CENTER FOR TECHNOLOGY IN LEARNING

Technology-Enhanced Elementary and Middle School Science II (TEEMSS II)

Research Report 1

Prepared August 15, 2006

Patty A. Kreikemeier
Larry Gallagher
William R. Penuel
Reina Fujii
Veronica Wheaton
Marianne Bakia

P15811.200

CONTENTS

Executive Summary

Technology-Enhanced Elementary and Middle School Science II (TEEMSS II) is a National Science Foundation-funded project (Grant IMD-0352522) whose goal is to bring the power of information and communication technology to science education in grades 3 through 8. The project staff at the Concord Consortium (CC) is creating instructional materials that address important, standards-based science content and that aim to be easily and inexpensively integrated into any science program. The learning strategy is based on student investigations of real phenomena using sensors and of virtual environments using computer models.

Researchers at the Center for Technology in Learning at SRI International (SRI) have prepared this evaluation report on the outcomes of student learning assessments administered to students both in classrooms implementing TEEMSS and in non-implementing classrooms during the 2004-05 school year. The evaluation relied on a quasi-experimental design with nonequivalent groups. Teachers who volunteered to be part of the study agreed to use TEEMSS curriculum in their science classes. Some teachers agreed to implement TEEMSS units relevant to the grade level they taught during 2004-05 (Group 1). Others agreed to delay implementation until 2005-06 (Group 2). Group 2 students served as a comparison group for Group 1 students. A total of 17 Group 1 and 25 Group 2 teachers completed both surveys and tests. The number of teachers teaching a given topic ranged from 0 to 12. The number of students in a particular group who took pre- and post-tests ranged from 0 to 251.

For 12 TEEMSS units, SRI researchers and CC staff collaboratively developed tests that aligned closely with the content of the units. The team drew primarily from released items developed for national or state standardized tests, as well as developing additional items to supplement the released items. The tests for which data in this report were analyzed covered the following eight topics included in TEEMSS units: Sound, Electricity, Human and Electronic Sensing, Temperature, Levers and Machines, Monitoring a Living Plant, Pressure, and Motion. An outside science expert reviewed all items for scientific accuracy. SRI researchers partnered with CC staff to field-test all items, subsequently eliminating items that had been shown to be likely to be too easy before administering the tests in the evaluation study.

Teachers selected the TEEMSS units to implement. Both Group 1 and Group 2 teachers administered pre- and post-tests to students in their classrooms for the units taught. Evaluators analyzed the test results in four ways: (1) by measuring gains from pretest to posttest for each unit by group; (2) by comparing gains of Group 1 and Group 2 students for unit tests in which students from both groups took tests; (3) by modeling effects of a group on results by accounting for student clustering in classrooms; and (4) by calculating effect sizes for unit tests in which students from both groups took tests. The first analysis focused on within-group gains, measuring improvement from pre- to post-test; the other analyses focused on comparisons between groups with respect to the significance and magnitude of gains. Analysis 3 corrected for an important limitation in the other analyses—namely, that they did not consider how classrooms contribute to variation in scores.

Results of the first analysis indicated that for all of the tests that Group 1 students took—Sound, Sensing, Temperature, Pressure, and Motion—students on average made statistically significant gains from pretest to posttest. In contrast, Group 2 students made significant gains

on half of the eight tests they took—Sound, Electricity, Levers, and Monitoring a Living Plant. Of the tests taken by both groups, Group 1 students made statistically significant gains on all five tests, whereas Group 2 showed statistically significant gains on three tests.

The second and third analyses focused on the five tests that both Group 1 and Group 2 took by comparing the scores of both groups. Both analyses found that for the Temperature unit test Group 1 students made gains that were significantly greater ($p < .05$) than gains for Group 2 students. For the other four tests, the gains made were not significantly different from one another across groups. Regression analysis that controlled for teacher effects found a similar pattern.

The fourth analysis showed positive and significant effects for the Temperature test, meaning that the confidence interval did not include zero.

Table ES-1 summarizes the results for the five unit tests that Group 1 students took and compares them with Group 2 Scores.

Table ES-1. Group 1 Test Results from Analyses of TEEMSS

	Gains from Pretest to Posttest	Comparison of Gain Scores with Group 2	Regression Analysis Comparing Scores with Group 2	Effect Size Analysis
Sound	Positive and statistically significant	No difference from Group 2	No difference from Group 2	Nonsignificant
Sensing	Positive and statistically significant	No difference from Group 2	No difference from Group 2	Nonsignificant
Temperature	Positive and statistically significant	Significantly greater gains than for Group 2	Significantly higher scores than for Group 2	+0.33 (significant)
Pressure	Positive and statistically significant	No difference from Group 2	No difference from Group 2	Nonsignificant
Motion	Positive and statistically significant	No difference from Group 2	No difference from Group 2	Nonsignificant

Overall, these results point to two conclusions:

1. TEEMSS students improved on all five unit tests they took from pre- to post-test—Sound, Human and Electronic Sensing, Water and Air Temperature, Pressure, and Motion.

2. The test for Temperature indicates that, for this group of teachers, students in Group 1 classrooms outperformed students in the comparison Group 2 classrooms.

Two important caveats apply to interpreting these results. First, teachers were not randomly assigned to condition. Thus, selection bias could have affected the results, and several meta-analytic studies do suggest that experimental studies give more accurate estimates of impact than do quasi-experimental studies (Glazerman, Levy, & Myers, 2003). Second, in controlling for effects of clustering in our third analysis, we increased the power of the study to detect significant effects but doing so limited its generalizability to the teachers in the sample. Thus, for more generalizable results, a cluster randomized trial with greater numbers of teachers implementing each unit would be necessary.

