performed statistically significantly different in the **Active Learning Teaching** and **Lesson Planning** (Table 4). The teachers in Bagh, Muzaffarabad, and Mansehra obtained average scores of 59% (*satisfactory*), 45% (*satisfactory*), and 33% (*unsatisfactory*) in the post-test as compared to their equivalent baseline estimated score of 23% in the **Active Learning Teaching** cluster variable. In **Lesson Planning**, teachers in all three districts were rated *satisfactory* in the post-test as opposed to *unsatisfactory* in the baseline. For the remaining four cluster variables, although teachers in Bagh outperformed their counterparts in Muzaffarabad and Mansehra, the differences were not statistically significant; they were rated either *satisfactory* or *excels*. Overall, teachers in Bagh progressed highest among the three districts, followed by Muzaffarabad and Mansehra.

Gender Comparison: When the performance of teachers in the post-test was analyzed by gender, it was revealed (Table 5) that overall male teachers in grade 4 (male=57%, female=50%) and female teachers in grade 8 (male=61%, female=62%) performed relatively better than their respective counterparts, though female teachers in grade 4 scored higher than male teachers in the baseline (male=31%, female=37%). In the post-test, the difference between male and female teachers in grade 4 was particularly statistically significant in **Active Learning Teaching**

(male=41%, female=29%); for the remaining five cluster variables although males outperformed females, the differences were non-significant. When the teachers' performance in the post-test was compared with their baseline scores, it was quite noticeable that male teachers made higher progress than the females in all six cluster variables. In the baseline both male and female teachers were rated *unsatisfactory* in four cluster variables and *satisfactory* in the other two; whereas in the post-test they received ratings of either *satisfactory* or *excels* in all six cluster variables with only the exception of females in *Active Learning Teaching* in which they rated *unsatisfactory*.

Table 5: Teacher Observation Indicators by Gender

Gr.	Indicator	Baseline			Baseline (Summer zone)			Post-test (Summer zone)		
		Total	Female	Male	Total	Female	Male	Total	Female	Male
4	Active Learning Teaching	24%	27%	21%	24%	28%	21%	36%*▲	29%▲	41%▲
4	Lesson Planning	20%	21%	20%	20%	23%	18%	48%▲	44%▲	50%▲
4	Presentation Tech.	24%	26%	21%	25%	29%	21%	62%▲	57%▲	65%▲
4	Content Knowledge	26%	24%	27%	29%	32%	25%	55%▲	53%▲	57%▲
4	Teacher-Student Relationship	60%	60%	59%	62%	64%	61%	71%▲	71%▲	72%▲
4	Positive Teaching Behavior	38%	39%	36%	40%	43%	37%	52%▲	48%▲	55%▲
4	Overall	32%	33%	31%	33%	37%	31%	54%▲	50%▲	57%▲
8	Active Learning Teaching	24%	28%	21%	23%	28%	21%	47%▲	48%▲	46%▲
8	Lesson Planning	22%	21%	22%	23%	27%	21%	59%▲	60%▲	58%▲
8	Presentation Tech.	26%	29%	25%	30%	36%	27%	66%▲	67%▲	65%▲
8	Content Knowledge	33%	34%	33%	37%	40%	36%	68%▲	67%▲	68%▲
8	Teacher-Student Relationship	60%	63%	57%	61%	68%	58%	72%▲	73%▲	71%▲
8	Positive Teaching Behavior	36%	41%	33%	38%	44%	35%	58%▲	58%▲	57%▲
8	Overall	34%	36%	32%	35%	41%	33%	62%▲	62%▲	61%▲

Note: 0% = No Evidence, 1%-33% = Unsatisfactory, 34%-66%= Satisfactory, 67%-100% = Excels; * denotes statistically significant difference at $p < 0.05$; ▲ represents improvement.

In the grade 8 post-test, female teachers scored higher in five of the six cluster variables, with the exception of *Content Knowledge* (Table 5). However, none of the differences were statistically significant. When they were compared with their baseline scores, it was revealed that improvement of the male teachers was much higher than that of female teachers in all cluster variables; the differences in scores between male and female reduced substantially in the post-test (62% - 61% = 1%) than it was in the baseline (41% - 33% = 8%). Teachers were also rated higher in the post-test (mostly *satisfactory* and *excels*) than they were in the baseline (mostly *unsatisfactory*, a few *satisfactory*, and *excels*).

