
F i n a l R e p o r t

Prentice Hall **2005 *Science Explorer*** **Pilot Study**



Prepared by:
Miriam Resendez, Senior Researcher
Curtis Stauffer, Research Consultant
Mariam Azin, President

PRES Associates
292 E. Pearl Street
Jackson, WY 83002
(307) 733-3255
www.presassociates.com

© 2005 PRES Associates, Inc.

Executive Summary

Prentice Hall Publishers asked PRES Associates, an external, independent educational research firm with over fifteen years of experience in applied educational research and evaluation, to conduct a pilot study on the effectiveness of the 2005 *Science Explorer* program in conjunction with Claremont Graduate University. The pilot study was conducted at one middle school site from December 2004 to May 2005. The primary purpose of the pilot study was to develop the instrumentation and inform the design of a multi-site, yearlong randomized control trial (RCT) to be conducted on the program during the 2005-6 school year¹.

This report provides information on the design of the study and procedures used; background information on the pilot study site, including teachers and students; and preliminary findings from the study. The report also summarizes the lessons learned during the pilot study and how such findings will be integrated into the design of the 2005-6 RCT.

As a result of the instrument development and piloting that occurred during the pilot study, PRES Associates has been able to develop and select reliable, valid instruments and procedures for the RCT that will enable researchers to measure:

- *The impacts of the 2005 Science Explorer program on student performance.* PRES Associates selected the science portions of the Iowa Test of Basic Skills (ITBS) and developed an assessment derived from items released from the 2003 Trends in International Math and

Science Study (TIMSS). While each measure emphasizes different constructs, both demonstrate high reliability and validity and good item difficulty. Using both of these measures will increase the sensitivity of the RCT in detecting differences and increase the validity of results.

- *Other affective outcomes related to the use of 2005 Science Explorer and external effects that could impact the RCT.* PRES Associates has developed reliable and valid teacher and student surveys and teacher interview protocols that will measure affective student outcomes (e.g. enjoyment of science, satisfaction with science program materials) and teacher attitudes and classroom practices (e.g. satisfaction with program materials, preferred teaching practices). Additionally the surveys gather data on teacher characteristics and other external factors that could potentially influence study results (e.g. teaching experience and background, level of student parental support, etc.). If significant differences between treatment and control groups are found in these areas during the RCT, statistical controls will be employed.
- *The fidelity of implementation of the 2005 Science Explorer program and contamination.* PRES Associates will employ multiple measures (classroom observations, monthly online teacher classroom activity logs, implementation checklists, teacher surveys and exit interviews) to monitor how 2005 *Science Explorer* is implemented in treatment group classrooms as well as how science is taught in control group classes. These measures were piloted and proven to be effective during the pilot study. With input from Prentice Hall staff, PRES Associates also developed implementation guidelines and training recommendations that will ensure that RCT participants can implement the 2005 *Science Explorer* program with fidelity.

Because of the short duration of the pilot study, the small sample size and the fact that randomization did not take place, this pilot study was not expected to produce rigorous quantitative evidence upon which strong conclusions can be drawn and that meets the criteria of the What Works Clearinghouse's

¹While the pilot study for *Science Explorer* has been contracted to CGU, with PRES Associates as a subcontractor, it is important to note that the RCT planned for *Science Explorer* during 2005-6 will be the sole responsibility of PRES Associates

(WWC) Study DIAD.² Nevertheless, researchers can report preliminary findings regarding the utility and efficacy of the Science Explorer program. These findings include the following.

- *Positive effects were observed from the use of 2005 Science Explorer.* Users of 2005 Science Explorer demonstrated significant growth in performance on the TIMMS-F test (overall, Life Science, and Physical Science). Though there was no significant difference between control and treatment group performance as a whole, female treatment group students had better Total and Physical Science TIMSS-F scores than female control group students. Additionally, low socio-economic status students from the treatment group outperformed their counterparts in the control group on Total, Life Science, and Physical Science scores of the TIMSS-F.
- *Users expressed satisfaction with 2005 Science Explorer.* Treatment group teachers were more satisfied with the 2005 Science Explorer program than control group teachers were with their curriculum. Treatment teachers generally found the 2005 Science Explorer program to be an effective pedagogical tool, and were impressed with many of the ancillary resources. Only a few weaknesses were noted. Students using 2005 Science Explorer reported greater satisfaction with their curriculum than students using the control texts.

Overall, the pilot study proceeded smoothly. During the course of the study, treatment group teachers mostly implemented the 2005 *Science Explorer* program with fidelity, and training and monitoring procedures proved to be effective. The only major issue encountered during the pilot study was the gradual release of program materials over the course of the study. The lack of Spanish language resources at the onset of the study prevented random assignment, and the absence of several technological components limited their use by participating teachers during the study period. That these materials were not present at the outset of the study also limited the hands-on training teachers received.

The pilot study has enabled PRES Associates to design procedures and instruments for the 2005-6 RCT that will ensure that the effects of 2005 *Science Explorer* can be reliably and effectively measured in keeping with the standards established by the WWC.

² Valentine, J.C. & Cooper H. (2003). *What Works Clearinghouse Study Design and Implementation Assessment Device*. Washington, DC: US Department of Education.

Table of Contents

List of Tables	5
List of Figures.....	6
Project Background	7
Project Overview.....	8
Study Design	10
Site and Sample Characteristics	13
Results	19
Major Findings.....	19
Detailed Findings	20
Conclusions.....	31
Project Summary.....	31
Lessons Learned and Implications for RCT	31
Appendix A - Implementation Guidelines	33
Appendix B - Teacher Ratings of Program Usefulness of the Program	40
Appendix C - 2005 <i>Science Explorer</i> Resources Used.....	42
Appendix D - Development of the TIMMS-F and Psychometric Properties	45

List of Tables

Table 1. Key Pilot Study Questions and Findings.....	9
Table 2. Study Site Student Demographics	13
Table 3. Site MCAS Test Results as Compared to Statewide Averages.....	14
Table 4. Participating Student Demographics.....	14
Table 5. Participating Student and National Sample Performance on ITBS	15
Table 6. Teaching Experience	16
Table 7. Percent of Time Spent (Weekly) on Each Class Activity by Group	18
Table 8. Psychometric Properties of the TIMSS-F	25
Table 9. Reliability of Student Survey Subscales	26
Table 10. Reliability of Teacher Survey Subscales	26
Table B1. Teacher Ratings of Usefulness of Their Respective Program Components	41
Table C1. Number of Times Treatment Teachers Reported Using the Following Lab Activities....	43
Table C2. Number of Times Treatment Teachers Reported Using the Following Technology Resources	43
Table C3. Number of Times Treatment Teachers Reported Using the Following Print Resources	44
Table D1. Content Breakdown of TIMSS and TIMSS-RE	48
Table D2. Cognitive Breakdown of TIMSS and TIMSS-RE	48
Table D3. Content Breakdown of TIMSS and TIMSS-F	48
Table D4. Cognitive Breakdown of TIMSS and TIMSS-F	49
Table D5. IRT Parameters for Pre- and Post-Test TIMSS-F and TIMSS	49
Table D6. Pearson Correlations.....	50
Table D7. Cronbach's Alpha: Pre- and Post-Test.....	52
Table D8. Concurrent Validity of TIMSS-F.....	53

List of Figures

Figure 1. Teacher Attitudes on the 2005 <i>Science Explorer</i> Program	20
Figure 2. Student Attitudes on Program Components	21
Figure 3. Science Performance Change Among Treatment Students.....	22
Figure 4. Science Content Domain (TIMSS-F) Change Among Treatment Students.....	22
Figure 5. Science Performance Between Treatment and Control Students	23
Figure 6. Science Control Domain (TIMSS-F) Performance Between Treatment and Control Students	23
Figure 7. Science Performance Between Female Treatment and Control Students.....	23
Figure 8. Science Performance Between Low SES Treatment and Control Students.....	24
Figure D1. 30 Item Information Function	51
Figure D2. Scree Plot: Post TIMSS-F	52

Project Background

There is a profound need to learn “what works” in science education. In an ambitious effort to improve mathematics and science achievement across the country, the U.S. Department of Education launched a major five-year Mathematics and Science Initiative in February 2003. A key goal of this initiative is to develop a comprehensive, academic research base “to improve our knowledge of what boosts student learning in mathematics and science in the classroom.”³ A component of this goal is to identify effective science interventions through research that investigates specific teaching methods and curriculum materials.

Additionally, the No Child Left Behind Act of 2001 (NCLB) mandates that educational materials purchased with public funds must be proven by scientific research to improve student achievement in the classroom⁴.

Given the need to show that its programs are research-based, Prentice Hall Publishers asked Claremont Graduate University (CGU), in conjunction with PRES Associates⁵ to conduct pilot studies on the effectiveness of their Prentice Hall *Biology* (10th Grade) and *Science Explorer* (8th Grade) programs. These pilot studies were intended to inform the design and implementation of multi-site, yearlong randomized control

trials (RCTs) to be conducted on these programs during the 2005-6 school year. These randomized control trials are designed to meet the study review standards⁶ established for curricular research by the What Works Clearinghouse (WWC). While both pilot studies fell under the auspices of CGU, PRES Associates was primarily responsible for conducting the *Science Explorer* pilot study⁷, while CGU ran the *Biology* pilot study. This report only addresses the 8th Grade *Science Explorer* pilot study.

The Prentice Hall 2005 *Science Explorer* program is a comprehensive, basal program designed for grades 6-8 that allows users to select from five life science modules, five earth science modules, five physical science modules, and one science and technology module to meet their local curricular requirements. The program, aligned with the National Science Education Standards, emphasizes the development of science inquiry skills in addition to the acquisition of content knowledge. The program integrates math and reading support into the curriculum, and emphasizes ongoing assessment with a range of methods for assessing student knowledge, including scaffolded questions. The program also offers teachers strategies for differentiated instruction. 2005 *Science Explorer* provides users with a wide range of ancillary materials, including workbooks, materials kits, online resources, and lab videos.⁸

³ US Department of Education (2003). *Description of Mathematics and Science Initiative*. (Washington, DC: Author). Downloaded from the web on February 16, 2005 from <http://www.ed.gov/rschstat/research/progs/mathscience/describe.html>.

⁴ The What Works Clearinghouse (WWC) has been established, in part, as a result of NCLB. The articulated purpose of WWC is to “provide educators, policymakers, and the public with an independent source of scientific evidence of what works in education.”

⁵ PRES Associates is an external, independent educational research firm with over fifteen years of experience in applied educational research and evaluation. For more information, please visit www.presassociates.com.

⁶ The WWC’s Study Review Standards reflect study characteristics originally contained in the Study Design and Implementation Assessment Device (Study DIAD v1.0). The standards can be found at www.whatworks.ed.gov/reviewprocess/study_standards_final.pdf.

⁷ While the pilot study for *Science Explorer* has been contracted to CGU, with PRES Associates as a subcontractor, it is important to note that the RCT planned for *Science Explorer* during 2005-6 will be the sole responsibility of PRES Associates.

⁸ For more information on the 2005 *Science Explorer* program, please refer to www.phschool.com/science. For a list of ancillary materials provided to teachers during the pilot study, please refer to the *Implementation Guidelines* found in Appendix A of this report.

Project Overview

As noted above, a rigorous, quantitative study on the effectiveness of the 2005 Prentice Hall *Science Explorer* program is planned for 2005-2006. This yearlong study will consist of a randomized control trial (RCT) that has been designed to fully meet the quality criteria put forth by the WWC.⁹ In preparation for this larger yearlong study, PRES Associates conducted a pilot study in the Spring semester of 2005.