Context for the Evaluation

Technology-Enhanced Elementary and Middle School Science II (TEEMSS II) is a National Science Foundation-funded project (Grant IMD-0352522) whose goal is to bring the power of information and communication technology to science education in grades 3 through 8. The project staff at the Concord Consortium (CC) is creating instructional materials that address important, standards-based science content and that aim to be easily and inexpensively integrated into any science program. The learning strategy is based on student investigations of real phenomena using sensors and of virtual environments using computer models.

A critical aspect of the TEEMSS II project is its focus on the development of integrated *units* of study aligned to standards. In the past, powerful and innovative educational technologies of the kind that are integrated into the TEEMSS units (e.g., probes, sensors, computers) have not always been widely adopted, in part because technology has not been well-integrated into curriculum materials (Blumenfeld et al., 2000). In contrast, the TEEMSS materials are integrating technology into a framework of science inquiry and into specific units of study that are aligned with the *National Science Education Standards* (NRC, 1996).

TEEMSS II is producing 15 two-week units, with five units each for Grades 3-4, 5-6, and 7-8. Each unit consists of 2 one-week investigations and employs technology both as a support to learning and as a support to recording and analyzing student work. Table 1 shows the different units being developed and their alignment with science domains that are part of the *National Science Education Standards*. Metcalf (2006) presents more details about the program design

Table 1. Units by Grade and Science Domain

Domain	Grades 3-4	Grades 5-6	Grades 7-8
Inquiry	Sound	Water and Air Temperature	Air Pressure
Physical Science	Electricity	Levers and Machines	Motion
Life Science	Sensing	Monitoring a Living Plant	Adaptation
Earth and Space Science	Weather	Sun, Earth, Seasons	Water Cycle
Technology/Engineering	Design a Playground	Design a Greenhouse	Design a Measurement

Adapted from Metcalf (2006)

Studies undertaken as part of the TEEMSS pilot project indicated the potential of the approach, which was accordingly incorporated in the current TEEMSS II project (Metcalf & Tinker, 2003). However, as new units were developed, new assessment tools for measuring student learning were needed. In addition, the TEEMSS II project needed evaluation studies to investigate whether the expanded set of units would produce statistically significant and measurable effects on student learning.

As part of its proposal to the National Science Foundation, CC approached researchers at the Center for Technology in Learning (CTL) at SRI International (SRI) about developing

assessments for the project and assisting with the evaluation. SRI's work has included development of 12 unit tests in collaboration with the CC and analysis of student learning gains as part of the first full year of implementing the units with teachers and students. This evaluation report describes how SRI researchers undertook both activities, and it presents results of our analysis of learning gains.

Evaluation Questions

The primary evaluation questions addressed in this report through our analyses of student learning gains are:

> *Do TEEMSS students' scores on unit assessments improve from pretest to posttest?*
>
> *Compared with students in a comparison group, do changes in TEEMSS students' scores differ from pretest to posttest?*

We hypothesized that TEEMSS students would show statistically significant gains from pretest to posttest in all units and that those gains would be greater for TEEMSS students than for students in the comparison group.

Methodology

Sample for the Study

Teachers volunteered to be part of the TEEMSS II project. The project staff assigned some teachers to implement the curriculum units with their students in 2004-05 (Group 1) and asked another group to delay implementation until 2005-06 (Group 2). A primary basis for selecting teachers to be part of Group 1 was the availability of technology in their classrooms. For purposes of this report, then, Group I can be considered to be the "treatment" group and Group 2 the "comparison" group. We do not use the term "control group" here because teachers were not randomly assigned to a condition and may not be equivalent in regard to all variables that may have affected the student results.

For all measured background variables that we analyzed using data from the teacher survey, however, the two groups of teachers were not significantly different statistically from one another. As Table 2 shows, Group 1 and Group 2 teachers did not differ significantly with respect to their experience, pedagogical approach, or comfort with or use of technology. Two variables did approach statistical significance, however: teachers in Group 2 had taught more years at their level than had Grade 1 teachers ($p = .06$), and teachers in Group 1 had taught more years in science ($p = .06$).

Table 2. Characteristics of Teachers in the Study

Domain	*M*	*SD*
Years Teaching		
Group 1 ($n = 17$)	6.9 years	5.4 years
Group 2 ($n = 25$)	11.3 years	9.2 years
Years Teaching Science		
Group 1 ($n = 17$)	6.4 years	4.8 years
Group 2 ($n = 25$)	10.6 years	9.2 years
Years at Current Grade Level		
Group 1 ($n = 17$)	4.9 years	4.8 years
Group 2 ($n = 25$)	7.1 years	6.0 years
Use of Hands-On Activities		
Group 1 ($n = 17$)	3.8 (out of 5)	0.9
Group 2 ($n = 25$)	3.5 (out of 5)	0.9
Frequency of Technology Use		
Group 1 ($n = 16$)	2.9 (out of 5)	1.2
Group 2 ($n = 25$)	2.6 (out of 5)	1.2
Experience with Sensors		
Group 1 ($n = 16$)	25%	
Group 2 ($n = 25$)	12%	
Confidence in Trying Technology		
Group 1 ($n = 16$)	4.4 (out of 5)	0.6
Group 2 ($n = 25$)	4.6 (out of 5)	0.6

M = Mean; *SD* = Standard Deviation. All group differences were non-significant.

