

PISA Year 2 Pre- and Post-test Data Analysis
Karen Guo
Susan Lowes
Institute for Learning Technologies
Teachers College/Columbia University
July 13, 2009

Teacher test: 26 questions, including 22 science questions and 4 engineering questions.
 Student test: 23 questions, including 19 science questions and 4 engineering questions.

Teacher Test Analysis

Background

Initially, forty-nine treatment teachers and thirty-six comparison teachers participated in Year 2 of the PISA project. However, six treatment teachers left in the middle of the summer workshop, and one comparison teacher left during the school year, resulting in forty-three valid tests in the treatment group and thirty-five valid tests in the comparison group. Of the forty-three treatment teachers, thirty had been treatment teachers in Year 1, two had been comparison teachers in Year 1, and eleven are newly recruited. Of the thirty-five comparison teachers, thirteen had been comparison teachers in Year 1, one had been a treatment teacher in Year 1, and twenty-one were newly recruited:

Group	Count (N=78)	From Year 1 treatment	From Year 1 comparison	Year 2 new recruited
Treatment	43	30	2	11
Comparison	35	1	13	21

Most of the teachers participating in this project had more than two years of teaching experience:

Treatment Group

Years teaching	Count (N=43)	Percent
New (1-2 years)	2	5%
Somewhat experienced (3-5 years)	12	28%
Experienced (6-10 years)	11	26%
Veteran (11 years and up)	18	42%

Comparison Group

Years Teaching	Count (N=35)	Percent
New (1-2 years)	1	3%
Somewhat experienced (3-5 years)	8	23%

Experienced (6-10 years)	15	43%
Veteran (11 years and up)	11	31%

Question: Will including non-lead treatment teachers bias the test results?

Of the forty-three treatment group teachers, five were not lead teachers: two were co-teachers, two were supporting teachers, and one is a technology teacher.

A T-test was used to determine if there would be difference in the group pre-test mean scores if these five non-lead treatment teachers are included. The mean scores are similar when the five non-lead teachers are included (M=16.12, SD=3.033) and when they are not included (M=15.68, SD=2.914). The difference was not significant statistically ($t(79) = .652, p = .517 > .05$). In other words, non-lead teachers can be included in the treatment group analysis without affecting the test results.

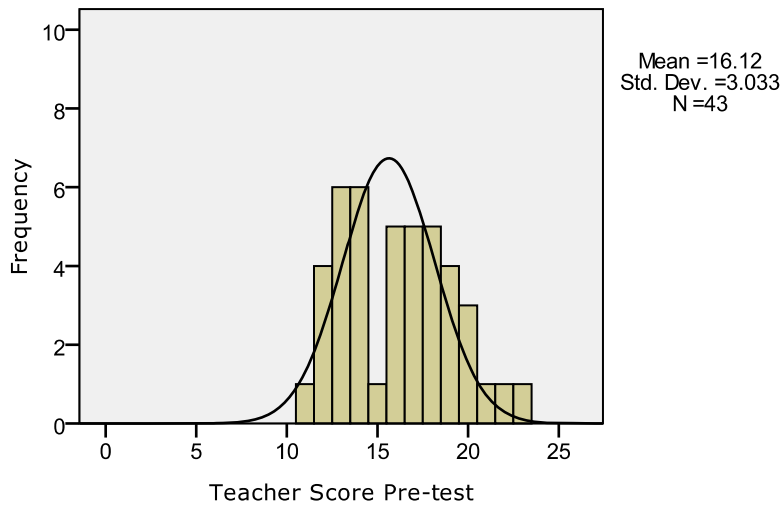
Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Teacher Score Pre-test	with nonlead	43	16.12	3.033	.463
	without nonlead	38	15.68	2.914	.473

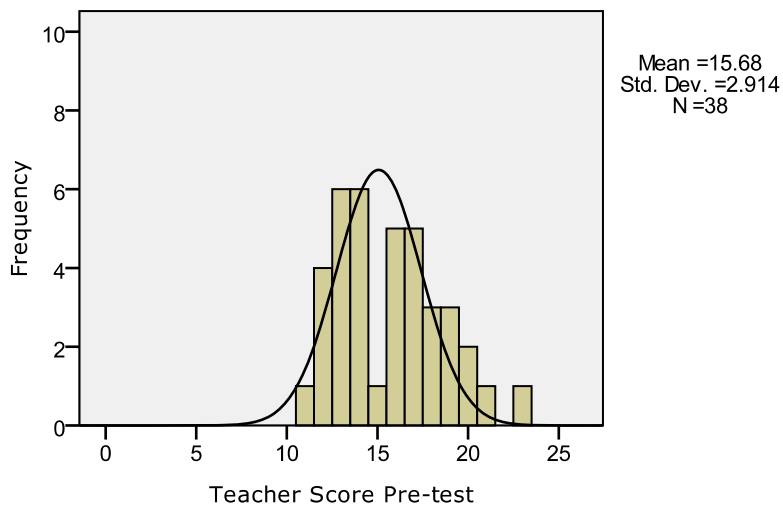
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Teacher Score Pre-test	Equal variances assumed	.090	.765	.652	79	.517	.432	.663	-.888	1.752
	Equal variances not assumed			.653	78.431	.515	.432	.661	-.885	1.749

Group: with nonlead



Group: without nonlead



Question: Did the treatment group and comparison group have the same baseline knowledge?