The primary purpose of the pilot study was to inform the design and help finalize the instrumentation for the full-scale RCT; it was *not* designed to produce conclusive evidence on the effectiveness of the *Science Explorer* program. That is, due to the short duration of the pilot study, and the fact that randomization did not take place, it was not expected to produce rigorous quantitative evidence upon which strong conclusions can be drawn and which meets the criteria of the WWC's Study DIAD.¹⁰ Nevertheless, researchers examined preliminary information about the relationship between the *Science Explorer* program and student science performance. Specifically, PRES Associates could address the key questions outlined in the following table during the pilot study.

⁹ The What Works Clearinghouse (WWC) has been established, in part, as a result of NCLB. The articulated purpose of WWC is to "provide educators, policymakers, and the public with an independent source of scientific evidence of what works in education."

¹⁰ Valentine, J.C. & Cooper H. (2003). *What Works Clearinghouse Study Design and Implementation Assessment Device*. Washington, DC: US Department of Education.

Table 1. Key Pilot Study Questions and Findings

Pilot Study Questions	Findings
<i>What did the users of the Prentice Hall 2005 Science Explorer think of it?</i>	<ol style="list-style-type: none"> 1) Treatment group teachers were more satisfied with the 2005 <i>Science Explorer</i> program than control group teachers were with their curriculum. Treatment teachers generally found the 2005 <i>Science Explorer</i> program to be an effective pedagogical tool, and were impressed with many of the ancillary resources. Only a few weaknesses were noted. 2) Students using 2005 <i>Science Explorer</i> reported greater satisfaction with their curriculum than students using the control texts.
<i>Was there any relationship between use of the program and improvement in science performance?</i>	<ol style="list-style-type: none"> 1) There was significant growth in student science performance among students who used the PH 2005 <i>Science Explorer</i> program as measured by the Trends in International Math and Science Study-F (TIMSS-F) test (overall, Life Science, and Physical Science). However, there was a significant decline in student performance on the Earth Science content domain. This is likely due to the treatment teachers focusing on physical science during the course of the pilot study. Treatment teachers drew from 3 chapters in the <i>Chemical Building Blocks</i> module, 2-3 chapters (depending on the teacher) in the <i>Motion, Forces, and Energy</i> module, and 1-2 chapters in the <i>Chemical Interactions</i> module. 2) Overall there were no significant differences between control and treatment students as measured by all tests. However, subgroup analyses indicated that female treatment students performed better than female control students on TIMSS-F Total and Physical Science score. In addition, low socioeconomic status (SES) treatment students performed better than low SES control students on the TIMSS-F Total, Life Science, and Physical Science scores.
<i>Which assessments will best measure the effects of Science Explorer?</i>	<ol style="list-style-type: none"> 1) Based upon our analysis of piloted assessments, it is recommended that the RCT assessment package include: 1) the Iowa Test of Basic Skills-Level 14 Complete Battery (science portion only), and 2) an assessment, the TIMSS-F, developed from the released items from the 2003 TIMSS exam.
<i>Do the measures demonstrate validity, reliability, and sensitivity to detect differences?</i>	<ol style="list-style-type: none"> 1) According to test publishers' documentation and our analyses of pilot data, the selected assessments are reliable, valid, and sensitive in measuring students of varying ability levels. In terms of construct validity, the assessments are tailored toward nationally and internationally accepted norms and standards for student performance. 2) The student and teacher surveys demonstrated moderate to high levels of reliability.
<i>What are the key characteristics of 2005 Science Explorer implementation and training?</i>	<ol style="list-style-type: none"> 1) Treatment teachers were provided with implementation guidelines to ensure that key components of the 2005 <i>Science Explorer</i> program were used. Generally, treatment teachers did not feel restricted by the implementation guidelines. In addition, they felt that the guidelines provided adequate direction regarding implementation requirements. These guidelines include the following key components: <ul style="list-style-type: none"> ▪ Follow pacing guidelines ▪ Chapter science project (each chapter) ▪ Lab activities (at least one demonstration or hands-on activity for every chapter) ▪ Pre-teach activity (each section) ▪ Check understanding (throughout each section) ▪ Assess knowledge (at the end of each chapter) ▪ Close the lesson (each lesson) ▪ Independent practice (30 minutes each school night, as feasible) ▪ <i>Math: Analyzing Data and Writing in Science</i> (each section) ▪ Differentiated instruction (throughout sections, as needed) 2) Trainings should consist of the following elements: a) a thorough discussion of the philosophy and research-base behind the program; b) walking the teachers through the instruction of a typical chapter while pointing out all key program components and the ways in which they can be met, and providing program resources and training prior to the start of the school year, to the extent that this is feasible.

This pilot study will help ensure that PRES Associates can address the following research questions during the 2005-6 RCT.

- Does student performance in science improve as a result of participation in the 2005 Prentice Hall *Science Explorer* program?
- Do students of teachers who use *Science Explorer* perform better in science as compared to students of teachers who employ other types of science programs?
- Do effects on student achievement differ across types of students or settings? How well does *Science Explorer* work with different types of students at different levels of abilities?
- Does use of the *Science Explorer* program result in other positive student outcomes (e.g., positive attitudes towards science, school, etc.)?

This report provides information on the procedures used for the pilot study, background information on the pilot site (including its teachers and students), and the preliminary findings from the pilot study. The report concludes with a list of accomplishments, the lessons learned during the pilot study and the implications of these findings to the RCT.

Study Design

One large middle school in Massachusetts was recruited to participate in the study in October 2004. In order to mimic the conditions that would be employed during the 2005-6 RCT, the school was asked to randomly assign its 8th grade science teachers and their classes into treatment and control groups. Because the 2005 *Science Explorer* Spanish-language materials would not be released until the Spring, the school requested to assign its teachers with English Language Learner

“immersion” classes to the control group so that ELL students would have continued access to the Spanish language resources available for the control text. Thus, random assignment of teachers and control groups could not occur under these conditions. *Random assignment* of teachers within schools¹¹ will occur during the 2005-6 RCT.

Treatment group teachers received new 2005 *Science Explorer* materials for their classes and training on the program in November 2004, and began using the program with their classes the first week of December 2004. Treatment group teachers drew their lessons from the following 2005 *Science Explorer* materials during the study period.

- *Chemical Building Blocks* (Physical Science)
- *Chemical Interactions* (Physical Science)
- *Motion, Forces, and Energy* (Physical Science)

Physical science was the only content area covered by 8th grade science teachers in the Spring Semester, thus, program level effects on student performance most likely will only occur in this domain.

Methods

A range of data was collected in the pilot study, including descriptive information, program implementation data, and outcome data. These activities spanned from December 2004 to June 2005. Copies of all instruments are available upon request from PRES Associates.

¹¹ There are a number of reasons that PRES Associates chose assignment to treatment conditions be done at the teacher level within schools for both the pilot and the upcoming RCT. The most important reason for selecting this level of assignment is that such a design helps to establish causality by eliminating the threat that school level factors could have potentially contributed to differences between treatment and control groups.

Student Rosters: Student rosters for participating classes were obtained in January 2005 from the school. Basic demographic information was obtained for each participating student, including gender, ethnicity, free/reduced lunch participation, special education status, and Limited English Proficiency status. Student rosters were continually updated based on student transfers, using online teacher logs (see below). These demographic data were collected to allow researchers to describe the pilot students and will be used for the RCT to conduct analyses on the effects of the *Science Explorer* program within different subgroups.

Teacher Logs/Implementation Checklist: Teacher implementation (for treatment teachers) and instructional practices (for all teachers) were monitored on a weekly basis with the use of a web-based data collection system. Teachers used the log to: 1) add/remove students from their student roster; 2) inform PRES Associates of their weekly classroom activities, including implementation of program components; 3) inform PRES of the resources used; and 4) communicate to PRES Associates any issues that they may have encountered in regards to the study or 2005 *Science Explorer* program. Different surveys were employed for treatment and control teachers. In addition, a teacher implementation checklist was developed so that teachers could note the percent of lessons that were covered in their respective science program(s).

Classroom Observations: Classroom observations were conducted in January and April of 2005. Observations focused on how class activities were structured, what and how materials were used, characteristics of the students (including student engagement), classroom environment and culture, and teachers' abilities to teach effectively. In addition, teachers were

interviewed after the observations to obtain more specific information on the representativeness of the lesson, resources used, ability levels of the students, and assessment, pacing, independent practice, and test preparation strategies. PRES Associates developed a classroom observation protocol for use during observations.¹²

Teacher Survey: All participating teachers completed the teacher survey during January (pre) and June (post) 2005. The survey, constructed by PRES Associates and partly based on other surveys,¹³ was developed to collect information on:

- current and past classroom and instructional practices;
- teacher knowledge of effective teaching practices (including those specific to science instruction);
- student learning;
- attitudes about science curricular resources; and
- general background and demographic information.

For the RCT, items from this survey will be used as measures of potential additional outcomes (i.e., teacher practices) as well as to gather background information (e.g.,

¹² The Classroom Observation Form was derived largely from a form constructed and validated in the 2004 Prentice Hall *Algebra I* Pilot Study. For more information on this instrument, refer to Resendez, M.G. and Manley, M. (2004). *Final Report: Prentice Hall Algebra I Program Pilot Study*. The instrument was modified to reflect content typical of 8th grade science classes and implementation of critical components of the Prentice Hall *Science Explorer*, using items from the program implementation guidelines, Horizon Research's *Local Systematic Change Professional Development Classroom Observation Protocol*, and the *Texas Collaborative for Excellence in Teacher Preparation Classroom Observation Protocol*.

¹³ Items in this survey have been developed by PRES Associates and modified from the *Trends in International Mathematics and Science Study (TIMSS) 2003 Teacher Questionnaire Science Grade 8* (Washington, DC: National Center For Education Statistics) and the *2000 National Survey of Science and Mathematics Education Science Questionnaire* (Rockville, MD: Westat). Items were also modified from a survey constructed and validated in the 2004 Prentice Hall *Algebra I* Pilot Study (refer to the reference in the footnote above).

number of years of teaching, etc.) that may subsequently be used as covariates.

Student Survey: Pre- and post-versions of a student survey were created to gather information on their attitudes about science and school, perceptions of effort and motivation, perceptions of science ability, perceived relevance of science to daily life, parental knowledge and support, classroom experiences, opinions on teacher practices, and background and demographic information¹⁴. Like the teacher survey, these were administered in January and June 2005. For the RCT, scales will be used as measures of potential additional outcomes (i.e., affective student outcomes) or covariates (e.g., parental involvement).

Teacher Exit Interview: Interviews were conducted with all participating teachers at the conclusion of the study period to gather detailed information on teaching strategies employed, satisfaction with materials, and for treatment teachers, feedback on the 2005 *Science Explorer* program. This interview also gathered feedback on study procedures employed, and issues encountered as a result of participation in the pilot study.

Student Assessments: Initially, CGU selected a contractor to provide assessments containing selected and constructed response items for both the *Biology* and *Science Explorer* pilot studies. However, when the *Science Explorer* pilot study was nearing its start date, the assessment materials and necessary technical information (e.g., content, reliability, validity) had not been provided to PRES Associates. Therefore, it was deemed necessary to select and/or develop alternative instruments. Important

criteria used in assessment selection and development included: the validity, reliability, and sensitivity of the instrument(s); alignment to national 8th Grade Science standards; a mix of selected and constructed response items; administration time; and cost. Additionally, assessments had to touch upon all 8th Grade Science content areas (rather than those taught in the pilot site), due to the wide range of topical areas that could be tested in the 2005-6 RCT as a result of *Science Explorer's* modular system.