Student Assessment

As it was stated earlier that the teachers did not have adequate time (about 4 - 5 months as opposed to 8-9 months) to complete the whole syllabus due to the short school year in 2008-2009, students' performance on the interim test would have been underestimated. Moreover,

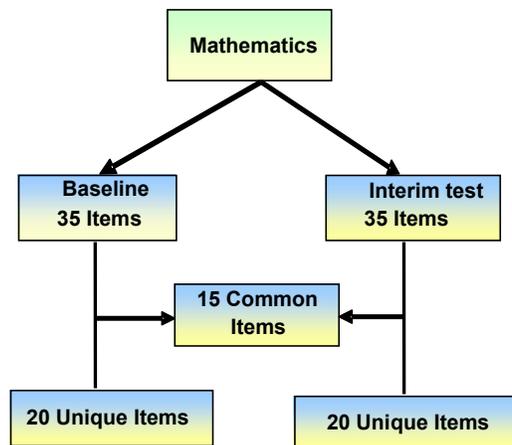
only a sample of summer zone schools was included in the interim test. Therefore, the student assessment results may be less robust with limited power. Before reporting the student assessment results, it is necessary to present the characteristics (psychometric quality) of the interim tests that may help in interpreting the results. They are described in the following.

Grades 4 and 8 English, Mathematics, and Science Interim Test: Test blueprints, which were created in 2007 and are based on the national curriculum, were used as a guide to ensure that the tests in both grades and all three subjects represented measurable objectives in the curriculum. The test blueprints, or test content matrices, will be maintained throughout the multi-year assessment period.

Test forms containing 35 multiple choice items each were created for all subjects: English, mathematics, and science. Items on these interim test forms were either pilot tested or field tested prior to this administration and had acceptable psychometric properties; item discrimination was considered as a criterion when selecting items for these interim tests.

From each baseline test form, a subset of 15 items was chosen to carry over from the 2008 baseline test to the 2009 interim test form. Psychometric experience has shown that this is an adequate number to provide the basis for statistically equating forms within each subject area and grade level. The equating items were selected so that each subset mirrors the baseline test in terms of content domain and difficulty. A sample baseline (2008) and interim test (2009) structure is presented in figure 2.

Figure 2: A Sample Baseline and Interim test (2009) Structure



Psychometric Quality: To examine the psychometric quality of the tests, both item level and test level quantitative analyses were conducted. Each item was evaluated with respect to its difficulty (or p-value) and discrimination (or point-biserial correlation) values. Each test was assessed based on the reliability coefficient of internal consistency (Cronbach, 1951)².

² Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16, 297–334.

Table 6 presents the average p-value³ of items both on the baseline and interim tests. Overall, the item p-values for both grades 4 and 8 tests were within acceptable and expected ranges (0.20 – 0.90). However, students in grade 4 found the English (baseline = 0.28, interim test = 0.32) and mathematics (baseline = 0.32, interim test = 0.35) in the interim test relatively easier as compared to the baseline, whereas they found science (baseline = 0.43, interim test = 0.41) relatively more difficult in the interim test. So, for each item on the grade 4 English test, on average about 28% students in the baseline and 32% of in the interim test got the item right. A similar pattern was also observed for grade 8 tests.

Table 6: Overall Test Difficulty Estimates by Subject Area

Grade	Subject Area	No. of Items	Baseline Test			Interim Test		
			P-value	Discr.	Rel	P-value	Discr.	Rel
4	English	35	0.28	0.21	0.40	0.32	0.21	0.44
4	Math	35	0.32	0.29	0.68	0.35	0.24	0.53
4	Science	35	0.43	0.33	0.76	0.41	0.30	0.72
8	English	35	0.35	0.32	0.75	0.35	0.26	0.67
8	Math	35	0.42	0.34	0.78	0.43	0.26	0.59
8	Science	35	0.44	0.30	0.70	0.41	0.23	0.48

Although the tests in the baseline and the interim test appear to be varied in difficulty (e.g., grade 4 English and Mathematics got easier and science got more difficult in the interim test), the differences in difficulty are adjusted when comparing student performance in the interim test as compared to their performance in the baseline. In other words, students’ scores both in the baseline and interim test are brought on to the same measurement scale so that any improvement or decline in their performance due to project-supported teacher training can be evident. The process of adjusting the test difficulty and bringing students’ scores on the same scale is called test equating. Through the equating procedure, psychometricians produce an answer to this question: if a student is taking the grade 4 interim test, what would have been his/her score on the baseline if he/she had taken the baseline test. In other words, we do not want to draw conclusions based on the interim test form being easier than the baseline test form. Students’ equated interim test scores (that are converted into the baseline) have been presented in this report.