To examine whether the treatment group and comparison group teachers had the same baseline knowledge, a T-test was used to compare their mean pre-test scores. No significant difference was found between the two group's mean scores ($t(76)=1.316$, $p=.192>.05$), although the treatment group had a slightly higher mean score ($M=16.12$, $SD=3.033$) than the comparison group ($M=15.17$, $SD=3.294$).

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Teacher Score Pre-test	treatment	43	16.12	3.033	.463
	comparison	35	15.17	3.294	.557

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Teacher Score Pre-test	Equal variances assumed	.028	.867	1.316	76	.192	.945	.718	-.485	2.374
	Equal variances not assumed			1.305	70.095	.196	.945	.724	-.499	2.389

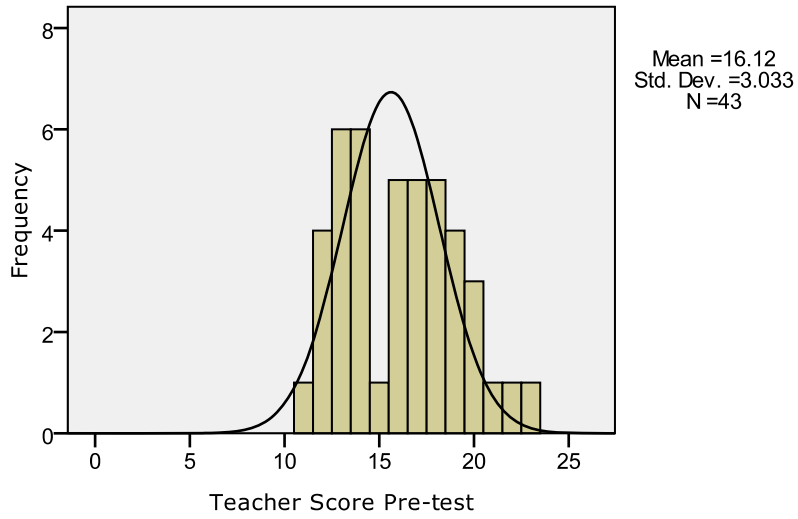
Teachers' pre-test score distribution

The treatment group teachers had a minimum pre-test score of 11 out of 26 and a maximum score of 23. The comparison group teachers had a minimum score of 4 and a maximum score of 21.

Teacher Pre-test Score

Group	Minimum Score	Maximum Score
Treatment	11	23
Comparison	4	21

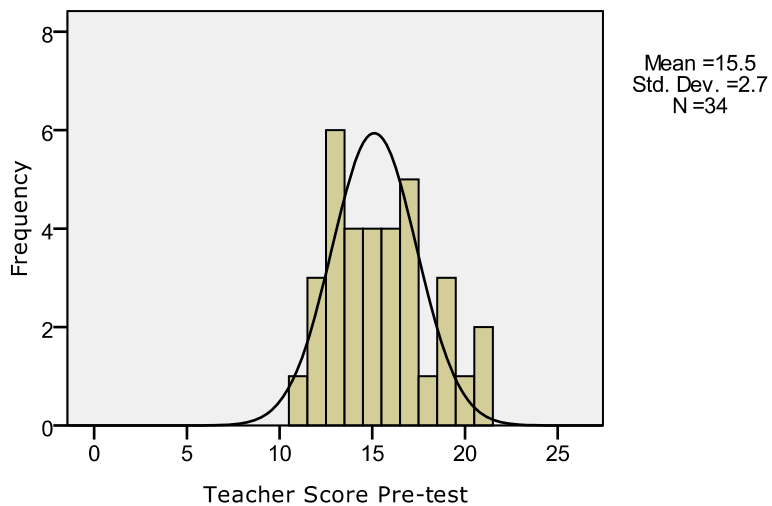
Group: experimental



Treatment Group

Score (max=26)	Count (n=43)	Percent
11	1	2.3%
12	4	9.3%
13	6	14.0%
14	6	14.0%
15	1	2.3%
16	5	11.6%
17	5	11.6%
18	5	11.6%
19	4	9.3%
20	3	7.0%
21	1	2.3%
22	1	2.3%
23	1	2.3%

Group: comparison



Comparison Group

Score (max=26)	Count (n=35)	Percent
4	1	2.9%
11	1	2.9%
12	3	8.6%
13	6	17.1%
14	4	11.4%
15	4	11.4%
16	4	11.4%
17	5	14.3%
18	1	2.9%
19	3	8.6%
20	1	2.9%
21	2	5.7%