After a thorough literature review of existing standardized group-administered assessments, the following assessments were piloted by PRES Associates: 1) the Iowa Test of Basic Skills (ITBS)-Complete Battery, Level 14 Form A (science portions only), and 2) an assessment, developed by PRES Associates, using released items from the 2003 Trends in International Math and Science Study (TIMSS) exam, TIMSS-RE. Both assessments were administered at pre- and post-testing, although slightly different versions of the TIMSS-RE assessment were used in order to pilot a larger pool of assessment items. Based upon item analyses conducted on pretest results, a final version of the TIMSS-based assessment, called TIMSS-F was developed.

ITBS: This National Science Teachers Association (NSTA) and American Academy for the Advancement of Science (AAAS) standards-aligned assessment covers 4 major content areas: scientific inquiry (20 items), life science (9 items), earth and space science (7 items), and physical science (8 items). The test focuses on “presenting real life science investigations with questions emphasizing thought processes used in designing and constructing research and analyzing data.”

¹⁴ Items in this survey were adapted from the: 2003 *TIMSS Student Questionnaire-8th Grade*; O'Neill and Abedi (1996) *Reliability and Validity of a State Metacognitive Inventory* (Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST)); the *Indiana Mathematics Beliefs Scale*, and the *Fennema-Sherman Math Attitude Scale*.

TIMSS-based: The content of the original 2003 TIMSS assessment, which contains both selected response and constructed response (rubric-scored) items, was developed by an international panel of expert science educators. Like the 2003 TIMSS science assessment, the first shortened version created by PRES Associates, TIMSS-RE, is framed by two organizing dimensions, a content domain and a cognitive domain. There are five content domains (life science, chemistry, physics, earth science, and environmental science) that define the specific science subject matter covered by the assessment. Three cognitive domains (factual knowledge, conceptual understanding, and reasoning and analysis) define the sets of behaviors expected of students as they engage with science content.

The TIMSS-RE version had a longer administration time than anticipated and was subsequently reduced from 54 items to 30 items based on items' psychometric properties. The revised version of this exam, used for the post-test, is hereby referred to as the TIMSS-F. For more detailed information on the development of the TIMSS-F and its psychometric properties, see the results section of this report or Appendix D.

Site and Sample Characteristics

About the Site

The pilot study site is a Title I school located in a lower-middle class neighborhood near the center of a small city in Massachusetts. The only regular public middle school for the district (there is also a small charter school), the site is a large, multi-story complex consisting of an older building and a newer addition. The school

has been fully renovated within the last 15 years. Enrollment for the pilot study site, as of September 2004, was 1061 students in grades 6-8. The school's student population is diverse. Demographics are reported as follows.

Table 2. Study Site Student Demographics

N=1061 (Grades 6-8)	
Category	Percentage
White	61%
Hispanic	25%
African American	7%
Asian	3%
Other	4%
Free/ Reduced Lunch	43%
Special Ed. (w/ IEP)	19%
Limited English Proficient	6%

The site offers immersion classes for English language learners and translated texts. Immersion classes have an immersion teacher who provides assistance to ELL student (including translation of lessons), in addition to the primary teacher. Additionally, most math and science classes have either a "Title I" or an immersion teacher assigned to assist the primary teacher.

The school is divided into three "wharves," (in keeping with the town's nautical history) where teams of teachers work collaboratively and report to assistant principals who oversee each "wharf." Students take all their classes within a particular wharf. With the exception of ELL immersion classes, class groupings in science are heterogeneous. In math, approximately one-half of 8th grade students takes Algebra, while the other half takes a standard 8th grade math program.

The site operates on a block schedule, using a 6-day cycle. Science is taught daily, but classes alternate in length. Classes are 60 minutes on 3 days of the cycle and 30-40 minutes on the other 3 days of the cycle. On days with the shorter science class, classes are typically split in half, with half the class attending world language classes for one portion of the block while the other half attends science. There is one science lab available for *all* 8th grade teachers, regardless of “wharf.” Classes typically have access to the lab for one period per 6-day cycle. Most teachers prefer to conduct labs during shorter blocks, when they can work with one-half of the class at a time.

The state tests for different subjects at different grade levels. In 2003-4, results for the pilot site are as follows. Statewide results are provided for comparison. Site student performance on statewide assessments can be generalized as below average.

Table 3. Site MCAS Test Results as Compared to Statewide Averages

Test Results	Advan- ced	Prof- icient	Needs Impvmt.	Fail
<i>Site- Grade 7 English</i>	2%	53%	38%	7%
Statewide - Grade 7 English	9%	59%	25%	7%
<i>Site - Grade 8 Math</i>	5%	21%	37%	37%
Statewide - Grade 8 Math	13%	26%	32%	29%
<i>Site - Grade 8 Sci. & Tech.</i>	1%	15%	39%	45%
Statewide - Grade 8 Sci. & Tech.	5%	28%	36%	31%

About the Pilot Students

Approximately 286 eighth grade students in 14 classes (6 control and 8 treatment) participated in the *Science Explorer* pilot study. There were slightly more students in the treatment group ($N=161$, 56.3%) compared to the control group ($N=125$, 43.7%). The average class size was 20 students with a range from 17 to 25 students. Demographic information on the pilot study participants is presented in the following table.

Table 4. Participating Student Demographics

Participants	Number	% White	% Hispanic	% African Am.	% Asian	% Female	% Limited English	% Special Education	% Free/Reduced Lunch
<i>Treatment</i>	161	64.6	26.1	6.8	2.5	53.4	1.2	7.5	37.9
<i>Control</i>	125	64.8	26.4	4.0	4.8	48.4	4.8	23.2	34.4
Total	286	64.7	26.2	5.6	3.5	51.4	2.8	14.3	36.4

Generally, the participants’ distribution on demographic characteristics is similar to those found school-wide. In addition, there are some discrepancies in the distribution of demographic characteristics between the treatment and control students but this is likely due to the small sample sizes and the non-random assignment of students into groups (which resulted in a larger proportion of LEP students in the control group). Also, to reiterate, the purpose of this study was not to produce conclusive results using a generalizable sample. However, this will be an important consideration for the RCT.

Teachers generally report that all classes are of average ability, with the exception of the control teachers with immersion classes, who report that those classes are of lower ability than the general student population.

This is consistent with comparisons between the pilot sample of students and national student performance data from the ITBS. As shown in the figure below, treatment students were generally of higher ability compared to control students and the national sample. This is unsurprising since the English Language Learners (ELL) could not use the new *Science Explorer* program because the resources were not yet available.

Table 5. Participating Student and National Sample Performance on ITBS

	Spring Scale Score
Sample Treatment Students	256.02
Sample Control Students	245.93
National Sample	251.50

Table 5 shows that treatment students were generally of higher ability compared to control students during Spring testing. Preliminary analyses showed that the treatment students (M=244) performed *marginally* significantly higher than control students (M=252) on the ITBS pretest, $t(233)=1.77, p=.08$. To control for these differences and to increase the power of the test to detect differences, groups were statistically equated in subsequent analyses.

About the Teachers

There were 7 teachers participating in the study (3 control and 4 treatment), each with 2 classes. Five teachers are female and all are white¹⁵. In regards to educational background, teachers had Bachelor's or Master's degrees in Elementary Education and Educational Leadership. Four teachers had completed 6 to 35 hours of professional development in the past 12 months and 6 teachers had received this amount of professional development in the last 3 years.

Teaching experience ranged among the participating teachers.¹⁶ It will be important to examine whether these varied backgrounds are associated with student performance.¹⁷ Two teachers have Bachelor's and PhD/JD degrees, respectively, while the remaining three have Master's degrees. The type of degree earned also varied among the teachers; only two have science specific degrees, while the remaining teachers have degrees in education (2) and other subjects (2). As shown in the table below, teacher experience also varied among the participating teachers. In particular, only two teachers had extensive teaching experience (over 10 years) and both were treatment teachers. However, due to the non-random nature of this pilot study and the small number of teachers in the pilot study, this is not atypical. For the RCT, random assignment *should* produce equivalent groups and, if not, statistical controls can be employed.

¹⁵ One teacher refused to respond to the ethnicity item.

¹⁶ During the course of the study one control teacher left the school. This teacher was replaced by a short-term substitute and then by a long-term substitute who finished the year. The initial teacher was male, and the long-term substitute was female, so at the conclusion of the study, there were 6 female teachers in the sample.

¹⁷ Prior research has shown that teacher experience and quality can have significant effects on student performance (Sanders, W. L. & Rivers, J. C. (1996). *Cumulative and Residual Effects of Teachers on Future Student Achievement*. Knoxville: University of Tennessee.)

Table 6. Teaching Experience

	Min.	Max.	Avg.
Years taught science in general	1	26	7.43
Years taught 8 th grade science	0	18	4.29
Years taught science in this school	1	15	4.64
Years taught 8 th grade science in this school	0	15	3.79
Number of other schools taught science	0	2	na
Number of other schools taught 8 th grade science	0	2	na

Science Curriculum and Resources

Both control and treatment group teachers use a modular science program. The control group curriculum consists of 1999 and 2000 editions of the PH *Science Explorer* curriculum.¹⁸ The 2005 edition of *Science Explorer* used by the treatment group has several major differences, including integrated math and reading support activities, more scaffolded questions, increased differentiated instruction, and more ancillary resources, including lab videos and online resources.

Control group teachers reported extensive use of older school-created resources, including lab activities and lesson plans. Treatment group teachers, on the other hand, relied almost exclusively on the 2005 *Science Explorer* program, and ranged in their use of supplemental materials. For more detailed information on their use of program resources, see Appendix C. Treatment teachers occasionally augmented the PH program with additional teacher

¹⁸ The school also has a very small number of 2002 student editions for one module, ordered to cover losses.

demonstrations and hands-on activities, especially when teaching density.

All teachers must follow district-issued pacing guidelines, which provide a window of time in which each module must be covered. These guidelines allow teachers a great deal of flexibility when planning lessons within a module. Both treatment and control teachers reported that they referred to pacing guides from their textbooks, but preferred to pace classes based on their assessments of student understanding and the framework provided by the district.

The school has one science lab available for 8th-grade science. Teachers typically accessed the lab once per block cycle with each of their classes. The lab appeared to be relatively well stocked, but students typically had to work in small groups of 2-4 to use lab equipment.

In addition the school has three science-related partnerships with the Audubon Society (for environmental studies), a local science museum (for physics), and a local foundation (to build science inquiry skills.) A local university also offers teachers professional development in mathematics.

Science Instructional Practices and Strategies

As noted above, science classes were either 30-40 minutes or 60 minutes, depending on the block day. Teachers varied in their individual instruction approaches, but several commonalities were noted among all teachers, including:

- Lecture classes, when new material is taught, begin with a warm-up activity or teacher demonstration.

- All teachers tried to provide real-world connections to concepts taught, though they varied in how often they did so.
- Teachers augmented lecture classes, which can occur 1-4 times per week, with lab activities, demonstrations, long-term projects, and other hands-on or exploration activities.
- Teachers assigned approximately 60 minutes of homework and 60 minutes of in-class independent practice each week, on average.
- Homework was assigned three times per week on average, though one treatment and one control teacher reported assigning homework almost nightly.
- Treatment group teachers reported using the 2005 PH textbook and student workbook for assignments, while control group teachers drew from the textbook, departmental worksheets, other supplements, or self-created materials. A range of homework problems were assigned, including guided reading and note taking, multiple choice questions, short answer problems, extended response questions, lab reports, and long-term chapter project work.
- All teachers reported checking homework for completion, and most would grade homework on occasion.
- All lab reports and long-term projects were graded.
- All teachers used informal assessment on a regular basis, typically through oral questioning and independent practice checks. They varied in their use of formal assessment.

Note that, since the focus of this study is an entire science curriculum (and this curriculum must address the local scope and sequence of the district) it is expected that there will be some overlap between what occurs in treatment and control classrooms. Data is collected on what occurs in treatment and control classrooms so that this information can be used to help interpret quantitative results.