As seen in Table 6, the discrimination⁴ values in the interim tests are consistently lower than they were in the baseline tests. This happened because of the restriction of range problem in correlation; note that the item discrimination is nothing but a simple Pearson correlation between item score (0 for a wrong and 1 for correct) and total test score. The restriction of range problem appeared due to the fact that only the summer zone schools as opposed to both summer and winter zone schools were covered in the interim tests (i.e., the sample was restricted to the

³ Item difficulty is defined as the average proportion of points achieved on an item by the students. It is calculated by obtaining the average score on an item and dividing by the maximum possible score for the item. In general, the greater the percentage of students who answer the item correctly, the easier the item is considered to be.

⁴ Item discrimination refers to the process of contrasting performance between higher- and lower-performing students on an item. An item is said to have higher discriminating power when higher-performing students do better on the item compared to lower-performing students.

summer zone schools only); as a result correlation values (item discrimination) get underestimated.

Although average item discrimination values for the interim test were relatively lower than their values in the baseline, they exceeded the accepted minimum of 0.20. The average discrimination values for the baseline and the interim tests ranged from 0.21 to 0.34 and 0.21 to 0.30, respectively. An average discrimination value on a test would be interpreted by saying that for each item on the English test, the higher-performing students (or those with higher total scores) had a 21 percent higher chance of answering the item correctly compared to lower-performing students (or those with lower total scores).

When the tests were evaluated with respect to their reliability coefficients of internal consistency⁵, it was revealed that the baseline and the interim tests had values of (0.40 to 0.78) and (0.44 to 0.72) respectively. Again, the reliability values for most interim tests (except for the grade 4 English) were much lower than they were in the baseline. This was due to the fact that the interim tests had on average lower discrimination values (Table 6). Please note that both item discrimination and test reliability are inter-related; a test with higher discriminating items will have a higher reliability coefficient than a test with fewer discriminating items. The discrimination values for the interim tests were lower due to the restriction of range problem.

Student Test Performance: Student performance was rated on two measurement scales: raw score⁶ and scaled score (ranging from 100 to 500). The scaled score is more robust than the raw score in determining the growth of student performance from one year to the next, as it captures differences in test difficulty in both years. Note that reporting student performance using scaled scores does not change the order of student position on the raw score scale. In addition, the results of the students were also reported by performance level categories.

⁵ The reliability of internal consistency is used to judge the consistency of results across items on the same test. Essentially, we are comparing test items that measure the same construct to determine the tests internal consistency.

⁶ Raw Score: Sum of correct responses to the items on the test.

Table 7: Overall Student Performance by District

Gr.	Subject Area	Baseline (Overall)				Baseline (Summer Zone Only)				Interim Test (Summer Zone Only)			
		Total	Bagh	Muz	Mans	Total	Bagh	Muz	Mans	Total	Bagh	Muz	Mans
4	English	9.9 *	10.3	10.2	9.5	9.6*	10.0	10.1	8.7	11.3*	13.0	11.6	9.8
		(268)	(273)	(272)	(264)	(265)	(269)	(271)	(255)	(282*)▲	(299)▲	(285)▲	(269)▲
4	Math	11.5*	11.6	11.0	11.7	10.9*	11.4	10.8	10.6	12.2*	14.1	11.8	11.4
		(274)	(275)	(271)	(275)	(269)	(273)	(269)	(264)	(279*)▲	(292)▲	(277)▲	(274)▲
4	Science	15.1	15.4	15.3	14.8	14.7*	15.4	15.3	13.4	15.1*	17.3	15.5	13.4
		(266)	(269)	(268)	(264)	(263)	(268)	(269)	(251)	(266*)▲	(286)▲	(270)▲	(251)
8	English	12.4*	13.3	14.3	10.9	12.7*	13.4	14.3	10.5	12.2*	13.0	13.0	10.7
		(254)	(264)	(273)	(238)	(258)	(265)	(273)	(234)	(253*)▼	(260)▼	(261)▼	(237)▲
8	Math	14.5*	15.9	15.2	13.3	14.9*	15.9	15.2	13.7	15.0*	15.8	14.5	14.7
		(267)	(276)	(272)	(259)	(270)	(276)	(272)	(262)	(271*)▲	(276)	(268)▼	(269)▲
8	Science	15.4*	15.5	16.1	14.9	15.5*	15.5	16.1	14.9	14.4	14.7	14.5	14.0
		(268)	(270)	(275)	(264)	(270)	(269)	(275)	(264)	(260)▼	(262)▼	(261)▼	(257)▼

Note: Numbers in parenthesis represents avg. scaled scores; * denotes statistically significant difference at $p < 0.05$; ▲ represents improvement and ▼ represents decline.