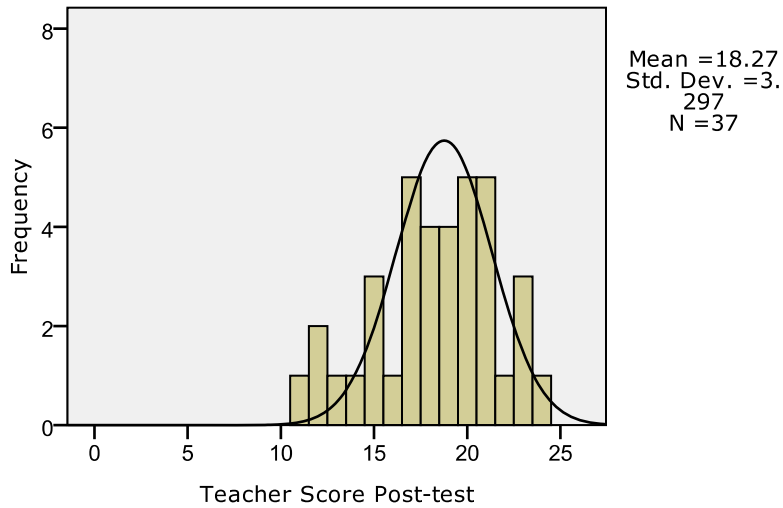
Teachers' post-test score distribution

The treatment group teachers had a minimum pre-test score of 11 out of 26 and a maximum score of 24. The comparison group teachers had a minimum score of 8 and a maximum score of 21.

Teacher Post-test Score

Group	Minimum Score	Maximum Score
Treatment	11	24
Comparison	8	21

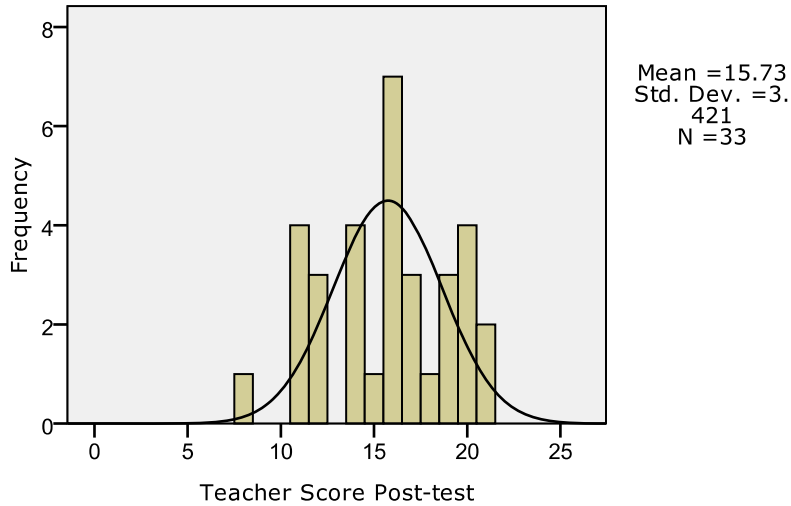
Group: Experimental Group



Treatment Group

Score (max=26)	Count (n=37)	Percent
11	1	2.7%
12	2	5.4%
13	1	2.7%
14	1	2.7%
15	3	8.1%
16	1	2.7%
17	5	13.5%
18	4	10.8%
19	4	10.8%
20	5	13.5%
21	5	13.5%
22	1	2.7%
23	3	8.1%
24	1	2.7%

Group: Comparison Group



Comparison Group

Score (max=26)	Count (n=33)	Percent
8	1	3.0%
11	4	12.1%
12	3	9.1%
14	4	12.1%
15	1	3.0%
16	7	21.2%
17	3	9.1%
18	1	3.0%
19	3	9.1%
20	4	12.1%
21	2	6.1%

Question: Are the matched-pairs a good match?

Thirty-six treatment group teachers and thirty-five comparison group teachers were matched by the grade level they taught. One of the comparison teachers was matched with two treatment group teachers, resulting in thirty-six matched pairs of teachers.

An ideal match would be when the difference of raw pre-test scores (number correct) between the treatment group teacher and the comparison group teacher is zero, or close to zero. However, only 36 percent of the pairs had a difference between 0 and 1 in their pre-test scores. Therefore, the analysis that follows will not look at matched pairs of teachers but will look at treatment and comparison teachers as two groups.

Raw Pre-test Score Difference

Treatment Group – Comparison Group	Count (n=36)
-7	3
-6	1
-3	3
-2	4
-1	5
0	4
1	4
2	1
3	2
4	3
5	2
6	1
7	1
8	1
10	1

Question: Did the matched groups of teachers have the same baseline knowledge?

To examine whether the matched groups of treatment and comparison teachers had the same baseline knowledge, a T-test was used to compare their pre-test mean scores. No significant difference was found between the two group's mean scores ($t(70) = .595$, $p = .554 > .05$), with the treatment group having a mean score ($M = 15.75$, $SD = 2.98$) very close to that of the comparison group ($M = 15.31$, $SD = 3.35$).