Comparison of Science Instructional Practices Between Treatment and Control Teachers

Treatment and control teachers covered similar content during the study period, with some variance in one control class where a teacher was replaced mid-way through the year. This was to be expected given that teachers had to follow district curriculum guidelines. In terms of materials used, treatment teachers relied almost exclusively on 2005 *Science Explorer*, drawing on outside resources mostly to provide additional demonstrations and hands-on activities.¹⁹ Control teachers reported using department-created resources and other supplemental programs with much greater frequency.

In terms of fidelity of implementation, treatment group teachers mostly adhered to the implementation guidelines provided (see Appendix A). Three of four teachers reported “mostly” following the implementation guidelines closely, while one “somewhat” followed them closely. All teachers tried to use each required component of the program, though there was some variation in the frequency of use, especially for one teacher. For example, not all teachers assigned a long-term chapter project for each chapter, or assigned the suggested 30 minutes of independent practice per day. Teachers also varied greatly in their use of *Math: Analyzing Data* and *Writing in Science* activities. The largest departure from the implementation guidelines involved the use of the program pacing guides. All treatment teachers noted that they primarily paced their classes on the needs of their students and the district curriculum guidelines. Only one reported

¹⁹ However, all treatment teachers reported using outside resources significantly when teaching density, as they felt that 2005 *Science Explorer* did not offer enough content on the topic.

referring to pacing guidelines with regularity.

While both groups exhibited many similarities in classroom/instructional practices, some differences were observed. Based on analyses of weekly teacher logs, teacher surveys, and classroom observations, treatment teachers were slightly more likely than control teachers to regularly spend time on higher level skills and activities, including student formulation of hypotheses, student writing about observations, observation of teacher-led demonstrations, and hands-on student activities, such as labs. In contrast, control teachers were more likely than treatment teachers to regularly spend time on review, individualized instruction, and assistance to students with limited reading and writing abilities. Many of these differences may be attributable to the fact that control teachers taught immersion and special-education classes that required more re-teaching and individualized assistance.

While the sample of teachers is too small to make any definitive conclusions about the impact of the 2005 *Science Explorer* program on teaching practices and vice versa, it will be important to measure this and take into consideration similarities (and differences) in strategies employed by treatment and control teachers during the 2005-6 RCT.

Table 7. Percent of Time Spent (Weekly) on Each Class Activity by Group

Activity	Control	TX
Lecture to Class	18%	23%
Experiments/Labs	13%	23%
Review Homework	10%	8%
Tests/Quizzes	8%	6%
Reteaching/Reviewing	17%	10%
Video/Multimedia Presentations	4%	4%
Student Independent Work Time	20%	21%
Classroom Management Activities	4%	3%
Other	15%	21%

Based on analysis of teacher logs, classroom observations and discussions with teachers and staff, there was no evidence of contamination between control and treatment teachers. The teachers clearly understood our discussion of contamination and the department chair helped ensure that contamination did not occur through tight control of program materials.

Treatment teachers mostly implemented the 2005 Science Explorer program with fidelity. In addition, there was no evidence of contamination.

Results

This section is organized by the key pilot questions and reviews major findings first, followed by a more detailed account of the results.

Major Findings

Again, it is important to note that the primary purpose of the pilot study is to inform the design of the 2005-6 RCT and that the major findings presented below do not provide conclusive evidence on the effectiveness of 2005 *Science Explorer*.

- What did the users of the Prentice Hall 2005 Science Explorer think of it?

Treatment group teachers were more satisfied with the 2005 *Science Explorer* program than control group teachers were with their curriculum. Treatment teachers generally found the 2005 *Science Explorer* program to be an effective pedagogical tool, and were impressed with many of the ancillary resources. Only a few weaknesses were noted. Students using 2005 *Science Explorer* reported greater satisfaction with their curriculum than students using the control texts.

- Were any positive effects observed from the use of 2005 Science Explorer?

Users of 2005 *Science Explorer* demonstrated significant growth in performance on the TIMSS-F test (overall, Life Science, and Physical Science). Though there was no significant difference between control and treatment group performance as a whole, female treatment group students had better Total and Physical Science TIMSS-F scores than female control group students. Additionally, low socio-economic status students from the

treatment group outperformed their counterparts in the control group on Total, Life Science, and Physical Science scores of the TIMSS-F.

- Which instruments will best measure the effects of 2005 Science Explorer?

Based on our analyses of pilot study data, two assessments, the Iowa Test of Basic Skills Complete Battery (science portion only) and the TIMSS-F, are recommended for the RCT to increase the sensitivity of the study in detecting differences in student performance. Both instruments demonstrate sufficient reliability and validity and have good item difficulty ratings.

Teacher and student surveys employed in the pilot study are reliable and will provide valuable information during the RCT on teacher and student background, classroom practices, student and teacher opinions, and other potential outcomes stemming from use of 2005 *Science Explorer*.

- What are the key characteristics of 2005 Science Explorer training and implementation?

Implementation guidelines developed by PRES Associates and Prentice Hall proved to provide sufficient direction for teachers in identifying the necessary practices needed to use the 2005 *Science Explorer* with fidelity. Training sessions given by Prentice Hall staff provided teachers with a thorough understanding of 2005 *Science Explorer* materials available to them.²⁰ Monitoring tools, including online teacher activity logs, classroom observations, and implementation checklists, let

²⁰ It is important to note that due to the gradual release of 2005 *Science Explorer* components over the course of the 2004-5 school year, treatment teachers were not able to receive hands-on training or use all ancillary components (particularly technological resources).

researchers measure the extent to which treatment teachers employed the suggested instruction model and helped monitor for evidence of contamination.

Detailed Findings

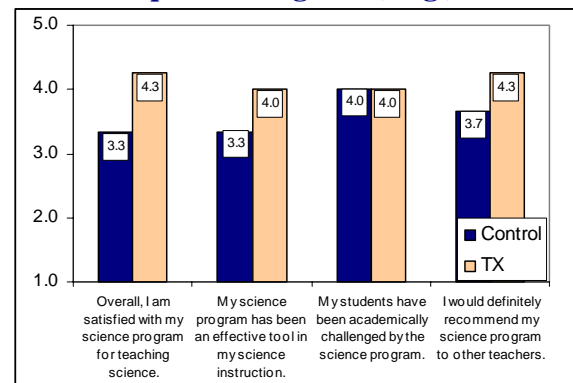
What did the users of the Prentice Hall 2005 *Science Explorer* think of it?

Information obtained from interviews and surveys were analyzed to obtain preliminary information on what users thought of the Prentice Hall 2005 *Science Explorer* program.

Preliminary Teacher Findings

As shown in Figure 1, treatment teachers enjoyed using the program and believed it was an effective tool for science instruction. In addition, their satisfaction with the 2005 *Science Explorer* program was greater than control teachers' satisfaction with their program, $p=.06$. When asked what specific components they liked best, they noted the *All-In-One* teacher's resources, *Exam View Test Generator*, student workbooks (especially the *Guided Reading and Study Workbook*), lab and field trip videos, and *Section Assessment* questions. Other teacher ratings of the usefulness of the 2005 *Science Explorer* components can be found in Appendix B.

Figure 1. Teacher Attitudes on the 2005 *Science Explorer* Program (Avg.)*



* N=4

Treatment teachers liked the 2005 Science Explorer program more than control teachers liked their curriculum. Treatment teachers felt that the program helped them assess student knowledge, individualize instruction, provide intervention to those who needed it, and make connections between science and the real world.

The teachers noted that 2005 *Science Explorer* aided their own pedagogical practices as well. For instance, they reported that the program helped them: assess student knowledge on an on-going basis; individualize instruction to average students (and, to a lesser extent, to below average and advanced students); provide intervention to those who needed it; and make connections between science and the real world. They also noted that the program saved them time in lesson preparation, helped them plan and execute lab activities, and helped them select and assign independent practice.

Comparisons were also made between practices employed and knowledge or preparation for effective teaching strategies at pre-test and post-test. Results showed no significant differences among treatment

teachers, $p > .05$. This is not unexpected given the limited time teachers used the program. During the RCT, we will examine if the program results in greater knowledge and use of effective teaching practices in science education.

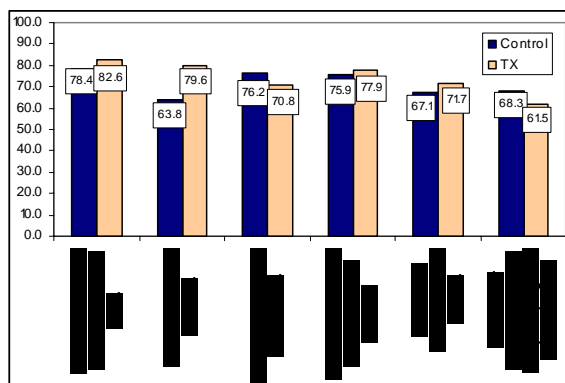
Preliminary Student Findings

Treatment teachers reported that the program was a useful tool for helping their students:

- develop reading and writing skills relevant to science;
- make real-world connections to lessons taught;
- prepare for tests;
- increase higher-order cognitive skills; and
- increase overall science knowledge.

As shown in Figure 2, over 80% of the treatment students were satisfied with their science book. In addition, there was a statistically significant difference between treatment and control students in terms of their satisfaction with their science textbook, $p < .05$. There were no other significant differences.

Figure 2. Student Attitudes on Program Components (Percent Who Were Satisfied)



Treatment students also used the technology provided with the program to some extent (a few times during the semester). This included the *Student Edition CD-ROM* and the *Science Explorer* website (*PHSchool.com* or *Go Online* links from textbook). For more information on the technology used in treatment science classes, please see Appendix C.

Weaknesses

One major complaint noted by treatment teachers was that the consumable and non-consumable kits lacked materials that were supposed to be supplied (as described in materials lists in Teacher’s Editions). Teachers also noted that several lab and demonstration activities did not turn out as planned, and wanted more specific instructions for teacher demonstrations. In addition, some teachers noted that the program needs *more* workbooks, activities, review questions, real-world applications and quizzes, and most importantly, options for teacher demonstrations. All treatment teachers stated that the density lessons provided in the new text were insufficient when compared to the old curriculum. All teachers reported augmenting their density lessons significantly.

Additionally, some of the treatment teachers found some of the content and language used to be inappropriate for 8th graders. As one teacher noted, “there are problems with the language used in the text - - at times it is too technical [especially in definitions of new vocabulary words]. It’s not 8th-grade English.” One teacher also observed that many online links provided in the student edition seemed intended for teachers, and wanted more references to resources (either online or other reference texts) that could *support* students as they worked.

Teachers observed some weaknesses with the program, particularly with the contents of materials kits, a lack of “8th-grade English” in definitions, insufficient directions for teacher demonstrations, and too little emphasis on density.

Were any positive effects observed from the use of 2005 Science Explorer?

Although no conclusive findings about the effectiveness of the program are possible given the nature of the pilot study, analyses were performed to provide preliminary, descriptive information on the relationship between the PH Science Explorer program and student science performance. In addition, it is important to note that treatment teachers used only the *Chemical Building Blocks*, *Chemical Interactions*, and *Motion, Forces, and Energy* modules from the Physical Science series during the study period. Consequently, any relationship between the program and student science performance is likely to be strongest in the area of physical science.