*The *n* reported here refers to teachers who completed surveys and tests.

Table 3 shows the number of students in each classroom who completed both pretests and posttests by unit. Note that some teachers implemented more than one unit; therefore, the numbers of teachers below do not match the figures cited in Table 2. For five of the tests—Sound, Sensing, Temperature, Pressure, and Motion—both Group 1 and Group 2 students completed tests, allowing for comparison of group gain scores.

Table 3. Number of Student and Teacher Participants by Unit

	TEEMSS unit [a]	Number of teacher participants (number of students) used in data analysis	
		Group 1	Group 2
Grades 3-4	1 Sound	2 (38)	10 (154)
	2 Electricity	0 (0)	12 (185)
	3 Human and Electronic Sensing	7 (126)	1 (35)
	4 Weather		
	5 Design a Playground		
Grades 5-6	6 Water and Air Temperature	5 (253)	4(149)
	7 Levers and Machines	0 (0)	6 (120)
	8 Monitoring a Living Plant	0 (0)	6 (193)
	9 Seasons		
	10 Design a Greenhouse		
Grades 7-8	11 Pressure	1 (30)	2 (42)
	12 Understanding Motion	3 (245)	2 (44)
	13 Evolution	0 (0)	0 (0)
	14 The Water Cycle		
	15 Design a Measurement	Evaluation via portfolio in Year 3	

[a]Curricula for units 4 Weather, 9 Seasons, and 14 The Water Cycle will not be ready until Year 3; evaluation for units 5 Design a Playground, 10 Design a Greenhouse, and 15 Design a Measurement will be undertaken via portfolio in Year 3.

Measures

The research study relied on two primary sources of data: (1) a teacher survey that gathered information on teachers' backgrounds, and (2) unit tests. Given the report's focus on the unit test results, this section discusses the nature of these assessments and how they were developed.

Identification of Items

The TEEMSS assessment development task consisted of collaborative development by the curriculum developers and researchers at CC and the assessment developers and evaluation researchers at CTL.

After the CC curriculum team completed the initial content description, the CTL assessment team culled content-possible items from released assessment sources (see Appendix A). More than 1,500 items were identified as potentially relevant as judged by broad concept mapping to the initial TEEMSS II curriculum unit content descriptions.

Alignment of Items with TEEMSS II Curriculum

The more than 1,500 multiple-choice, short constructed response and extended constructed response items identified as potentially relevant were then classified into categories more closely associated with the TEEMSS curriculum units (e.g., TEEMSS Unit 12 Understanding Motion, does not include concepts such as circular motion and momentum). The remaining 686 items were organized into assessment notebooks, each of which corresponded to a specific TEEMSS curriculum unit. To determine item appropriateness, the TEEMSS curriculum developers then assessed the items in the notebooks for their degree of alignment with the TEEMSS curriculum. Curriculum developers identified approximately 380 items as sufficiently aligned with the content under development and the proposed content for the 15 TEEMSS units (see Appendix B). As content development proceeded, we eliminated some items or revised them so that they aligned more closely with the completed units.

Editing of Items

The items identified by the TEEMSS research and curriculum team for the nine units under current development were further refined. Because the items had been culled from released test items for various grade levels, the CTL assessment team edited items to ensure comprehensibility the appropriate grade level. In addition, because several of the nine TEEMSS units did not have adequate numbers of potential assessment items, CTL and CC collaborated to develop unit-specific questions to supplement those already collected.

Small numbers of students who were not part of the study were asked to "think aloud" while answering test questions in order to identify student comprehension of questions. Modifications were made as appropriate to items on the basis of that student input.

Review of Items for Scientific Accuracy

Dr. George Miller, an outside science expert, reviewed all items for subject-matter accuracy and appropriateness. Dr. Miller serves as science advisor for many K-12 projects, including principal investigator for the California Science Project and assessment director for the FOCUS Math Science Partnership at the University of California, Irvine. In addition, he has worked on assessment projects in California and Kentucky, and served as a science advisor for many other science assessment projects. Items that Dr. Miller rejected were not scored on subsequent tests.

Field Testing

Potential test items and scoring rubrics were field-tested in November 2004, with 60 to 100 non-TEEMSS students completing each unit test. Validity testing was used to select questions that were appropriate for the target grade level and to compare student performance on matched pre- and post-test variations of questions. The items that performed the best across this range of priorities were included on the final tests. Items to which more than 75% of

students responded correctly were eliminated from further consideration in the pre- and post-test TEEMSS assessments.

Pretest and Posttest Forms

The differences between the pretest and posttest forms used in the study were minimal, limited primarily to the order of the answer choices and the presentation of slightly different surface features (e.g., changing values of temperature readings for the prompts on a multiple choice test). Because of these differences, we have analyzed the results by treating the test forms in two ways: (1) as if they are identical (analysis of gains), and (2) as if the pretest were distinct (covariate analysis, in which we control for clustering effects).

We note that several items were included on the final pre- and post-test forms that had not been field-tested or reviewed for scientific accuracy. We included these items at the request of the client. Although these items were included on the tests, they were not scored and are not included in our analyses.

Appendix C presents all pre- and post-test forms. Items that were not scored are indicated with an asterisk (*).

Procedure

A third-party subcontractor for the TEEMSS II project, SuccessLink, facilitated field collection of data by providing teachers with appropriate pretests and posttests and collecting students' completed test data from teachers. Following instructions developed jointly by SuccessLink, CC, and SRI staff, teachers administered all tests to students.