	Group	N	Mean	Std. Deviation	Std. Error Mean
Teacher Score Pre-test	treatment	36	15.75	2.980	.497
	comparison	36	15.31	3.345	.558

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper

Teacher Score Pre-test	Equal variances assumed	.010	.919	.595	70	.554	.444	.747	-1.045	1.934
	Equal variances not assumed			.595	69.084	.554	.444	.747	-1.045	1.934

Question: How much did each group improve its test scores?

The pre-test for the treatment group teachers was administered at the beginning of summer workshop; the pre-test for the comparison group teachers was administered at the beginning of the fall semester. Both groups had their post-tests administered at the end of the spring semester 2009. During the semester, some teachers left the program, while others did not return both tests. The total number of matched-tests is 37 for the treatment group and 33 for the comparison group.

	Pre-test	Post-test	Matched-tests
Treatment Group	43	37	37
Comparison Group	35	33	33

When looking at each group separately, the treatment group had a significant increase in their test scores, from 16.11 to 18.27 points ($t(36)=-3.991, p<.01$).

	Mean	N	Std. Deviation	Std. Error Mean
Teacher Score Pre-test	16.11	37	2.706	.445
Teacher Score Post-test	18.27	37	3.297	.542

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Teacher Score Pre-test - Teacher Score Post-test	-2.162	3.296	.542	-3.261	-1.063	3.991	36	.000

The comparison group had virtually no increase:

	Mean	N	Std. Deviation	Std. Error Mean
Teacher Score Pre-test	15.33	33	3.323	.578
Teacher Score Post-test	15.73	33	3.421	.596

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Teacher Score Pre-test - Teacher Score Post-test	-.394	3.201	.557	-1.529	.741	.707	32	.485

Question: Did the treatment group improve significantly more than the comparison group?

Although it would appear from the above that the treatment group improved significantly more than the comparison group, an analysis of covariance (ANCOVA) was used in order to control for differences in pre-test scores. When their pre-test scores are used as a covariate, these are a significant predictor of post-test scores ($F(1,66) = 19.321, p < .01$) for both groups.

The interaction effect between teachers' pre-test scores and Group variable (treatment or comparison) was not significant ($F(1,66) = .073, p = .788 > .05$). Therefore, the ANCOVA without the interaction term was used for this analysis.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	292.035 ^a	3	97.345	10.952	.000	.332	32.857	.999
Intercept	174.420	1	174.420	19.624	.000	.229	19.624	.992
TeacherScorepre	171.721	1	171.721	19.321	.000	.226	19.321	.991
Group	5.849	1	5.849	.658	.420	.010	.658	.126
Group * TeacherScorepre	.648	1	.648	.073	.788	.001	.073	.058
Error	586.608	66	8.888					
Total	21279.000	70						
Corrected Total	878.643	69						

a. R Squared = .332 (Adjusted R Squared = .302)

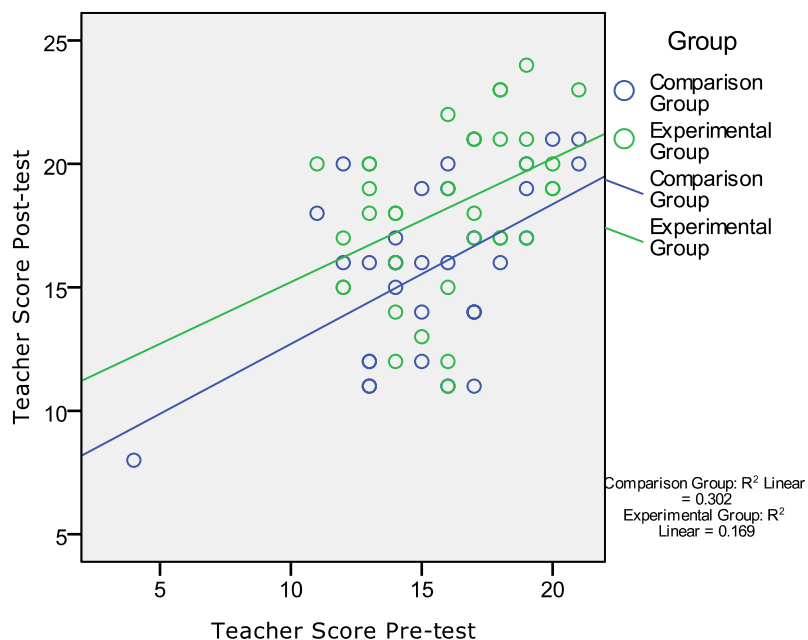
b. Computed using alpha = .05

The ANCOVA without the interaction component (Group*TeacherScorePre) showed that the difference in post-test scores between the two groups was significant ($F(1,67) = 8.846, p < .01$) when the pre-test scores are held constant. In other words, the treatment teachers' post-tests scores improved significantly even when their slightly higher pre-test scores are taken into account:

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	291.387 ^a	2	145.693	16.622	.000	.332	33.244	1.000
Intercept	175.775	1	175.775	20.054	.000	.230	20.054	.993
TeacherScorepre	178.586	1	178.586	20.375	.000	.233	20.375	.994
Group	77.534	1	77.534	8.846	.004	.117	8.846	.834
Error	587.256	67	8.765					
Total	21279.000	70						
Corrected Total	878.643	69						

a. R Squared = .332 (Adjusted R Squared = .312)

b. Computed using alpha = .05



Put another way, before teachers' pre-test scores were held constant (in ANCOVA), the treatment group teachers had higher post-test scores (M=18.27, SD=3.297) than the comparison group teachers (M=15.73, SD=3.421).