Total Raw Score Comparison: Student performance showed some variation in both baseline and interim tests. For the grade 4 baseline (summer and winter zone schools together), baseline (summer schools only), and the interim test (summer schools), the average raw scores for students were 9.9, 9.6, and 11.3 in English; 11.5, 10.9, and 12.2 in mathematics; and 15.1, 14.7, and 15.1 in science respectively out of a possible score of 35 (Table 7). Although the average raw scores in the interim tests were much higher in all three subject areas than they were in the baseline, it is important to note that the tests in the interim tests were also relatively easier for English and mathematics. So the improvement in the interim test over the baseline may be due to the project-supported teacher training or it could be because of the easier tests. The raw score comparison between the baseline and the interim tests therefore may not be very relevant; the scaled score comparison would be most relevant as they are brought on the same scale to assess the change due to the training. Their corresponding averaged scaled scores were estimated 268, 265, and 282 in English; 274, 269, and 279 in mathematics; and 266, 263, and 266 in science respectively (Table 7). It is evident from the scaled score comparison that students in grade 4 showed improvement in all three subjects due to the project-supported teacher training.

In contrast, performance of grade 8 students in the interim tests was not improved over their performance in the baseline. On average, students obtained total raw and scaled scores of 12.4 (254), 12.7 (258), and 12.2 (253) in the English subject area in the baseline for summer and winter zone schools together, the baseline for summer zone schools only, and the interim test for summer zone schools, respectively (Table 7). Their corresponding scores were 14.5 (267), 14.9 (270), 15.0 (271) in mathematics and 15.4 (268), 15.5 (270), and 14.4 (260) in science. There could be several factors which can be attributed to the no improvement situation. As was stated earlier, teachers in the summer zone schools had about five months to complete the syllabus

before the interim test assessments were administered. The length of time might not have been sufficient given the number of competencies in each subject-grade. Students might have been tested on certain competencies that might not have been taught in some schools before the interim tests were even administered.

Table 8: Student Performance: District-wise Comparison

Grade	Subject Area	Estimated Baseline Scores (summer zone)	Estimated Interim Test Scaled Scores (Summer Zone)			
			Total	Bagh	Muz	Mans
4	English	266.4	285.6	303.0*▲	283.6▲	270.4▲
4	Math	267.5	280.7	292.0*▲	279.1▲	271.1▲
4	Science	266.2	268.3	283.6*▲	268.6▲	252.9▼
8	English	259.7	252.4	258.5▼	254.9▼	244.0▼
8	Math	271.6	269.3	272.6▲	265.8▼	269.6▼
8	Science	271.3	259.2	262.7▼	256.6▼	258.2▼

Note: * denotes statistically significant difference among the districts at $p < 0.05$;

▲ represents improvement and ▼ represents decline.

District-wise Comparison: Student performance was also compared by district (Bagh, Muzaffarabad, and Mansehra). Table 8 shows the average raw and scaled scores on the baseline and interim test for each subject (English, mathematics, and science), district, and grade (4 and 8). The average scores are appeared to be varied both in the baseline and interim test. The differences among the districts for all subject areas in grades 4 and 8 were evaluated using an Analysis of Covariance (ANCOVA). In this statistical method, the baseline scaled score was used as a covariate for the interim test. The covariates make the districts statistically equivalent on the baseline test score so that the districts can be evaluated on an equal basis on the interim tests; it is similar to making sure that a race is fair by having two runners begin at the same starting line, and not in front or behind the other runner. For example, the scaled scores on the grade 4 English interim test were estimated for the districts (through the ANCOVA) after equalizing the baseline scores for the districts at 266.4; so, the difference among the districts (Bagh=303.0, Muzaffarabad=283.6, and Mansehra=270.4) on the interim test was noticeable after putting the districts at the same starting point on the baseline.