Comparison of Pre- and Post-test Among Treatment Students

Preliminary analysis was performed to determine whether there was growth among treatment students as measured by the ITBS and TIMSS-RE science assessments. Results showed significant change among students as measured by the TIMSS-F Total Score and its content domains: Life Science, Physical Science, and Earth Science, $t_{(140)} = 3.29, p < .001, d = .78$, LS $t_{(140)} = 3.05, p < .001, d = .52$, PS $t_{(140)} = 5.38, p < .001, d = .91$, ES $t_{(140)} = -3.06, p < .001, d = .52$. However, the change in Earth Science performance was negative (indicating a decrease in achievement) whereas all the areas showed a positive growth. This is likely the result of

the greater emphasis placed on Physical Science during the last 5 months of science instruction. Additionally, it is important to note that these content domain analyses are only descriptive (and will be for the RCT as well) given the low reliabilities found for the content domains. There was no significant difference as measured by the ITBS-science test, $t_{(128)} = .97, p = .34, d = .17$. With the exception of the ITBS test, the moderate to large effect sizes found are of note.

Figure 3. Science Performance Change Among Treatment Students

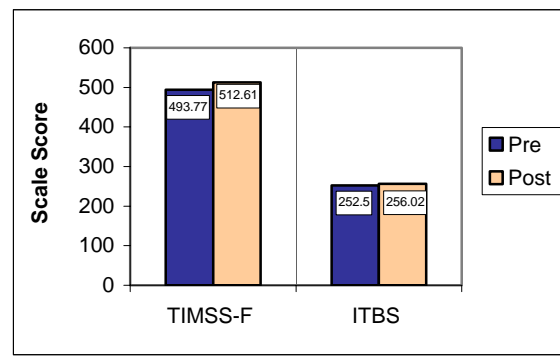
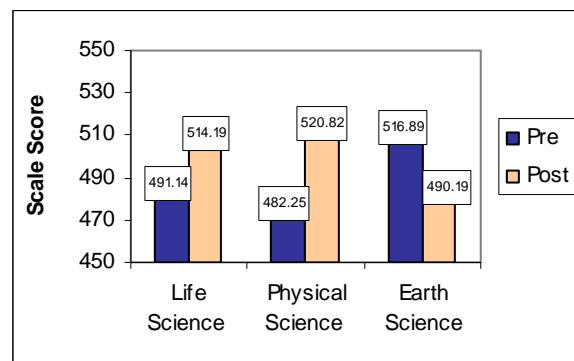


Figure 4. Science Content Domain (TIMSS-F) Change Among Treatment Students



Comparison of Treatment and Control Students

Preliminary analysis showed no significant differences between treatment and control students at post-test, after controlling for pre-test scores, as measured

by both the TIMSS-F and ITBS science tests. In addition, although treatment students tended to perform better than control students, these differences were not significant as measured by the TIMSS-F content domains. The range of effect sizes found in these analyses was .01 to .23.

Figure 5. Science Performance Between Treatment and Control Students

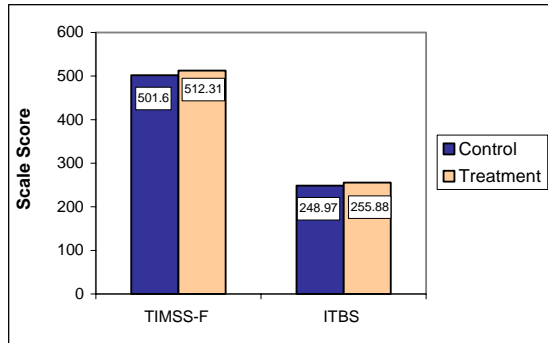
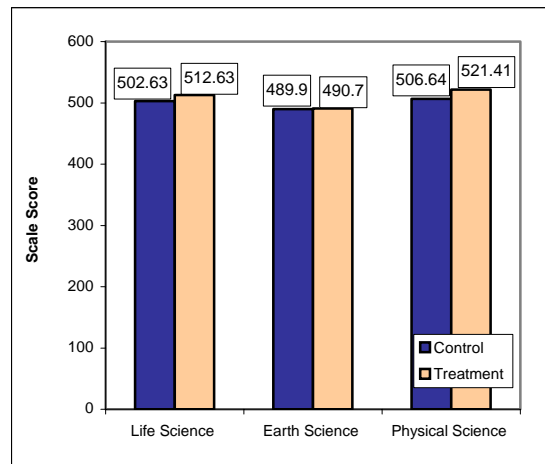


Figure 6. Science Content Domain (TIMSS-F) Performance Between Treatment and Control Students



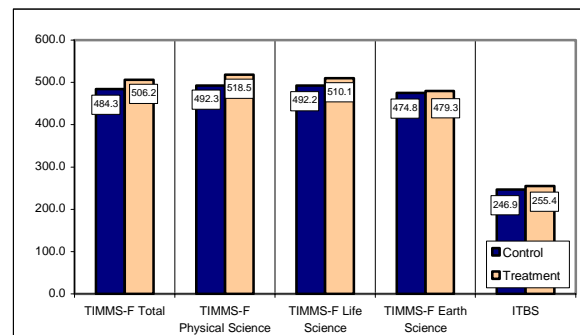
Conclusive findings about the relative effectiveness of the Science Explorer program are not possible given the nature of the pilot study. Nevertheless, preliminary analyses of pilot data revealed significant changes in performance among treatment students as measured by the TIMSS-F. However, no significant differences were found between the treatment and control groups.

Subgroup Analyses

Subgroup analyses were conducted to obtain preliminary information on the relative effectiveness of the program with special populations. These analyses were limited to gender and socioeconomic status only due to the limited number of special education and Limited English Proficiency students in the sample.

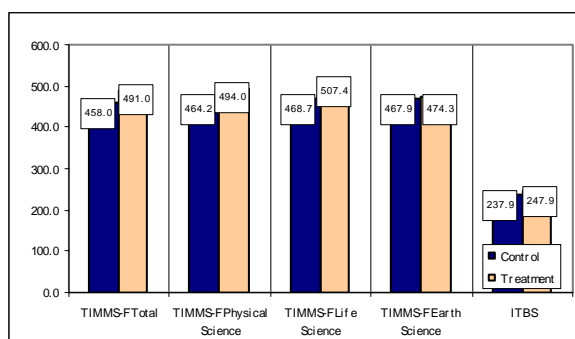
As shown in the figure below, among females there was a *marginally* significant difference between treatment and control students; treatment students tended to perform better on the TIMSS-F Total science test and TIMSS-F Physical Science content domain, $TIMSS-F F(1, 106) = 3.13 p = .08, d = .12$, and $PS F(1, 106) = 2.70 p = .10, d = .10$. There were no other significant differences as measured by the remaining tests.

Figure 7. Science Performance Between Female Treatment and Control Students



Among students of low socioeconomic status (SES), results again showed a significant difference between treatment and control students. Specifically, low SES treatment students performed better than low SES control students on the TIMSS-F Total and Life Science portion of the science test, TIMSS-F $F(1, 76)= 5.57$ $p=.02$, $d= .29$, and LS $F(1, 76)= 4.75$ $p=.03$, $d= .25$. The difference between the two groups on the TIMSS-F Physical Science test was *marginally* significant, $PS F(1, 76)= 3.13$ $p=.08$, $d= .16$. There were no other differences observed.

Figure 8. Science Performance Between Low SES Treatment and Control Students



Subgroup analyses showed a positive relationship between use of the program and improved performance on the TIMSS-F test among females and students of low socioeconomic status.

Effect Size Analysis

As previously noted, the effect sizes found between the treatment and control groups were small (.01-.29). This is not unusual given that the teachers used this program for about 5 months. Typically, effect sizes for educational programs range from small to moderate (or .20-.50).

During the RCT, it is very probable that these data will be analyzed via Hierarchical Linear Modeling (HLM) in order to account for the likely dependency issues in data collected in educational settings. That is, because the performance of students within a class are likely to be more similar to each other than the performance of students between classes, this violates the critical assumption of independence of observations used in General Linear Modeling (i.e., ANOVA, ANCOVA, etc.) of student-level data. Therefore, HLM (or some variation thereof) needs to be used in order to be in accordance with WWC Statistical Analyses criteria.

HLM requires that there be a sufficient number of level 2 subjects (or teachers in our case). Based on information reviewed on HLM power analyses, it appears that with the inclusion of covariates (such as socioeconomic status and pretest scores), the RCT will require at least 35 teachers (with at least 4 students each) to have a power of .50 to detect a small effect size of .30.²¹

²¹ This assumes an intraclass correlation (proportion of variance in the outcome measure that is between groups; a measure of dependency) of .50. As noted by Bloom (1995) this is a reasonable estimate based on prior educational research. Note also that this still does not ensure a power of at least .80, which is optimal. Due to fiscal limitations, the RCT is designed to include only 24 teachers. However, based on these results it is recommended that at least 35 teachers be included.

Which assessments will best measure the effects of **Science Explorer**?

Multiple assessments are recommended for the RCT. This will help increase the sensitivity of the study in detecting differences. The two assessments recommended are the Iowa Test of Basic Skills and the TIMSS-F. These tests place emphasis on different constructs, with the ITBS more focused on scientific process (e.g., conducting research, data analysis) and the TIMSS-F equally dealing with factual knowledge, conceptual understanding, and reasoning and analysis.

Do the measures demonstrate validity, reliability, and sensitivity to detect differences?

In order to increase the sensitivity of the RCT to detect differences, and increase the validity of the results (e.g., if positive findings are found across different measures), a battery of outcome measures is recommended. These are discussed in the following section.

Assessments

Iowa Test of Basic Skills (Complete Battery - Form A) - Level 14, Science Portions Only (ITBS): Science content of this widely used and recognized 43-item norm-referenced, selected response test is aligned to NSTA and AAAS standards. The science portion of Level 14 reports high reliability (.89 for Fall norms and .90 for Spring norms) and good item difficulty scores (mean p (fall) = .52, mean p (spring) = .55) indicating it is sensitive to the wide-range of ability levels evident in classrooms across the nation.

Assessment Derived from the 2003 Trends in International Math and Science Study 8th Grade Science Assessment (TIMSS-F):

The TIMSS 8th Grade Science Assessment was administered to over 215,000 students worldwide in 2003, including 8900 in the United States. It is used to compare student performance on science internationally. In the 2003 administration in the United States, the TIMSS 8th Grade science assessment displayed an internal reliability coefficient of .88.

Based on data from the pre- and post-test administration of the TIMSS-derived assessment during the pilot study, PRES Associates finalized a single version of the assessment to be used during the 2005-6 RCT. Psychometric properties of the TIMSS-F are summarized in the table below. For more detailed information on the development of this test and its psychometric properties, see Appendix D. As shown, the TIMSS-F is a psychometrically sound instrument, with sufficient reliability and validity. To review a copy of the TIMSS-F, contact PRES Associates.

Table 8. Psychometric Properties of the TIMSS-F

	Alpha (Reliability)	
	Pre	Post
Total TIMSS-F	.77	.80
Life Science	.52	.56
Earth Science	.54	.46
Physical Science	.55	.68
	Concurrent Validity (Correlation (r) with ITBS)	
Total TIMSS-F	0.65	0.70
Life Science	0.47	0.62
Earth Science	0.50	0.54
Physical Science	0.53	0.58

However, it is important to note that the TIMSS-F is *unidimensional* (i.e., generally, it measures one overall construct). Factor analyses showed that the first factor accounted for 63% of total variability. In addition, the low reliability values observed for the content domain areas also suggest that this measure is unidimensional and as such, RCT assessment analyses should be done using the total TIMSS-F score. Analysis by content domain would be used during the RCT for descriptive purposes only.

Student Surveys

A total of 282 (pre) and 280 (post) students completed the student survey (99% and 98% completion rate, respectively²²). Reliability analyses showed that the deletion of two items²³ would significantly improve upon the reliability of the student survey. With this change, overall reliability analysis revealed that the portion of the student survey measuring school and science-related attitudes (the top four listed in Table 9) had a high internal consistency²⁴ value (.87). The subscales revealed moderate to high internal consistency values. A value of .70 and greater is viewed as adequate; all subscales meet this criterion. To review a copy of the survey, please contact PRES Associates.