Group	Mean	Std. Deviation	N
treatment	18.27	3.297	37
comparison	15.73	3.421	33

When the teachers' pre-test scores are held constant, the treatment teachers still had higher post-test scores (M=18.074) than the comparison teachers (M=15.948) and the difference in the post-test scores between the two groups was significant.

Estimated Marginal Means

Group	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Treatment Group	18.074 ^a	.489	17.098	19.049
Comparison Group	15.948 ^a	.518	14.914	16.981

a. Covariates appearing in the model are evaluated at the following values: Teacher Score Pre-test = 15.74.

Student Test Analysis

1425 students from 73 classes took the pre-test. One of the classes is not included in the analysis since it was a special education class with a class size of four. Therefore, 1421 students from 72 classes comprise the dataset for Year 2 student pre-test analysis. There were 737 students in treatment group from 37 classes and 684 students in comparison group from 35 classes.

Group	Students (N=1421)	Classes (N=72)
Treatment	737	37
Comparison	684	35

Question: Did the treatment group and comparison group have the same baseline knowledge?

A T-test was used to determine if there was any difference in the pre-test scores between the treatment group and comparison group. The treatment group's pre-test scores (M=9.01, SD=3.271) were slightly higher than the comparison group's scores (M=8.84, SD=2.95), but the difference is not significant statistically ($t(1417.883) = .989, p = .323 > .05$).

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Student Score Pre-test	treatment	737	9.01	3.271	.120
	comparison	684	8.84	2.951	.113

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Student Score Pre-test	Equal variances assumed	7.509	.006	.986	1419	.324	.163	.166	-.162	.488
	Equal variances not assumed			.989	1417.883	.323	.163	.165	-.160	.487

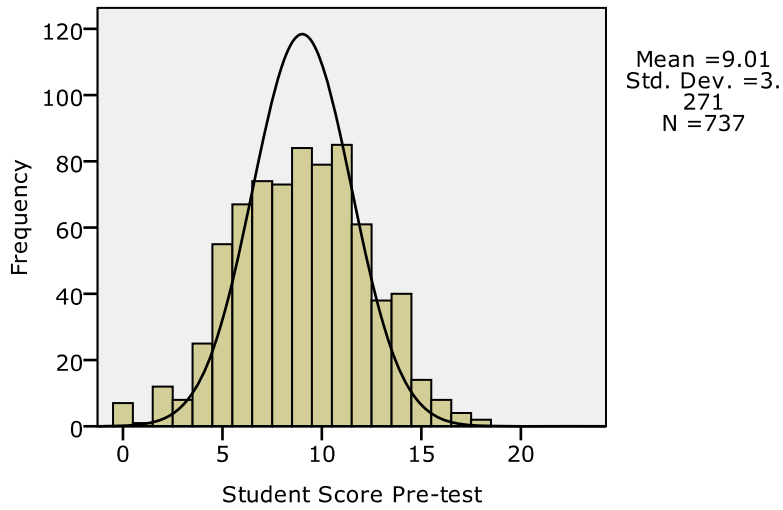
Students' pre-test score distribution

The treatment group had a minimum pre-test score of 0 out of 23 and a maximum score of 18. The comparison group had a minimum score of 0 and a maximum score of 17.

Student Pre-test Score

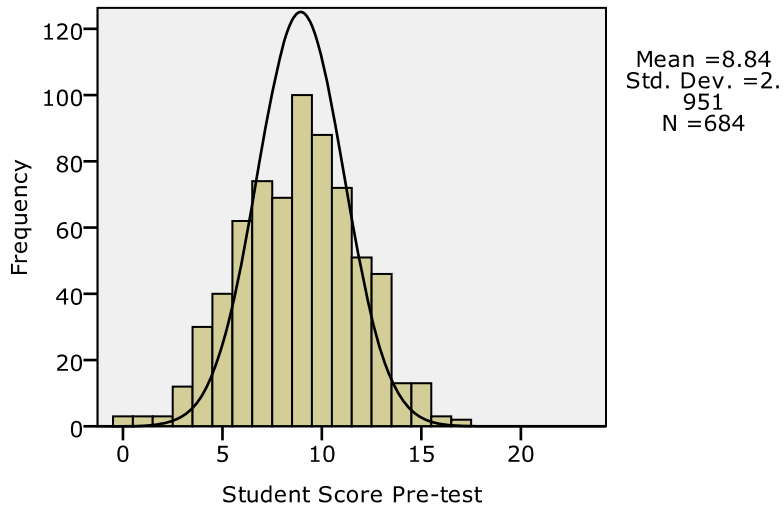
Group	Minimum Score	Maximum Score
Treatment	0	18
Comparison	0	17