The data collection process entailed several challenges, which became apparent both when data were being collected and when SRI received boxes of test data. First, although teacher participants were asked to choose two of the grade-level-appropriate units and to complete pretests before instruction and posttests after instruction, not all teachers heeded these instructions. Several reported on their teacher survey that they had taught the materials before giving the pretest and then spent only 1 to 2 days using the TEEMSS curriculum before giving the students the posttest.

A second data collection challenge was uneven teacher participation and completion of evaluation materials. In some cases, many more Group 2 than Group 1 teachers implemented a particular unit. For other units, we received no assessments for one group of teachers.

A third data collection challenge was wide variation in the level of understanding of participation requirements as shown in Table 4. Group 1, the teachers scheduled to use the TEEMSS curriculum after completing the online preparatory course, were less likely to satisfactorily complete the materials needed to evaluate their students. Only 16 of the 31 Group 1 teachers completed all data collection segments, compared with 24 of the 26 Group 2 teachers. Examples of noncompliance include submittal of pretests but no posttests, posttests but no pretests, and pretests associated with grade-level-appropriate units and posttests for inappropriate grade level units; and less than strict adherence to permission slip collection and correlation of the slips with returned tests (e.g., submittal of tests whose number exceeded the

number of permission slips without indicating which of the tests the researchers were permitted to use).

Appendix D presents data on teacher participation, by unit.

Table 4. Noncompliance Issues

	Number of teacher participants					
	Materials complete	**No pretests**	**No posttests**	**Pretest - Posttest mismatch**	**Number of tests > number of permission slips**	**No permission slips**
Group 1 ($n = 31$)	16	4	5	6	1	5
Group 2 ($n = 26$)	25	0	1	0	0	0

Changes to minimize these challenges in Year 3 include preparation of anonymous student identification labels to be used uniformly on pretests, posttests, and permission slips.

Scoring Student Work

Scoring of student work conformed to standard practices; namely, each rater scored a single item for all student samples before being trained, raters were then trained to score using anchor and discussion papers, raters were allowed to score actual student work only after scoring qualification samples with 80% reliability, and a minimum of 20% of student-constructed (open-ended) responses were scored by 2 raters whose scores were checked by a third person who resolved any discrepancies in the scores. Before the student data were analyzed, interrater reliability was verified for constructed response items scored by two people. The results indicated that average agreement was 74% across all units on the pretest and 76% across all units on the posttest, somewhat lower than the 80% usually desired.

One problematic item was on the Unit 3 test. The percentage of agreement ranged from 55% on item 8 for the Unit 3 pretest to 76% for the same item for the Unit 3 posttest. The low percentage of agreement suggests that the scoring rubric could not be reliably interpreted and that this item should be eliminated. The results reported here do not include scores from this item. To improve interrater reliability in Year 3, additional training tests (samples of test illustrating a range of student answers) will be identified and additional monitoring of raters will be undertaken.

Since the preparation of these analyses, we have identified two items that may need rescoring. One item from the Monitoring a Living Plant test and one from the Motion test are affected. Preliminary analyses suggest rescoring will not have an effect on the overall results for these tests.

Approach to Analyzing Results

Evaluators analyzed results in four ways: (1) by measuring gains from pretest to posttest for each unit by group; (2) by comparing gains of Group 1 and Group 2 students for unit tests that students from both groups took; (3) by modeling group effects on results when accounting for student clustering in classrooms; and (4) by calculating effect sizes for unit tests taken by

students from both groups. The first analysis focused on within-group gains, measuring improvement from pre- to post-test. We used paired *t*-tests to compare pretest and posttest scores for each group of students' scores to examine the statistical significance of the difference and direction of difference.

The second analysis focused on comparing the statistical significance of the gains made by both groups of students. Both Group 1 and Group 2 students were expected to improve from pretest to posttest, either because they had become familiar with test items or because everyday instruction in the topics would lead to improvement. It was therefore necessary in evaluating the curriculum to determine whether the treatment (Group 1) gains were greater than those that for the comparison group (Group 2) to estimate the effect of the TEEMSS units on student understanding.

The third analysis corrected for a problem that is common in educational research, but not always addressed as part of research design: the effect of clustering of students within classrooms. Our first two analyses assumed that each student's score was independent from those of other students; yet that conclusion is not strictly true. Some variation in a student's scores is likely to be attributable to that student's being a member of a particular classroom. Therefore, student scores are partly determined by being in a classroom with a particular level of implementation of TEEMSS, a specific composition of students, and a teacher of a particular background. Typically, projects like TEEMSS II cannot afford to conduct a cluster-randomized trial large enough to detect significant effects and account for the effect of clustering. An alternative is to treat classrooms as "fixed effects" and model the results as a regression—the approach we took in our third analysis, which treated classrooms in the study as fixed effects and used the pretest as a covariate.