²² The less than 100% completion rate is the result of student absences. Attendance will be measured and taken into account during the RCT. In addition, during the RCT we will request that students absent on the day of administration be given the measure when they return to class.

²³ The items “I learn things quickly in science” and “An adult at home expects me to do well in school” will be removed from the final version of the student survey.

²⁴ Internal consistency values (Chronbach’s alpha) is a measure of reliability. Reliability refers to the property of a measurement instrument that causes it to give similar results for similar inputs. The range of this value is 0 (no relationship) to 1 (complete relationship).

Table 9. Reliability of Student Survey Subscales

Subscales in Student Survey	Reliability
Enjoyment of Science	.82
Perceived Science Ability	.81
Importance of Science	.76
School and Science Motivation and Effort	.79
Parental Knowledge and Support	.75

The student survey is reliable and can be used to measure other potential outcomes of the Science Explorer program.

Teacher Surveys

The overall reliability of the portion of the survey measuring attitudes about effective teaching and learning practices showed a high internal consistency value of .80. Internal consistency of the sub-scales are provided in the following table. To review a copy of the survey, please contact PRES Associates.

Table 10. Reliability of Teacher Survey Subscales

Subscales in Student Survey	Reliability
Importance of Effective Science Instruction	.75
Attitudes about Learning Science	.77
Limitations in Teaching as a Result of Student Characteristics/Behavior	.89

Teacher surveys are an effective tool for the collection of information on teacher background, classroom practices, knowledge, and opinions.

What are the key characteristics of 2005 *Science Explorer* training and implementation?

To ensure that all treatment teachers participating in the study had sufficient knowledge and skills to successfully implement this program from the beginning of the study period, teachers were given implementation guidelines and provided training prior to implementation. In addition, monitoring procedures (via weekly instructional logs completed by teachers and classroom observations) were developed to measure the extent to which treatment teachers were implementing a similar instructional model. Because Prentice Hall released 2005 *Science Explorer* materials to the school gradually over the study period, treatment teachers were not able to use some of program materials at all, and use of others was limited by time²⁵. This was especially true of technological components.

Implementation

Through guidance and consultation with Prentice Hall staff, key components of the 2005 *Science Explorer* program were identified. Based on this information, implementation guidelines were developed for use by the treatment teachers. Teachers were provided with the guidelines during the first training session and instructed to follow the guidelines while using the program. Identified key components of the 2005 *Science Explorer* program include the following:

- Follow pacing guidelines
- Chapter science project (each chapter)
- Lab activities (at least one demonstration or hands-on activity for every chapter)

²⁵ Technological components, Spanish-language materials, and some supplemental teaching resources were not released by Prentice Hall until late Spring 2005.

- Pre-teach activity (each section)
- Check understanding (throughout each section)
- Assess knowledge (at the end of each chapter)
- Close the lesson (each lesson)
- Independent practice (30 minutes each school night, as feasible)
- *Math: Analyzing Data and Writing in Science* (each section)
- Differentiated instruction (throughout sections, as needed)

Please see Appendix A to review the implementation guidelines.

Teachers felt that the implementation guidelines were clear and provided guidance on the key components of the program and study procedures. However, after the initial training session, they generally only felt the need to refer to them occasionally.

Training

In order to implement this program successfully, it is essential that teachers have a thorough understanding of the materials provided by the 2005 *Science Explorer* program. Therefore, in November 2004, the treatment teachers met with a Prentice Hall professional trainer for approximately 4 hours prior to implementation of the program in their classes. In addition, a follow-up session (approximately 3 hours) was conducted at the end of January 2005. During the initial training session, the trainer described the key components of the 2005 *Science Explorer* program, went over the Teacher's Edition and available ancillary resources, reviewed resources available through the PHSchool.com website, and offered examples of when to use certain

materials. The follow-up training was more informal, providing teachers with the opportunity to review issues encountered during implementation with the trainer and to receive more training on using the website resources and the test generator. Representatives of PRES Associates attended both trainings.

Feedback provided on the training by the teachers was generally positive. It should be noted that training was not provided on general effective teaching strategies or on science content but rather on how to use the 2005 *Science Explorer* program (i.e. what is available and when/how to use). The treatment teachers felt that the training provided was sufficient. It should also be noted that because the teachers previously used an earlier version of *Science Explorer* and used 2004 Prentice Hall products for mathematics, they were already familiar with the general organization of the curriculum and some of the technological resources offered by Prentice Hall. As such, they may have required less training than other teachers unfamiliar with Prentice Hall products may require.

The following are key aspects of the training that should be incorporated for the RCT.

- Training on the technology available and website should be conducted such that each teacher is able to work with his/her own computer and software (e.g., in a computer lab) to the extent that this is feasible.
- Training should begin with a thorough discussion of the program's research base and philosophy. This will help set the stage for how the program works to improve student science skills.
- Trainers should walk teachers through the typical instruction of a chapter while being open to modifications according to teachers' needs or preferences.
- All key program components (and resources available to meet the key component) should be thoroughly discussed during the above demonstration.
- To the extent that it is feasible, initial training should occur 2 weeks prior to the start of school and teachers should have access to the materials ***1-2 months prior*** to this training so that they can review the materials.

These changes will help ensure that teachers are ready to begin implementing the *Science Explorer* program effectively at the start of the school year.

Trainings should include a thorough discussion of the program's research-base and philosophy, walking teachers through the instruction of a chapter while point out all key components, and providing materials 1-2 months prior to the start of school.

Monitoring and Fidelity of Implementation

To determine the extent to which treatment teachers implemented the program with fidelity, and to examine whether contamination occurred between control and treatment teachers (via sharing of 2005 *Science Explorer* resources), teacher log/implementation checklists, classroom observations and information obtained from the interviews were analyzed. As previously noted, teachers mostly implemented the *Science Explorer* program with fidelity and no contamination was observed. It should be noted that information obtained from these three sources were highly consistent. That is, for the most part, our observations concurred with what was being reported by the teachers.²⁶

Do the monitoring procedures capture what is occurring in teachers' classrooms?

Teacher Logs – Teacher logs were developed so that program implementation could be monitored on a weekly basis. Average completion rate for the 20-week period in which teachers were asked to complete logs was 79% (range was 60% to 100%). Other studies conducted by PRES Associates using a very similar template have shown that the web-based data collection is suitable for providing information on what is occurring in each teacher's classroom, including implementation of their program, use of technology and exercises (including homework and independent practice), and other resources used during the class week. However, it also has become apparent from these studies that completion of teacher logs on a weekly basis can be problematic for

many teachers, even with frequent reminders from PRES Associates and school staff members. Completion rates of less than 75% can produce unreliable information. Analyses by PRES Associates has shown that it will be sufficient to gather teacher logs on a monthly basis, which should reduce the burden on teachers. In addition to teacher logs, an implementation checklist was developed so that teachers could indicate the percent of lessons taught from their science program. Treatment teachers were also asked to note which *Science Explorer* lessons were taught. This additional method helps ensure that we can measure implementation fidelity among treatment teachers and can compare the extent of program implementation between control and treatment teachers (thereby meeting WWC guidelines).

Researchers obtained feedback on the teacher logs and implementation checklist to assess the ease of use by the teachers. While there were some complaints about the length of the online teacher log, most teachers observed that the interface was easy to use. No comments were noted about the implementation checklist (though one module used by the treatment group teachers was inadvertently omitted, forcing teachers to write in chapters used from that module). However, it is important to note that implementation checklists will have to be developed for each participating site in the RCT due to the modular nature of the *Science Explorer* program.

The following should be incorporated into the implementation log for the upcoming year.

²⁶ Our prior research has shown a high agreement on implementation ratings between classroom observations and logs (94%).

- Completion of teacher logs should be collected monthly.²⁷ Items will be modified to request averages for the month or incorporate a scale relating to frequency of use during the past month (e.g., daily, 2-3 times per week, once a week, 2-3 times per month, once a month, not used at all).
- Items relating to implementation of key components by treatment teachers should directly reflect the criteria set forth in the implementation guidelines.

These changes should reduce time demands and should, therefore, increase completion rates from teachers who are reluctant to complete these due to lack of time.

Teacher logs should be completed monthly as opposed to weekly to increase teacher completion rates. In addition, the RCT should employ the use of implementation checklists to account for teachers who do not complete logs on a regular basis.

Classroom Observation – Two classroom observations took place during the pilot, in January and April 2005. An evaluator observed a science class session for each teacher. Results showed that the instrument is efficient in capturing what is occurring in the classroom and easy to use. To review a copy of the observation protocol, please contact PRES Associates.

In planning site visits for the 2005-6 RCT, efforts will be made to ensure that observers see classes where students are learning new material, as opposed to review or laboratory classes. Based on the experiences of the PRES evaluator and feedback from the Prentice Hall trainer, it was decided that such introductory classes provide the best picture of what the classroom environment is like.

The teachers indicated that they did not mind the intrusion of the observations and that they were not greatly affected by our presence, though some noted that the presence of an observer affected the behavior of some students.

²⁷ This information has been collected monthly in other PRES evaluation studies and we have found that it does not negatively impact the validity of the data.

Conclusions

Project Summary

The 2005 *Science Explorer* pilot study served to develop and test various procedures and instruments that will be used during the RCT. In addition, it provided valuable information on issues that may affect the full-year study. In this section, study accomplishments are summarized as are the lessons learned from the study and their implications for the RCT.

Accomplishments

- We finalized all instruments and assessments to be used for the RCT. After modifications, all quantitative measures demonstrated good reliability and validity. Effect size analysis showed that at least 35 teachers will ideally be selected for participation in the RCT.
- Implementation guidelines and monitoring procedures to be employed during the RCT were shown to be effective.
- We identified training requirements for the treatment teachers.
- We were able to draw some tentative findings regarding the utility and efficacy of the 2005 *Science Explorer* program. Students using 2005 *Science Explorer* demonstrated growth on the overall TIMSS score, and some positive differences were noted between some subgroups of treatment group students when compared to control group students. Both teachers and students using 2005 *Science Explorer* expressed

greater satisfaction with their materials than those using the control program.

Lessons Learned and Implications for RCT

The Spring 2005 Pilot Study on the Effectiveness of the 2005 Prentice Hall *Science Explorer* provided PRES Associates with valuable information for the RCT.

- Treatment teachers mostly implemented the 2005 *Science Explorer* program with fidelity. In addition, no contamination was observed. It will be important for the RCT to continue to monitor classroom activities for both treatment and control group teachers. PRES Associates will intervene with treatment group teachers who fail to implement the program with fidelity. A variety of measures, including classroom observations, monthly teacher logs, implementation checklists, and interviews will be used to monitor implementation and contamination.
- Differences and similarities in teaching styles and practices between treatment and control teachers as well as within the treatment group were observed. Therefore, it will be important to measure and take this into account during the RCT as it can have an impact on the effect sizes observed and will consequently assist us in interpreting the quantitative data. In addition, quantitative information on teachers' practices will be collected via the teacher survey. Additionally, if

major differences in teacher background and teaching preferences are observed between treatment and control groups, statistical controls will be employed to account for them.

- The student survey demonstrated a high level of reliability after two items were dropped. The portions of the teacher survey measuring teacher attitudes (as opposed to background information) demonstrated high reliability as well.
- Both the ITBS and the TIMSS-F exhibited good reliability and validity, as well as good item difficulty. Based on teacher feedback, the original TIMSS-derived test was modified to minimize administration time. We learned that the TIMSS-F is unidimensional and that analysis by content domain subscores should be used for descriptive purposes only.
- Based on feedback from participating teachers and our experience with other studies, online teacher logs will only be required monthly as opposed to weekly to increase response rates. An implementation checklist will be provided to teachers at the conclusion of the RCT to help verify information provided in the logs and serve as proxies for teachers with low teacher log response rates.
- The classroom observation protocol proved to be a useful tool for monitoring science class activities. Classroom observations should occur during classes in which new material is taught, as opposed to lab or review classes.
- Triangulation of information by using multiple sources (researchers, teachers *and* students) is important for the RCT.