Group: experimental



Pre-test Score (Max=23)	Count (n=737)	Percent
0	7	.9%
1	1	.1%
2	12	1.6%
3	8	1.1%
4	25	3.4%
5	55	7.5%
6	67	9.1%
7	74	10.0%
8	73	9.9%
9	84	11.4%
10	79	10.7%
11	85	11.5%
12	61	8.3%
13	38	5.2%
14	40	5.4%
15	14	1.9%
16	8	1.1%
17	4	.5%
18	2	.3%

Group: comparison



Pre-test Score (Max=23)	Count (n=684)	Percent
0	3	.4%
1	3	.4%
2	3	.4%
3	12	1.8%
4	30	4.4%
5	40	5.8%
6	62	9.1%
7	74	10.8%
8	69	10.1%
9	100	14.6%
10	88	12.9%
11	72	10.5%
12	51	7.5%
13	46	6.7%
14	13	1.9%
15	13	1.9%
16	3	.4%
17	2	.3%

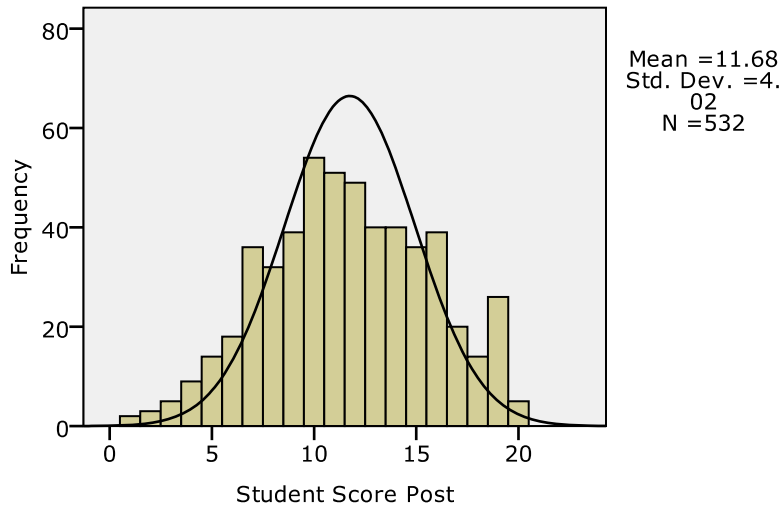
Students’ post-test score distribution

The treatment group of students had a minimum post-test score of 0 out of 23 and a maximum score of 18. The comparison group of students had a minimum score of 0 and a maximum score of 17.

Student Post-test Score

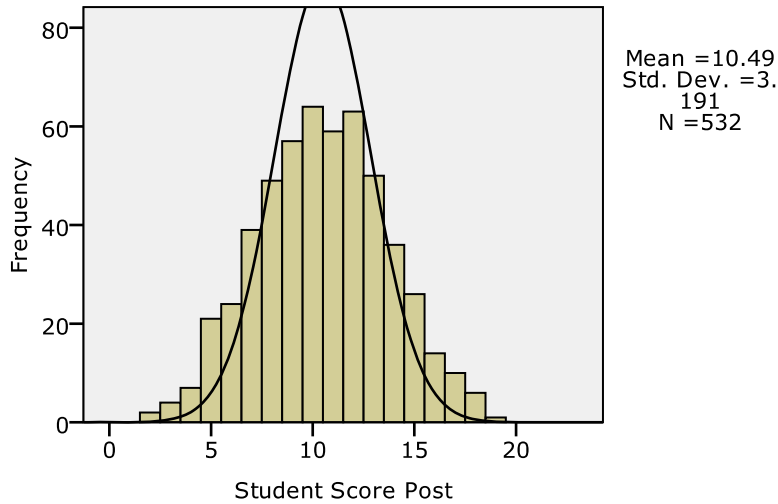
Group	Minimum Score	Maximum Score
Treatment	1	20
Comparison	2	19

Group: Experimental Group



Post-test Score (Max=23)	Count (n=532)	Percent
1	2	.4%
2	3	.6%
3	5	.9%
4	9	1.7%
5	14	2.6%
6	18	3.4%
7	36	6.8%
8	32	6.0%
9	39	7.3%
10	54	10.2%
11	51	9.6%
12	49	9.2%
13	40	7.5%
14	40	7.5%
15	36	6.8%
16	39	7.3%
17	20	3.8%
18	14	2.6%
19	26	4.9%
20	5	.9%

Group: Comparison Group



Post-test Score (Max=23)	Count (n=532)	Percent
2	2	.4%
3	4	.8%
4	7	1.3%
5	21	3.9%
6	24	4.5%
7	39	7.3%
8	49	9.2%
9	57	10.7%
10	64	12.0%
11	59	11.1%
12	63	11.8%
13	50	9.4%
14	36	6.8%
15	26	4.9%
16	14	2.6%
17	10	1.9%
18	6	1.1%
19	1	.2%

Question: How much did each group improve its test scores?