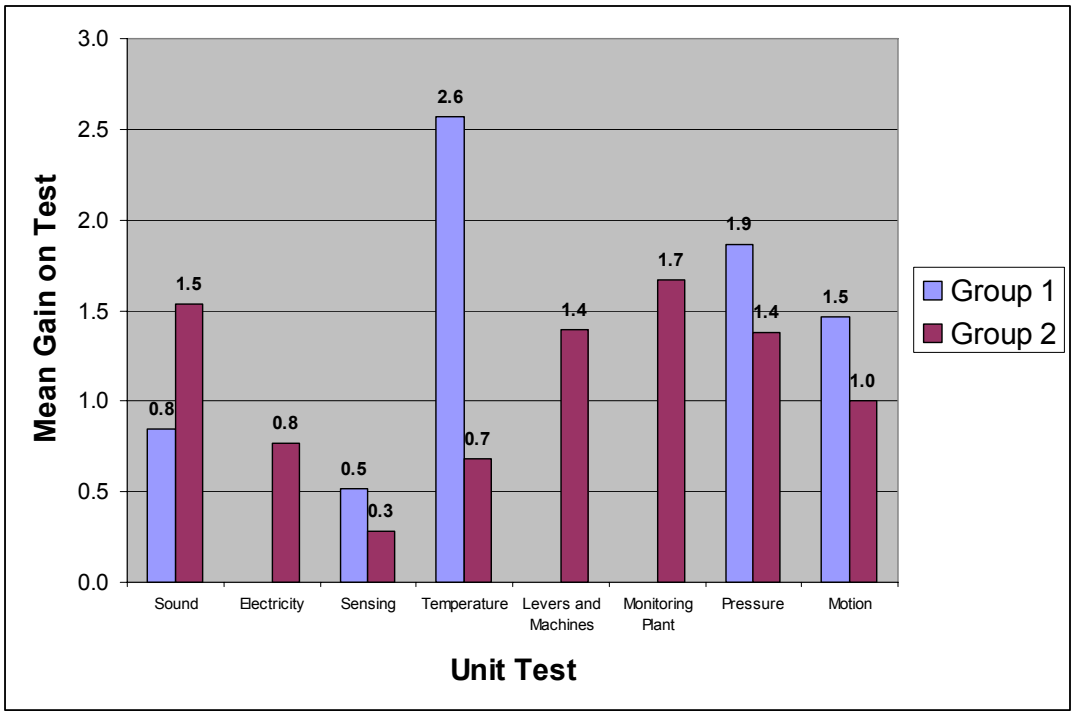
Using the results of the third analysis, we calculated an effect size; that is, a measure of the *magnitude* of gains made by students. When significant, the effect size can also be used to help determine the appropriate sample size for a cluster-randomized trial. We report the magnitude for all tests. Another standard approach consists of considering the 95% confidence intervals when interpreting effect size: if the interval includes zero, the effect is interpreted as statistically nonsignificant or no different from zero.

Results

Within and Between Group Gains (Analyses 1 and 2)

Figure 1 depicts the mean gains students made on each of the tests. In the sections that follow, we present statistical analyses of the gains for each of the tests administered in determining whether the differences in gains are statistically significant.

Figure 1. Gain Scores of Group 1 and Group 2 Students by Test



Results for the Sound Unit Test

The Sound unit test consisted of nine items, with the points for each item distributed as indicated in Table 5. The total number of points possible for this test was 21.

Table 5. Sound Unit Test Item Scoring Algorithm

Item	Format	Possible Points
1	MC*	1
2	CR**	5
3	CR	4
4a	CR	2
4b	CR	2
4c	CR	2
5	CR	3
6	MC	1
7	MC	1
Total Points		21

* multiple-choice item, ** constructed-response item

Both Group 1 and Group 2 students made gains from pretest to posttest on the Sound unit test, and gains for both groups were statistically significant (see Table 6).

Table 6. Results for Sound Unit Test

	Pretest		Posttest		Difference	
	M	***SD***	***M***	***SD***	***M***	***SD***
Group 1 (n = 38)	12.1	2.7	12.9	2.4	+0.8*	2.6
Group 2 (n = 154)	11.9	2.4	13.5	2.4	+1.5**	2.5

* $p < .05$, ** $p < .01$

A *t*-test comparing the gains showed that gains of Group 1 were no different from the gains of Group 2, however (t = -1.5, df = 190, p = .13).

Results for the Electricity Unit Test

Because of concerns about particular items, few items were scored for this test, resulting in limited coverage of the content of the TEEMSS II unit on this topic. However, we did score the items that experts judged to have good construct validity, and results for Group 2 suggest that it was sensitive to instruction. This test had four items, worth a total of 6 points (see Table 7).

Table 7. Electricity Unit Test Item Scoring Algorithm

Item	Format	Possible Points
1	MC	1
2	MC	1
3	MC	1
4	CR	3
Total Points		6

As Table 8 shows, Group 2 made statistically significant gains from pretest to posttest.. No Group 1 classrooms returned completed pre- and post-tests for this unit.

Table 8. Electricity Unit Test Results

	Pretest		Posttest		Difference	
	M	***SD***	***M***	***SD***	***M***	***SD***
Group 2 (n = 185)	2.8	1.01	3.6	1.0	+0.8**	1.4

**$p < .01$

Results for the Sensing Unit Test

The Sensing unit test had seven items scored, with the points for each item distributed as indicated in Table 9. The total number of points possible for this test was 9.

Table 9. Sensing Unit Test Item Scoring Algorithm

Item	Format	Possible Points
1	CR	3
2	MC	1
3	MC	1
4	MC	1
5	MC	1
6	MC	1
7	MC	1
Total Points		9

Group 1 and Group 2 students made gains from pretest to posttest on the Sensing unit test, but only the gains for Group 1 were statistically significant (see Table 10).

Table 10. Sensing Unit Test Results

	Pretest		Posttest		Difference	
	M	***SD***	***M***	***SD***	***M***	***SD***
Group 1 (n = 126)	3.7	1.5	4.2	1.4	+0.5**	2.0
Group 2 (n = 35)	3.8	0.9	4.1	1.4	+0.3	1.7

** $p < .01$

A *t*-test comparing the gains showed that gains of Group 1 were no different from the gains of Group 2, however (t = .63, df = 159, p = .53).