In contrast, none of the three districts (except Bagh for grade 8 mathematics) showed improvement over their baseline scaled scores for the grade 8 interim test; however, the improvement was statistically non-significant (Table 8).

Gender Comparison: Student performance was also compared by gender (Table 9). In the grade 4 baseline (both summer and winter zones together), male and female students performed better than their counterparts in mathematics (male=275, female=273) and science (male=263, female=269) respectively; they obtained exactly the same scores in English (268). When the analysis was done only for the baseline summer zone schools, it was revealed that male students received higher average scores than female students both in mathematics (male=271, female=267) and science (male=263, females=262) and females students received higher scores in English (male=264, female=266). In the interim test, male students outperformed females both

in English (male=283, female=281) and mathematics (male=283, female=275) and females outperformed males in science (male=265, female=267).

Table 9: Student Performance by Gender

Gr.	Subject Area	Baseline (Overall)			Baseline (Summer Zone)			Interim Test (Summer Zone)		
		Total	Female	Male	Total	Female	Male	Total	Female	Male
4	English	9.9	9.9	9.9	9.64	9.78	9.50	11.3	11.2	11.4
		(268)	(268)	(268)	(265)	(266)	(264)	(282)▲	(281)▲	(283)▲
4	Math	11.5	11.4	11.7	10.9*	10.6	11.1	12.2*	11.5	12.7
		(274)	(273)	(275)	(269)	(267)	(271)	(279*)▲	(275)▲	(283)▲
4	Science	15.1*	15.4	14.8	14.7	14.9	14.6	15.1	15.2	15.0
		(266)	(269)	(263)	(263)	(262)	(263)	(266)▲	(267)▲	(265)▲
8	English	12.4*	13	11.9	12.7*	13.8	12.0	12.2*	13.1	11.8
		(254)	(260)	(248)	(258)	(269)	(250)	(253)▼	(262)▼	(248)▼
8	Math	14.5*	14.1	14.9	14.90	14.8	15.0	15.0*	14.3	15.3
		(267)	(265)	(270)	(270)	(270)	(270)	(271*)▲	(267)▼	(273)▲
8	Science	15.4*	16.2	14.7	15.5*	16.4	14.9	14.4	14.8	14.2
		(268)	(276)	(262)	(270)	(278)	(264)	(260)▼	(263)▼	(258)▼

Note: Numbers in parenthesis represents avg. scaled scores; * denotes statistically significant difference at $p < 0.05$; ▲ represents improvement and ▼ represents decline.

A similar pattern of results was observed for students in the grade 8 baseline test (both summer and winter zone together); females outperformed males both in English (male=248, female=260) and science (male=262, female=276) and males outperformed females in mathematics (male=270, female=265). When baseline summer zone schools were only considered, female students performed better than males both in English and science, with no difference in mathematics. In contrast, in the interim test female students received higher average scaled scores than males both in English (male=248, female=262) and science (male=258, female=263) and males secured higher scores in mathematics (male=273, female=267) (Table 9). Both groups' performance declined (except for males in mathematics) in the interim test as compared to the baseline. As it was stated earlier, either teachers did not have adequate time to finish the syllabus or some competencies may have been tested that were not taught by the time the interim test was administered.

Performance Level Categories: Scores on the tests only provide information about how students performed on the test. They did not provide specific information about how much students at different score points know and are able to do. In order to gather that level of information, we classified students in the baseline into four performance level categories (i.e., *unsatisfactory*, *needs improvement*, *satisfactory*, and *advanced*) using a procedure called standard setting. This is similar to classifying students by letter grade (A, B, C, etc.), except that the categories are mapped out on a scale that does not change, and the different forms are equated. Another reason for doing standard setting is to keep track of student growth from one year to the next by

comparing the percentage of students in each category. A modified version of the Angoff Yes/No method (Plake, Ferdous, Buckendahl, & Impara, 2005)⁷ was used for standard setting.