The pre-test was administered to all students at the beginning the fall semester and the post-test was administered at the end of the spring semester (2009). During the school year, some teachers left the program while others did not return the student tests. The total number of student tests that could be matched (pre with post) was

532 in the treatment group and 532 in the comparison group (the equal numbers are pure chance).

	Pre-test	Matched-tests
Treatment Group	737	532
Comparison Group	684	532

When looking at each group separately, the treatment group had a significant increase in its post-test scores (from $M=9.19$ to $M=11.68$, $t(531)=-15.638$, $p<.01$). This was a 27 percent increase.

	Mean	N	Std. Deviation	Std. Error Mean
Student Score Pre	9.19	532	3.201	.139
Student Score Post	11.68	532	4.020	.174

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Student Score Pre - Student Score Post	-2.489	3.671	.159	-2.801	-2.176	15.638	531	.000

Although the comparison group had a significant increase in their post-test scores as well (from $M=9.01$ to $M=10.49$, $t(531)=-10.488$, $p<.01$), the increase was only 16 percent.

	Mean	N	Std. Deviation	Std. Error Mean
Student Score Pre	9.01	532	2.932	.127
Student Score Post	10.49	532	3.191	.138

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Student Score Pre - Student Score Post	-1.483	3.262	.141	-1.761	-1.205	10.488	531	.000

Question: Did the treatment group improve significantly more than the comparison group?

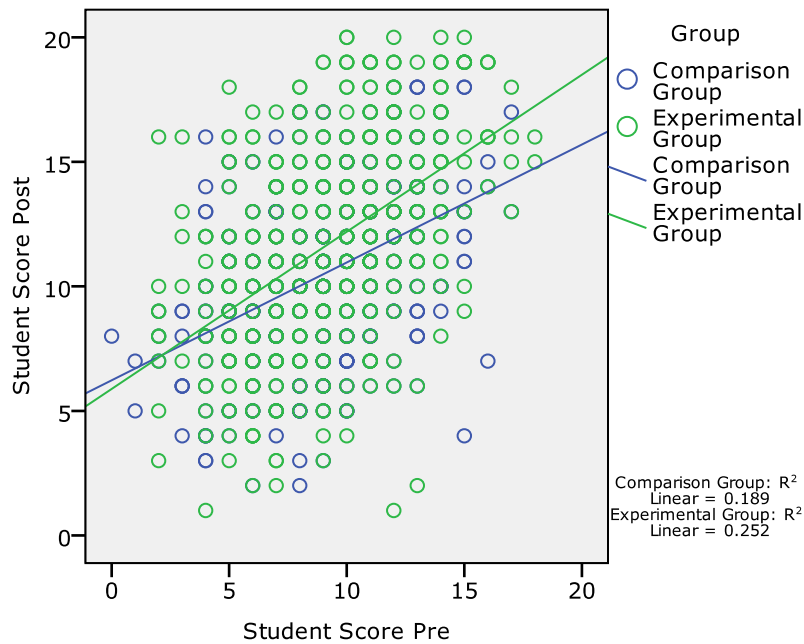
Although it would appear from the above that the treatment group improved significantly more than the comparison group, an analysis of covariance (ANCOVA) was used in order to control for differences in pre-test scores. When pre-test scores were used as a covariate, they were a significant predictor of post-test scores ($F(1,1060)=297.287$, $p<.01$). However, the interaction effect between the students' pre-test scores and Group variable (treatment or comparison) was significant ($F(1,1060)=6.037$, $p=.014<.05$), with the treatment having a greater positive impact on those students who had better pre-test scores.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power ^b
--------	-------------------------	----	-------------	---	------	--------------------	-----------------------------

Corrected Model	3567.202 ^a	3	1189.067	116.739	.000	350.217	1.000
Intercept	3955.244	1	3955.244	388.313	.000	388.313	1.000
StudentScorePre	3028.074	1	3028.074	297.287	.000	297.287	1.000
Group	3.180	1	3.180	.312	.576	.312	.086
Group * StudentScorePre	61.491	1	61.491	6.037	.014	6.037	.690
Error	10796.843	1060	10.186				
Total	145140.000	1064					
Corrected Total	14364.045	1063					

a. R Squared = .248 (Adjusted R Squared = .246)

b. Computed using alpha = .05



Before students' pre-test scores were held constant (in ANCOVA), the treatment group had higher post-test scores (M=11.68, SD=4.02) than the comparison group (M=10.49, SD=3.19).

Group	Mean	Std. Deviation	N
treatment	11.68	4.02	532
comparison	10.49	3.19	532

After students' pre-test scores were held constant statistically, the treatment students still had higher post-test scores (M=11.624) than the comparison students (M=10.535).