This will ensure the accuracy of observations.

- Treatment teachers did not feel overly restricted by the implementation guidelines. In addition, they felt that the guidelines adequately informed them of what they had to implement as part of the study and provided them with an overview of the 2005 *Science Explorer* program. However, they only felt the need to occasionally refer to the document after receiving training.
- Treatment teachers felt that the training was sufficient. However, based on our experiences with this study and others, we feel that trainings should include a thorough discussion of the program's research-base and philosophy, walking teachers through the instruction of a chapter while pointing out all key components. Most importantly, teachers should be provided with ***all materials 1-2 months prior to the start of school*** so they have a chance to familiarize themselves with the product and its components.

In conclusion, the pilot study has enabled PRES Associates to design a 2005-6 RCT that will ensure that the effects of 2005 *Science Explorer* can be reliably and effectively measured in keeping with the standards established by the WWC.

Appendix A:
Pilot Study Implementation Guidelines

2005 PRENTICE HALL SCIENCE EXPLORER PILOT STUDY **IMPLEMENTATION GUIDELINES FOR TEACHERS**

INTRODUCTION

Welcome, and thank you for participating in the Prentice Hall *2005 Science Explorer* Pilot Study. We believe your experience with our study will be rewarding and enjoyable. Not only will you contribute to cutting-edge research, but you will also benefit from first-rate professional development provided by Prentice Hall professional training specialists.

We understand that it may be challenging to change former practices and implement a new science program. Therefore, we greatly appreciate the time and effort you will be putting into making this study a success. However, we also realize that there will be obstacles and challenges as you begin to implement this program. Under these circumstances, we *want and need* to hear from you; we will make every attempt to guide you through those challenges. In fact, it is critical that any problems you encounter be addressed as soon as possible to ensure that this program is being implemented to its full potential. Feel free to contact Amy Blake, Research Assistant for PRES Associates at (307) 733-3255 or ablake@presassociates.com if you have any questions, problems, concerns and so forth.

The following provides answers to some common questions teachers may have related to this study. Please read through all of these questions/answers. Should you have further questions, please contact PRES Associates.

Why Is This Research Being Done?

As you are aware, the No-Child Left Behind (NCLB) of 2001 requires that educational materials and strategies used by educators in the classroom *must be proven by scientific research to improve student achievement in the classroom*. Prentice Hall has developed a strong research model for determining that their programs are scientifically-based and successful. As part of this ambitious research agenda, Prentice Hall has contracted with Claremont Graduate University, in conjunction with PRES Associates²⁸, an external educational research firm, to conduct a rigorous quantitative research study on the effectiveness of the Prentice Hall *2005 Science Explorer*. This Pilot Study will enable researchers to design an effective, yearlong randomized control trial to be conducted in 2005-6. Both studies will contribute to the growing research base behind Prentice Hall *2005 Science Explorer* and the effectiveness of different approaches to science instruction.

Why Do I Need Professional Development?

It takes more than a good curricular program to raise students' knowledge of science. It also takes good teachers with a thorough understanding of the curriculum and who are supported by professional development, school administrators, and parents/guardians. To this end, it is hoped that through the professional development training sessions provided by Prentice Hall on the use of its *2005 Science Explorer* program, all "treatment" teachers participating in the study will gain the knowledge and skills to successfully implement this program right from the start.

²⁸ PRES Associates is an external, independent, educational research firm with an established track record in conducting large-scale, rigorous evaluations on the effectiveness of research materials.

As you will soon learn, this science program provides numerous teaching resources and supports. In order to implement this program successfully, it is essential that teachers have a thorough understanding of the materials provided by the *2005 Science Explorer* program. Rather than having teachers figure it out on their own, professional trainers will guide you through this process, offering examples of when to use certain materials, how to manage and supplement classroom instruction, what types of assessments to administer, and so forth. In addition, training will be provided on the science technology available.

Why Do I Need To Follow These Implementation Guidelines?

The Teacher Implementation Guidelines were developed as part of the Prentice Hall *2005 Science Explorer* Pilot Study. The guidelines are designed for “treatment” teachers to use while implementing the new program. The guidelines point out key program components that *must* be implemented during science lessons. These key program components have the greatest influence on student learning and performance and therefore should be implemented. In addition, it is critical to ensure that all “treatment” teachers are implementing a similar instructional model. That is, if teachers are modifying the program to an extent that it no longer resembles the original program, the study becomes invalid. In sum, by providing these implementation guidelines, we are attempting to (1) maximize the potential of this science program, and (2) ensure that the program is being implemented with equal fidelity across teachers. To reiterate, *it is essential that all “treatment” teachers implement the program fully as prescribed in the following implementation guidelines.*

That said, we do not expect that all teachers will teach in the same style or manner, or even use all of the same ancillary program resources. We know that each teacher has different teaching preferences and different student needs. *We trust your professional judgment and ask that you try to implement the program as best you possibly can while meeting your instructional needs.*

Again, thank you for your participation in this study. You are an integral part of this study and we appreciate your assistance. We look forward to working with you.

2005 PRENTICE HALL SCIENCE EXPLORER

IMPLEMENTATION GUIDELINES

MATERIALS LIST

Please note that you will have numerous materials to draw from as you implement the program. We do not expect you to use all components. You will find that some materials work better for you and your classes than others, but we ask that you consider each program component and try them as feasible. Additionally, Prentice Hall has not yet released some of the components, and they will arrive over the course of the year as they become available (noted below). We appreciate your patience while waiting for these materials.

Materials that have been or will be provided to you for each *2005 Science Explorer Module* that you will teach from are as follows:

- Student Edition
- Teacher's Edition
- Guided Reading and Study Workbook
- All in One Teaching Resources
- Discovery Channel Videos
- Consumable Materials Kit
- Non-consumable Materials Kit
- Color Transparencies

As soon as they are available, you will also receive the following resources for all modules:

- Teacher Express CD-ROM (anticipated by March 2005)
- Presentation Express CD-ROM (anticipated by March 2005)
- Lab Zone Easy Planner CD-ROM (anticipated by April 2005)
- Adapted Reading and Study Workbook (anticipated by March 2005)
- Adapted Chapter and Unit Tests (anticipated by March 2005)
- Differentiated Instruction Guide to Labs and Activities (anticipated by February 2005)
- Student Edition on Audio CD (anticipated by December 2004)
- Spanish Student Edition (anticipated by Feb. or Mar. 2005)
- Spanish Guided Reading and Study Workbook (anticipated by Feb. or Mar. 2005)
- Spanish Teaching Guide with Tests (anticipated by Feb. or Mar. 2005)
- Interactive Textbook Online (anticipated by Feb 2005)
- ExamView CTB Test Generator CD-ROM
- Inquiry Skills Activity Books
- Lab Activity Videos
- Progress Monitoring Assessments
- Test Preparation Workbook
- Test-taking Tips with transparencies
- Inquiry Skills Activity Book III
- Reading Strategies for Science content
- Teacher's English Language Learner Handbook
- Teacher Online Access Pack
- Guided Reading and Study Workbook

Within each lesson, your TE will reference these resources when appropriate for use.

IMPLEMENTATION GUIDELINES

Please follow these guidelines as you implement the Prentice Hall *2005 Science Explorer* program. All of these checked items are considered critical to the success of the program.

- ✓ **PACING GUIDELINES.** In the TE for each *2005 Science Explorer* module you will find *Pacing Options*, which gives the suggested teaching time required for each chapter. ***Teachers should aim to teach each lesson within the time period suggested by the Pacing Options.*** However, these are simply guidelines and we realize that there will be variations in terms of how long you take teaching a particular chapter and which chapters you teach. The important thing is to progress through the texts in a somewhat timely manner.

- ✓ **CHAPTER SCIENCE PROJECT.** In order to develop inquiry and scientific laboratory skills, it is important to complete a science project for each unit/module. For each chapter taught, the *2005 Science Explorer* program offers a chapter project. This ongoing project, which students work on throughout the chapter activities, is an integral part of the program. Your *All in One Teaching Resources* provides a detailed Chapter project guide for each project, and includes teaching notes, 2 project worksheets, and a project scoring rubric. ***Students should at least one chapter project per module.***

- ✓ **LAB ACTIVITIES:** You can choose from six different lab activities that are available with each chapter. ***It is important that you provide students with at least ONE of these demonstrations or hands-on activities with each Chapter Section taught,*** as they encourage students to develop inquiry skills. Your Lab Zone Easy Planner CD-ROM, when available, will assist with lab planning. You will also be provided with Consumable and Non-Consumable materials kits to help you conduct these activities. Lab activities that are available for each chapter include:
 - **Ongoing**
 - *Chapter Project:* Provides opportunities for long-term inquiry

 - **Full Period** (some of these activities are available as labs your students can conduct or as Lab Zone Videos. They are designed to provide in-depth practice of inquiry skills or science concepts)
 - *Skills Lab*
 - *Technology Lab*
 - *Consumer Lab*
 - *Design Your Own Lab*

 - **Short Exercises**
 - *Skills Activity:* Directed activity that allows students to practice a specific skill
 - *Try This Activity:* Directed activity that reinforces key concepts
 - *Discover Activity:* Open-ended activity to explore lesson before reading.
 - *At Home Activities:* Quick, open-ended activities for home and family.
 - *Teacher Demo:* Designed for teachers to present concepts to students.

- ✓ **PRETEACH.** Teachers should introduce each Section in an engaging manner and provide some background on the topic to be presented. The *Build Background Knowledge* activities and *Discover Activities* provide good opportunities to do so, though you can use any other procedure or activity to meet this goal.

-
- ✓ **INTRODUCE KEY CONCEPTS AND TERMS.** Teachers must review all *Key Concepts* and *Key Terms* in each Chapter Section taught. You can teach/review these key words in any manner that you feel is appropriate (e.g. via lecture, exercises, etc.).
 - ✓ **CHECK UNDERSTANDING AND ASSESS KNOWLEDGE.** It is extremely important to determine students' level of understanding and to plan intervention accordingly prior to any formal assessment. You can check for student understanding by asking questions, providing a short quiz or exercises designed to check understanding. Within each Chapter Section, the program presents teachers with opportunities to check student understanding. For instance, the program offers checks for understanding for use in the middle of a lesson; these can be found in the TE under *Monitor Progress* and *Reading Checkpoint*. To assess understanding at the end of each lesson, the text offers *Reviewing Key Concepts* questions and *Performance Assessment* suggestions. These can be used as in-class group activities, in-class independent practice, or homework. ***Teachers should check understanding and assess knowledge throughout each Section.***

In addition, at the end of each chapter, ***teachers should formally assess students' knowledge of the material covered in the chapter.*** That is, some form of a chapter test should be given to students. This will help you determine how students are performing and their level of understanding. The *2005 Science Explorer* program offers various resources for assessing students at the end of a chapter, including Performance Assessments and Chapter tests in the *All In One Teaching Resources* book, or you can create your own by picking test items provided with your PH program materials. It is strongly recommended, but not required, that teachers use some of the *Standardized Test Prep* questions at the end of the chapter.