Results for the Temperature Unit Test

The Temperature unit test consisted of seven items, with the points for each item distributed as indicated in Table 11. The total number of points possible for this test was 21.

Table 11. Temperature Unit Test Item Scoring Algorithm

Item	Format	Possible Points
1	MC	1
2	MC	1
3	CR	3
4a	CR	11
4b	CR	3
5	MC	1
6	MC	1
Total Points		21

Group 1 and Group 2 students made gains from pretest to posttest on the Temperature unit test, and the gains for Group 1 were statistically significant (see Table 12).

Table 12. Temperature Unit Test Results

	Pretest		Posttest		Difference	
	M	***SD***	***M***	***SD***	***M***	***SD***
Group 1 (n = 253)	8.2	4.6	10.8	4.1	+2.6**	4.3
Group 2 (n = 149)	7.2	4.3	7.8	3.9	+0.6	4.3

** $p < .01$

A *t*-test comparing the gains showed that the gains of Group 1 were significantly higher than for Group 2 (t = 4.26, df = 400, $p < .001$).

Results for the Levers and Machines Unit Test

The Levers and Machines unit test consisted of five items, with the points for each item distributed as indicated in Table 13. The total number of points possible for this test was 9.

Table 13. Levers and Machines Unit Test Item Scoring Algorithm

Item	Format	Possible Points
1	MC	1
2	MC	1
3	MC	1
4	CR	2
5	CR	4
Total Points		9

As Table 14 shows, Group 2 made statistically significant gains from pretest to posttest,. No Group 1 classrooms returned completed pre- and post-tests for this unit.

Table 14. Levers and Machines Unit Test Results

	Pretest		Posttest		Difference	
	M	***SD***	***M***	***SD***	***M***	***SD***
Group 2 (n = 120)	3.9	1.9	5.3	1.8	+1.4**	1.8

***$p < .01$

Results for the Monitoring a Living Plant Unit Test

The Monitoring a Living Plant unit test consisted of seven items, with the points for each item distributed as indicated in Table 14. The total number of points possible for this test was 13.

Table 15. Living Plants Unit Test Item Scoring Algorithm

Item	Format	Possible Points
1	MC	1
2	MC	1
3	CR	5
4	CR	3
5	MC	1
6	MC	1
7	MC	1
Total Points		13

As Table 16 shows, Group 2 made statistically significant gains from pretest to posttest. No Group 1 classrooms returned completed pre- and post-tests for this unit.

Table 16. Living Plants Unit Test Results

	Pretest		Posttest		Difference	
	M	***SD***	***M***	***SD***	***M***	***SD***
Group 2 (n = 193)	5.0	2.4	6.6	2.7	+1.6**	3.1

**$p < .01$

Results for the Pressure Unit Test

The Pressure unit test consisted of four items, with the points for each item distributed as indicated in Table 17. The total number of points possible for this test was 14.

Table 17. Pressure Unit Test Item Scoring Algorithm

Item	Format	Possible Points
1	CR	4
2	MC	1
3	CR	5
4	CR	4
Total Points		14

Group 1 students made significant gains from pretest to posttest on the Pressure unit test, as did Group 2's scores (see Table 18).

Table 18. Pressure Unit Test Results

	Pretest		Posttest		Difference	
	M	***SD***	***M***	***SD***	***M***	***SD***
Group 1 (n = 30)	4.3	1.3	6.1	1.7	+1.9**	1.9
Group 2 (n = 42)	4.5	1.7	5.8	2.4	+1.4**	2.1

** $p < .01$

A t-test comparing the gains showed that the gains of Group 1 were not significantly different from the gains of Group 2 ($t = 1.0$, $df = 70$, $p = .31$).

Results for the Understanding Motion Unit Test

The Understanding Motion unit test consisted of eight items, with the points for each item distributed as indicated in Table 19. The total number of points possible for this test was 16.

Table 19. Motion Unit Test Item Scoring Algorithm

Item	Format	Possible Points
1	MC	1
2	MC	1
3	MC	1
4	MC	1
5	MC	1
6	CR	3
7	CR	3
8	CR	5
Total Points		16

Both Group 1 and Group 2 students made gains from pretest to posttest on the Understanding Motion unit test, but only the gains for Group 1 were statistically significant (see Table 20).

Table 20. Motion Unit Test Results

	Pretest		Posttest		Difference	
	M	***SD***	***M***	***SD***	***M***	***SD***
Group 1 (n = 245)	6.5	2.7	8.0	2.7	+1.5**	2.9
Group 2 (n = 44)	6.9	2.3	7.9	2.3	+1.0**	2.4

** $p < .01$

A *t*-test comparing the gains showed that the gains of Group 1 were no different from Group 2's gains (t = 1.0, df = 287, p = .32).

Effect Sizes in Comparing Group 1 and Group 2 Scores (Analyses 3 and 4)

We computed effect sizes to compare the magnitude of differences between Group 1 and Group 2 students' posttest scores using regression models. We analyzed results using pretests as a covariate in our analysis. Using pretests in this way is a common alternative to using gain scores to compare the performance of two groups' performance, particularly when pre-existing differences between groups may be significant. We report on effect sizes here from this simple regression approach and also from a model in which we included each teacher in the model (the "fixed effects" model) to attempt to control for individual teacher differences. Table 21 presents both model results.