Estimated Marginal Means

Group	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Treatment Group	11.624 ^a	.138	11.352	11.895
Comparison Group	10.535 ^a	.138	10.263	10.806

Group	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Treatment Group	11.624 ^a	.138	11.352	11.895
Comparison Group	10.535 ^a	.138	10.263	10.806

a. Covariates appearing in the model are evaluated at the following values: Student Score Pre = 9.10.

Question: Was a teacher's post-test score a significant predictor of his/her students' post-test scores?

When the teachers' post-test scores were added as another covariate, they were a significant predictor of their students' post-test scores ($F(1,1058)=24.867, p<.01$).

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	3827.057 ^a	5	765.411	76.854	.000	.266	384.268	1.000
Intercept	303.577	1	303.577	30.482	.000	.028	30.482	1.000
Group	27.328	1	27.328	2.744	.098	.003	2.744	.380
StudentScorePre	3000.622	1	3000.622	301.287	.000	.222	301.287	1.000
TeacherScorePost	247.664	1	247.664	24.867	.000	.023	24.867	.999
Group * StudentScorePre	40.837	1	40.837	4.100	.043	.004	4.100	.525
Group * TeacherScorePost	20.288	1	20.288	2.037	.154	.002	2.037	.297
Error	10536.988	1058	9.959					
Total	145140.000	1064						
Corrected Total	14364.045	1063						

a. R Squared = .266 (Adjusted R Squared = .263)

b. Computed using alpha = .05

However, this correlation was statistically significant for the treatment group ($r=.186, p=.000$), but not for the comparison group ($r=.059, p=.173$). This was not the case in the previous year and suggests that the test itself is better tied to the content being taught.

		Student Score Post	Teacher Score Post-test
Student Score Post	Pearson Correlation	1	.186**
	Sig. (2-tailed)		.000
	N	532	532
Teacher Score Post-test	Pearson Correlation	.186**	1
	Sig. (2-tailed)	.000	
	N	532	532

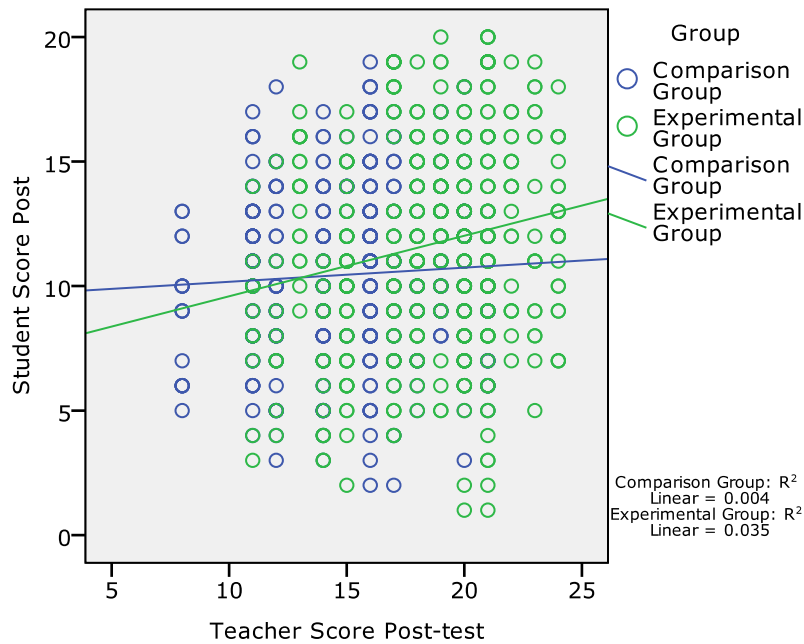
** . Correlation is significant at the 0.01 level (2-tailed).

a. Group = Treatment Group

		Student Score Post	Teacher Score Post-test
Student Score Post	Pearson Correlation	1	.059

	Sig. (2-tailed)		.173
	N	532	532
Teacher Score Post-test	Pearson Correlation	.059	1
	Sig. (2-tailed)	.173	
	N	532	532

a. Group = Comparison Group



Year 1 and Year 2 comparison

Teacher Test

	Year 1	Year 2
Did the treatment group teachers have significant increase in their test scores?	Yes	Yes
Did the comparison group teachers have significant increase in their test scores?	No	No
Were the teachers' pre-test scores a significant predictor for their post-test scores?	Yes	Yes
Did the treatment group teachers perform significantly better than the comparison group teachers?	Yes	Yes

Student Test

	Year 1	Year 2
Did the treatment group students have significant increase in their test scores?	Yes	Yes
Did the comparison group students have significant increase in their test scores?	Yes	Yes

Were the students' pre-test scores a significant predictor for their post-test scores?	Yes	Yes
Did the treatment group students perform significantly better than the comparison group students?	Yes	For most but not all students
Were the teachers' post-test scores a significant predictor of their students' post-test scores?	No	Only in the treatment group