Table 10: Student Performance Level Categories by Districts

Grade	Subject Area	Baseline (Overall)				Baseline (Summer Zone)				Interim Test (Summer Zone)			
		US	NI	S	A	US	NI	S	A	US	NI	S	A
4	English	8%	72%	20%	0%	10%	72%	18%	0%	7%▼	75%▲	18%	0%
4	Math	10%	70%	20%	0%	12%	72%	16%	0%	10%▼	75%▲	15%▼	0%
4	Science	6%	69%	24%	1%	7%	70%	22%	1%	9%▲	71%▲	19%▼	1%
4	Overall	8%	70%	21%	1%	10%	71%	19%	0%	9%▼	74%▲	17%▼	0%
8	English	10%	72%	16%	2%	9%	72%	16%	3%	10%▲	78%▲	12%▼	0%▼
8	Math	8%	72%	20%	0%	7%	71%	22%	0%	3%▼	81%▲	16%▼	0%
8	Science	5%	70%	24%	1%	4%	70%	25%	1%	6%▲	85%▲	9%▼	0%▼
8	Overall	8%	71%	20%	1%	7%	71%	21%	1%	6%▼	81%▲	12%▼	0%▼

Note: ▲ represents improvement and ▼ represents decline.

Table 10 above shows the percentages of student scores in each performance category both in the baseline and interim test by grade level and subject. In the grade 4 English baseline test (both summer and winter schools together), about 8% of sample students were classified as *unsatisfactory*, 72% *needs improvement*, 20% *satisfactory* and none *advanced*. The corresponding percentages for students in the summer zone schools only in the baseline were 10%, 72%, 18% and 0% respectively. In contrast, for the interim test about 7% of the students were classified into the *unsatisfactory*, 75% into the *needs improvement*, 18% into the *satisfactory*, and 0% into the *advanced* categories. In looking at the comparison between the baseline and interim test scores for students in the summer zone schools, we find that students in the *unsatisfactory* category has dropped down to 7% in the interim test from 10% in the baseline and increased to 75% in the *needs improvement* category from 72% in the baseline. Percentages for the *satisfactory* and *advanced* categories remained the same for both the baseline and interim test. One possible interpretation would be that about 3% students who were in the *unsatisfactory* category in the baseline progressed to the *needs improvement* category in the interim test.

For grade 8 mathematics, about 3% of the students were classified into the *unsatisfactory* category in the interim test as opposed to 7% in the baseline; 81% as compared to 71% into *needs improvement*; 16% as compared to 22% into *satisfactory*; and none into the *advanced* category (Table 10). This would indicate that about 4% of the students who were in the *unsatisfactory* category in the baseline must have moved up to the *needs improvement* category, but about 6% of the students who were in the *satisfactory* category must have moved down to the *needs improvement* category. The students moving up from one category to the next or moving down from one to the lower category are usually called borderline students and must have received test scores that are much closer to the lower or upper cut off scores. This pattern of

⁷ Plake, B. S., Ferdous, A. A. Buckendahl, C., & Impara J. (2005). Setting Multiple Performance Standards Using the Yes/No Method: An Alternative Item Mapping Method. Paper presented to the meeting of the National Council on Measurement in Education, Montreal, Canada.

results was also observed for most subject areas. Note that in an ideal situation for improvement, we would expect that the percentage of students (from the baseline to interim test) goes down for *unsatisfactory* and increases for *needs improvement*, *satisfactory*, and *advanced* categories.

IV. Conclusion

Student assessment and teacher observation interim test data were analyzed separately to report how students had performed on the interim tests as compared to their baselines. The comparisons were done both at the aggregated (overall between the baseline and interim test) and disaggregated levels (comparison by districts and gender). As the interim test includes only summer zone schools (because of the recent school year change in Pakistan), in order to evaluate the impacts of RISE teacher training on student learning outcomes and teacher classroom behavior, the baseline data were also reanalyzed separately for summer zone schools only, thus making one-to-one comparison between the baseline and the interim test among the summer zone schools.

Teacher Classroom Observation

Overall, teachers in both grades 4 and 8 performed substantially better in the interim test than in the baseline. In most cases, they rated either *satisfactorily* or *excels* in the interim test as opposed to *unsatisfactory* or *satisfactory* in the baseline, respectively. Teachers in grade 4 received a *satisfactory* rating in the interim test as compared to an *unsatisfactory* rating in the baseline in *Active Learning Teaching*, *Lesson Planning*, *Presentation Technique*, and *Content Knowledge*. In *Teacher Student Relationship*, they were rated *excels* and *satisfactory* in the interim test and baseline, respectively. They were also rated *satisfactory* in *Positive Teaching Behavior* in both the baseline and interim test. The same pattern of results was also observed for grade 8 teachers in the interim test. When the percentage of teachers (i.e., the same teacher observed both times) in each rating category in the baseline and interim test were compared, it was evident that teachers have made tremendous improvement in their classroom behavior (in all six cluster variables). Most of the teachers who were in the *no evidence* and *unsatisfactory* categories in the baseline had moved up to *satisfactory* category in the interim test. This pattern was observed both for grade 4 and 8 teachers.