Table 21. Effect Sizes for Group 1-Group 2 Comparisons using Regression Models

Test	Effect Size, Controlling for Pretest (95% Confidence Intervals)	Effect Size, Controlling for Pretest and Teachers (95% Confidence Intervals)
Sound	-.26 (-0.57 – + 0.06)	-.22 (-0.51 – +0.07)
Sensing	+ 0.07 (-0.31 – 0.46)	+ 0.03 (-0.35 – + 0.41)
Temperature	+ 0.56** (+0.40 – +0.72)	+0.33** (+0.16 – + 0.49)
Pressure	+0.28 (-0.34 – +0.89)	+0.25 (-0.35 – +0.85)
Motion	+0.09 (-0.21 – +0.38)	+ 0.10 (-0.19 – + 0.39)

**$p < .01$ (refers to the significance of the regression model coefficient)

As the table shows, the results are consistent with the *t*-tests performed on gain scores, in terms of the differences that were significant. The trend of all the effects, except for Sound, was positive in favor of the treatment group (Group 1), but the effect was significant only for the Temperature test. In general, when confidence intervals for effect sizes include 0, the effect is considered nonsignificant. The magnitude of the effect was slightly larger when we did not take into account the effects of clustering. To determine adequate sample size for a future experimental study, the smaller estimate of effect would be the better estimate to use in planning such a study.

Interpreting the Results

Results of our first analysis indicate that for all of the tests that Group 1 TEEMSS II students took—Sound, Sensing, Temperature, Pressure, and Motion—students on average made statistically significant gains from pretest to posttest. In contrast, Group 2 students made significant gains on just over half of the tests—Sound, Electricity, Levers, Monitoring a Living Plant, and Motion—they took. Of the five tests taken by both groups, Group 1 students made significant gains on all five tests, and Group 2 showed significant gains on three tests.

Results of the second and third analyses focused on the five tests for which we had both Group 1 and Group 2 scores to compare. Both analyses found that for the Temperature unit test, Group 1 students made gains that were significantly greater ($p < .05$) than the gains for Group 2 students. For the other four tests, the gains made were not significantly different from one another across groups. Regression analysis that controlled for teacher effects found a similar pattern. In this analysis, the Temperature test showed a significant effect ($p < .01$). The consistency of these findings points to converging evidence that the Temperature unit had a significant effect on students in this sample of classrooms.

The fourth analysis showed positive effects for four of the tests; using the criterion that the confidence interval does not include zero, the Temperature test was significant. For that test, the effect was relatively large for commonly evaluated educational interventions. As a point of

comparison, the effects of class size reduction as measured by the Tennessee Star Schools study was between +0.14 and +0.17. In contrast, the Temperature unit's effect was roughly one-third of a standard deviation or +0.33.

Two important limitations to the analyses we conducted caution against interpreting these results as evidence of impact of the curriculum units: First, students and teachers were not randomly assigned to condition. Although we found differences on pretest scores only for the Motion test and no significant measured differences between teachers in the two groups, unmeasured differences between groups could bias estimates of the impacts of the units. Past studies show that randomized control trials offer more unbiased estimates of effects than do quasi-experimental studies (Glazerman, Levy, & Myers, 2003).

As is appropriate for projects such as TEEMSS II that are still under development, we controlled for the effects of clustering of students in classrooms by using a fixed effects model instead of attempting to achieve sufficient power to analyze teacher effects. Such studies are prohibitively expensive for development studies, as recognized by the Institute of Education Sciences at the U.S. Department of Education in its recent call for research proposals in mathematics and science education (see http://ies.ed.gov/ncer/funding/). Using a fixed-effects model, however, entails a significant trade-off: it is not possible to generalize results beyond the sample used in the study.

These limitations do not prevent evaluation results from being interpretable, however, given that quasi-experimental designs that account statistically for pretest differences are a generally interpretable form of evaluation research (Shadish, Cook, & Campbell, 2002). Therefore, the results for the Temperature unit are indeed promising and constitute important preliminary evidence of the efficacy of that TEEMSS II unit.

Appendix A

TEEMSS Assessment Resources

Assessment / Organization	Web Location
National Assessment of Educational Progress (NAEP)	http://nces.ed.gov/nationsreportcard/science/
Trends in International Math and Science Study (TIMSS)	http://nces.ed.gov/timss/Educators.asp
Florida	http://www.firn.edu/doe/sas/fcat/fcatit03.htm
Illinois	http://www.isbe.net/assessment/PDF/2003ScienceSample.pdf http://www.isbe.net/assessment/PDF/2002ScienceSample.pdf http://www.isbe.net/assessment/PDF/2000ScienceSample.pdf http://electron-net.eztest.eppg.com/ISBE/PSAE/Science/
Louisiana	http://www.doe.state.la.us/lde/uploads/3788.pdf http://www.doe.state.la.us/lde/uploads/3786.pdf http://www.doe.state.la.us/lde/uploads/3785.pdf
Maryland	http://www.mcps.k12.md.us/curriculum/science/elem/scisummatives.htm http://www.mcps.k12.md.us/curriculum/science/forms/hsabiopubrelease03.pdf http://www.mcps.k12.md.us/curriculum/science/assess/hsa.htm
Massachusetts	(http://www.doe.mass.edu/mcas/testitems.html)
Michigan	http://www.michigan.gov/documents/Gr8Def_96523_7.pdf
New York	http://www.nysedregents.org/
North Carolina	http://www.ncpublicschools.org/accountability/testing/eog/ http://www.ncpublicschools.org/accountability/testing/eog/science/
Texas	http://www.tea.state.tx.us/student.assessment/resources/release/taks/2003/gr11taksscience.pdf http://www.tea.state.tx.us/student.assessment/resources/release/taks/2003/gr10taksscience.pdf http://www.tea.state.tx.us/student.assessment/resources/release/taks/2003/gr5taks.pdf http://www.tea.state.tx.us/student.assessment/resources/release/taks/2004/gr11taks.pdf http://www.tea.state.tx.us/student.assessment/resources/release/taks/2004/gr10taks.pdf http://www.tea.state.tx.us/student.assessment/resources/release/taks/2004/gr5taks.pdf http://www.tea.state.tx.us/student.assessment/resources/release/taas/release02/gr8.pdf http://www.tea.state.tx.us/student.assessment/resources/release/taas/release01/gr8.pdf http://www.tea.state.tx.us/student.assessment/resources/release/taas/release00/gr8.pdf
Washington	http://www.k12.wa.us/assessment/WASL/testquestions.aspx

Appendix B

TEEMSS Items Co-Development Summary

		TEEMSS Units	# of items			
			In review notebook	Field Tested		used and (scored items)
				revisions	New items	
Grades 3-4	1	Sound	44	12	8	7 (7)
	2	Electricity (and Magnetism*)	49	21	0	8 (4)
	3	Human and Electronic Sensing	22	4	11	9 (8)
	4	Weather	33	18	14	14**
	5	Design a Playground	9	Portfolios will be collected and evaluated by CC in year 3 to evaluate this design unit		
Grades 5-6	6	Water and Air Temperature	40	7	15	6 (6)
	7	Levers and Machines	52	24	4	6 (5)
	8	Monitoring a Living Plant	16	13	4	7 (7)
	9	Seasons	117	12	13	12 (12)**
	10	Design a Greenhouse	30	Portfolios will be collected and evaluated by CC in year 3 to evaluate this design unit		
Grades 7-8	11	Pressure	15	7	5	5 (4)
	12	Understanding Motion	46	11	9	8 (6)
	13	Evolution	73	26	0	4 (4)***
	14	The Water Cycle	49	17	9	8 (8)**
	15	Design a Measurement	91	Portfolios will be collected and evaluated by CC in year 3 to evaluate this design unit		

* Broader conceptual limits of the unit are identified in parentheses and include concepts found in early stages of curriculum development. Content was refined during iterative cycles of instructional and instrument development. The numbers identified in parentheses include the items linked to the initial content, including the content that was eventually dropped

**These curriculum units were still under development and not included in Year 1 research. It is anticipated that this number of items will be scored during Year 2 research.

***These items were not scored in Year 1 research because there was insufficient participant response to this unit.

Appendix C

Distribution of Students in Classrooms Included in Analysis

Sound	# Students
Group 1	38
Participant 11	23
Participant 13	15
Group 2	154
Participant 38	9
Participant 39	12
Participant 50	19
Participant 51	13
Participant 52	13
Participant 53	17
Participant 54	22
Participant 55	16
Participant 56	13
Participant 57	20

Electricity	# Students
Group 1	NONE
Group 2	185
Participant 36	9
Participant 38	10
Participant 39	13
Participant 41	19
Participant 50	19
Participant 51	14
Participant 52	14
Participant 53	15
Participant 54	22
Participant 55	16
Participant 56	14
Participant 57	20

Human & Electronic Sensing	# Students
Group 1	126
Participant 11	24
Participant 12	15
Participant 14	11
Participant 15	11
Participant 20	24
Participant 21	21
Participant 22	20
Group 2	35
Participant 41	19
Participant 42	16

Water &Air Temperature	# Students
Group 1	251
Participant 08	91
Participant 18	71
Participant 25	42
Participant 26	24
Participant 27	23
Group 2	148
Participant 43	49
Participant 48	14
Participant 58	30
Participant 59	55

Levers & Machines	# Students
Group 1	NONE
Group 2	120
Participant 33	22
Participant 34	16
Participant 35	20
Participant 40	25
Participant 45	22
Participant 48	15

Monitoring A Living Plant	# Students
Group 1	NONE
Group 2	193
Participant 35	21
Participant 40	24
Participant 43	54
Participant 45	10
Participant 58	32
Participant 59	52

Pressure	# Students
Group 1	30
Participant 30	30
Group 2	42
Participant 37	20
Participant 44	22

Motion	# Students
Group 1	244
Participant 6	64
Participant 7	92
Participant 17	88
Group 2	44
Participant 37	20
Participant 44	24

References

Blumenfeld, P., Fishman, B. J., Krajcik, J., Marx, R. W., & Soloway, E. (2000). Creating usable innovations in systemic reform: Scaling up technology-embedded project-based science in urban schools. *Educational Psychologist, 35*(3), 149-164.

Glazerman, S., Levy, D., & Myers, D. (2003). *Nonexperimental replications of social experiments: A systematic review*. Princeton, NJ: Mathematica Policy Research, Inc.

Metcalf, S. J., & Tinker, R. (2003, March). *TEEMSS: Technology Enhanced Elementary and Middle School Science*. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Metcalf, S. (2006). TEEMSS II: Technology Enhanced Elementary and Middle School Science - Year 1 Report. In S. A. Barab, K. E. Hay & D. T. Hickey (Eds.), *7th International Conference of the Learning Sciences* (Vol. 1, pp. 474-480). Mahwah, NJ: Erlbaum.

National Research Council. (1996). *National Science Education Standards*. Washington, DC: National Academy Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.