The classroom performance of grades 4 and 8 teachers varied substantially both in the baseline (with varying baseline estimates) and the interim test; it is difficult to make any inference about the district in which teachers improved the most. The results showed that grade 4 teachers in Bagh, followed by Muzaffarabad and then Mansehra in all cluster variables, except for *content knowledge* and *teacher-student relationship*. In *content knowledge* and *teacher-student relationship*, teachers in Mansehra outperformed the teachers in Muzaffarabad. The similar pattern was also observed for grade 8.

When the performance of teachers in the interim test was analyzed by gender, it was revealed that overall male teachers in grade 4 and female teachers in grade 8 performed relatively better than their respective counterparts, though female teachers in grade 4 scored higher than male teachers in the baseline. In the baseline both grade 4 male and females teachers were rated *unsatisfactory* in four cluster variables and *satisfactory* in the other two; whereas in the interim test they received ratings of either *satisfactory* or *excels* in all six cluster variables with the only

exception for females in *active learning teaching* in which they rated *unsatisfactory*. In contrast, grade 8 female teachers in the interim test scored higher in five of the six cluster variables, except in *content knowledge*. However, none of the differences were statistically significant.

Student Assessment

Students in grade 4 have shown some improvement, though they had only 4-5 months of school year at the time of interim post-test because of the change in school year in Pakistan, compared to a regular school year (8-9 months) needed to acquire all the expected grade 4 knowledge, skills, and abilities. Students in grade 4 revealed highest improvement from the baseline to the interim test in English followed by mathematics and then science. This finding was also supported by teachers' classroom performance in the post-test as compared to the baseline. Overall, teachers in the post-test obtained higher scores in all six cluster variables (i.e., teacher classroom behavior).

When the student assessment results for grade 4 were compared by district, the students in Bagh made the highest improvement (in all three subjects) in the interim test as compared to the baseline, followed by Muzaffarabad and Mansehra. The results were also supported by teacher observation results; teachers in Bagh outperformed teachers in Muzaffarabad and Mansehra in most cluster variables. Student performance was also compared by gender. It was revealed that both male and females students in grade 4 obtained higher scores in the interim tests than they obtained in the baseline. However, male students outperformed females in English and mathematics and females outperformed males in science in the interim tests.

In contrast, students in grade 8 did not perform well on the interim test. Students' scores in the interim test declined in English and science and showed about no improvement in mathematics. It could be due to number of factors: (1) in general, the grade 8 syllabi are much longer than the grade 4 ones so teachers did not have adequate time to complete the syllabus due to the short school year in 2008-09 (i.e., recent change in school year in Pakistan), (2) the interim test might have had a few items that tested knowledge, skills, and abilities (KSAs) that might not have been taught yet by the teachers in their classes. However, the difference in the scores between the baseline and the interim test are not too far apart. Please note that students' low performance on the interim test might be due to the second factor mentioned above.

With regard to the district-wise comparison, grade 8 students in Bagh district performed better than students in Muzaffarabad and Mansehra in all three subjects in the interim test, though the differences were not statistically significant. Note that students' performance in all three districts declined as compared to their statistically equivalent baseline scores, except for mathematics in which students in Bagh received about the same scores both in the baseline and interim test. In a gender-wise comparison, although female students received higher average scaled scores than male students both in English and science and male students secured higher scores in mathematics, both groups' performance declined (except for male students in mathematics) in the interim test as compared to the baseline.

It is to be noted that the student assessment interim test study has its own limitations, particularly with regard to lack of adequate classroom instruction time (short school year due to the change in school year). Therefore, it is recommended that the findings of the interim test student

assessment should be considered as interim. The teacher classroom observation findings, though, have shown an overall success of RISE teacher training for both grades 4 and 8. The next upcoming activity for the student assessment and teacher observation study is full-phased post-tests in November 2009 (winter zone schools) and January – February, 2010 (summer zone schools). The upcoming post-tests will include both summer and winter zone schools and the teachers will have the full school year for classroom instruction to complete the syllabi. These post-tests will provide a more accurate evaluation of the actual project impact on teacher behavior and student learning outcomes.