- ✓ **CLOSING THE LESSON.** Teachers should allow for time at the end of the lesson so that students and/or the teacher can bring an appropriate level of closure to the lesson. This may include students thinking about and discussing what they are learning as a result of the lesson, teachers reviewing what was taught and what students have said they have learned, or asking the questions provided in your TE under *Reteach*. The important thing is that you “pull it altogether” at the end of the lesson. This helps to not only reinforce the concepts at the end of the lesson but also gives the students a sense of accomplishment.
- ✓ **INDEPENDENT PRACTICE.** Teachers should assign homework and independent practice for each lesson. That is, students should have an opportunity to practice new science concepts/skills on their own. The *Guided Reading and Study Workbook* worksheets are designed to provide students with independent practice, though you can, obviously, use your own assignments. When feasible (e.g. not during testing) students should be assigned approximately **30** minutes worth of independent practice each school night. Note that when assigning homework, students should be able to take the textbook home.
- ✓ **MATH: ANALYZING DATA AND WRITING IN SCIENCE.** The *2005 Science Explorer* program offers activities designed to reinforce higher level math and writing skills through the *Math: Analyzing Data* and *Writing in Science* activities presented in the text. ***Teachers should try to assign at least one Math: Analyzing Data and one Writing In Science activity per chapter taught.***
- ✓ **DIFFERENTIATED INSTRUCTION:** ***Teachers should try to tailor instruction to meet the needs of all students.*** Any method that you feel will meet the needs of your students is acceptable. To assist you in meeting the needs of your students, the TE offers, throughout each lesson, coded questions and activities according to difficulty (L1, L2, L3). *Differentiated Instruction* boxes offer suggestions for activities designed to meet different student subgroups, such as “English Learners/Beginning,” “English Learners/Intermediate,” “Less Proficient Readers,” “Special Needs” and “Gifted and Talented.”

✓ **OTHER PROGRAM COMPONENTS.** All other materials and activities provided with this science program and not listed here are considered optional. However, the following Prentice Hall program components are highly recommended. If it's feasible and practical, consider using the following:

- ❑ *Reading Strategies for Science Content*
- ❑ *Inquiry Skills Activity Books*
- ❑ *Standardized Test Prep Workbook*
- ❑ *Progress Monitoring Assessments*
- ❑ *Test-Taking Tips with Transparencies*

As you are aware, we will also be monitoring implementation. We will also conduct two classroom observations, in part, to determine the extent to which teachers are implementing all key components. In addition, teachers will complete weekly logs to indicate the extent to which they used key and optional components. We will provide more detail on the teacher logs shortly. Together, these data will help us determine the fidelity of implementation.

Appendix B:
Teacher Ratings of Program Usefulness

Teacher Ratings of Program Usefulness

Table B1 shows the teachers' ratings of the usefulness of their respective program materials and resources used. Of note is the finding²⁹ that treatment teachers generally reported greater usefulness of 2005 *Science Explorer* components than control teachers did of their program.

Table B1. Teacher Ratings of Usefulness of Their Respective Program Components

	Treatment	Control
Professional development resources embedded in program	75%	50%
Formal assessments	100%	50%
Informal assessments	100%	100%
Review materials/exercises	100%	67%
Independent practice exercises	100%	100%
Remediation resources	75%	0%
Intervention resources	75%	0%
Enrichment resources	75%	0%
Technology in the program	50%	100%
Supplemental worksheets/books provided with the program (as a whole)	100%	100%
Program's Internet resources	50%	50%
Lesson planning resources	75%	0%
Teaching tips	75%	0%
Organization of the textbook	75%	67%
Ease of use of the textbook	100%	100%
Ability to make connections to other subject areas	100%	33%
Ability to make connections to the real-world	100%	33%
Increasing your students' higher order cognitive skills	100%	33%
Increasing your students' science skills and knowledge	100%	33%
Preparing your students to do well on tests	100%	0%
Preparing your students to do well in future science courses	100%	33%

²⁹ Caution should be used in generalizing these results since it is based on only 4 treatment teachers and 3 control teachers.

Appendix C:
2005 *Science Explorer* Resources Used

Science Explorer Resources Used

Lab Resources

As shown in Table C1, the PH lab resources that were used most often during pilot study period were *Discover*, *Teacher Demo* and *Try This* activities. Teachers also used a number of activities that were not a part of the program but that they had used in the past.

Table C1. Number of Times Treatment Teachers Reported Using the Following Lab Activities

	Number of times used ³⁰
Discover Activity (Short Exercise)	28
Lab Activity not from 2005 PH <i>Science Explorer</i>	16
Try This Activity (Short Exercise)	11
Teacher Demo (Short Exercise)	11
Skills Lab (Full Period)	8
Skills Activity (Short Exercise)	8
Chapter Project (Ongoing)	5
At Home Activity (Short Exercise)	3
Design Your Own Lab (Full Period)	2

Technology Resources

As shown in Table C2, the PH technology resources that were used most often during pilot study period were the consumable and non-consumable materials kit, *Exam View CD-ROM*, *Student Edition* on Audio CD, and internet resources (*PH.School.com/Go Links*).

Table C2. Number of Times Treatment Teachers Reported Using the Following Technology Resources

	Number of times used ³⁰
Consumable Materials Kit	26
Non-Consumable Materials Kit	26
Exam View CD ROM (Test Generator)	18
Student Edition on Audio CD	12
PHSchool.com or 'Go Online' links from textbook	10
Lab Activities Videos	6
Used display device (e.g. projector for transparencies)	4
Other non-PH software or internet resources for support	4
Discovery Channel Videos	3
Teacher Express CD ROM	3
Other non-PH software or internet resources for prof. development	2
Interactive Textbook Online	1

³⁰ This includes the actual count of resources used during the study period as reported on the weekly teacher logs. These numbers are likely higher given that some teachers did not complete the logs every week.

Print Resources

Other than the Teacher and Student Editions, the PH print resources that were used most often during pilot study period included the *All in One* teaching resources, *Guided Reading and Study Workbook*, and the transparencies. A number of non-PH resources were used as well; these mainly consisted of resources they were already familiar with and/or they created.

Table C3. Number of Times Treatment Teachers Reported Using the Following Print Resources

	Number of times used ³⁰
Teacher's Edition	65
Student Edition	64
All in One Teaching Resources	60
Guided Reading and Study Workbook	47
Other Non-PH Print Resources	22
Color Transparencies	15
Inquiry Skills Activity Book	3
Differentiated Instruction Guide to Labs and Activities	3
Progress Monitoring Assessments	2
Adapted Chapter and Unit Tests	2
Adapted Reading and Study Workbook	1
Reading Strategies for Science Content	1

Appendix D:
TIMSS-F Development and Psychometric Properties

Overview³¹ of the Original 2003 TIMSS

The Trends in International Mathematics and Science Study (TIMSS) is a project of the International Association for the Evaluation of Educational Achievement (IEA). The IEA is an independent international cooperative of national research institutions and government agencies that has been conducting studies of cross-national achievement since 1959. TIMSS 2003 is the most recent in the series of IEA studies to measure trends in students' mathematics and science achievement. Offered first in 1995 and then in 1999, the regular cycle of TIMSS studies provides countries with an unprecedented opportunity to measure progress in educational achievement in mathematics and science.

The overriding principle in constructing tests for the upcoming cycles of the TIMSS study is to produce assessment instruments that will generate achievement data that are valid for the purposes they are to be used for, and are reliable. An international panel of mathematics and science education and testing experts provided guidance for the general form the assessment frameworks should take. The U.S. National Science Foundation provided support for the meetings and the work of the expert panel. Using an iterative process, successive drafts were presented for comment and review by National Research Coordinators, national committees, and expert panel members.

The identified frameworks do not consist solely of content and behaviors included in the curricula of all participating countries. The aim of the extensive consultation on curriculum was to ensure that goals of mathematics and science education regarded as important in a significant number of countries are included. The ability of policy makers to make sound judgments about relative strengths and weaknesses of mathematics and science education in their systems depends on achievement measures being based, as closely as possible, on what students in their systems have actually been taught. This is also a prerequisite for valid use of the measures in many potential secondary analyses.

Based on the frameworks, the TIMSS tests are developed through an international consensus-building process involving input from experts in education, mathematics, science, and measurement. The tests contain questions requiring students to select appropriate responses or to solve problems and answer questions in an open-ended format.

Development of the TIMSS-F

TIMSS released selected items from the 2003 8th grade science assessment into the public domain. Using the released item set, PRES Associates created the TIMSS-RE, the first shortened version of the original 2003 TIMSS that was administered in January 2005. Four blocks of test items were created, each containing seven constructed-response items and six multiple-choice items, for 13 items in each block. Four different forms of the assessment, each with 3 blocks of test items, were developed to minimize testing time while increasing the pool of questions to be tested during the pilot study. The TIMSS-RE was then shortened further to 30 items to create the TIMSS-F; this modification is discussed further in this section.

³¹ The following summary was taken from TIMSS documents available online <http://timss.bc.edu/timss2003i>.

Using Item Response Theory (IRT), the items were selected to provide an equivalent distribution of item difficulty and item discrimination in each block. Before proceeding to the psychometric properties of this test and the IRT results, a primer in IRT is provided.

Primer in IRT

In Item Response Theory, item difficulty and student ability are linked and measured on the same scale. These values generally run between -3 and +3, much the same as z-scores, where a value of 0 indicates an item of average difficulty or a student of average ability. Smaller values represent easier items and weaker students; large values represent more difficult items and stronger students. A student of a given ability engaging an item of the same difficulty has approximately a 50% chance of answering the item correctly.

Item discrimination in IRT is defined as the slope (steepness) of the s-shaped item response function at the point where a student has approximately a 50% chance of answering the item correctly. In addition, IRT discrimination represents the degree to which an item differentiates between students of ability just above and just below the item's difficulty. Generally, IRT discrimination ranges between .5 and 1.5, where the higher values indicate a steeper s-shaped curve and higher differentiation between students. However, there is no expectation of a distribution of item discrimination

“Good” items are usually items that discriminate better than “poor” items. However, extremely good items (discrimination above 1.5) are often difficult to write, and test developers have found that items with discrimination scores of 0.5 to 1.5 are quite acceptable for use. Of course, the values of discrimination can exceed the stated range, and items exhibiting discrimination above 1.5 are often retained for examination construction, whereas items with values below .5 are often discarded because they provide little discrimination between students at any level of ability. However, the demands of both content coverage and appropriate examination difficulty often influence the decision regarding item retention, and items exhibiting discrimination somewhat below .5 are often included in examination construction with little or no detriment to the intended purpose of the assessment.

Comparability of the TIMSS-based Assessment with the Original 2003 TIMSS

The mean item difficulty of each block ranged from -.02 to +.02, indicating that the items were chosen carefully over the range of available difficulties, which was about -1 to +1, such that the blocks as a group, or testlet, were of “average” difficulty. The mean item discrimination of each block ranged from .858 to .889, indicating again that the items were chosen carefully over the range of available difficulties.

The “average” student (ability=0) will find the examination of “average” difficulty, although this average student will find some items to be easier than other items. Moreover, because the individual items discriminate reasonably between students of low and high ability, the examinations should perform in a similar manner. Hence, it is reasonable to expect that the examinations constructed from these blocks provide adequate assessment information regarding student learning.

In terms of the content domains, the content of the TIMSS-RE approximates the content of the TIMSS as closely as is possible with the reduced number of items.

Table D1. Content Breakdown of TIMSS and TIMSS-RE

	TIMSS	TIMSS-RE
Physical Science	40%	39%
Life Science	30%	31%
Earth Science	30%	31%

Note: The TIMSS-RE percents do not add to 100% due to rounding.

Similarly, the distribution of items across the cognitive domains reflects the composition of the original TIMSS as closely as the reduced number of items will permit. The percents presented below are approximate for Factual Knowledge and Conceptual understanding, where cognitive domain matching forced the conceptual domain matching to be off by one item in some blocks.

Table D2. Cognitive Breakdown of TIMSS and TIMSS-RE

	TIMSS	TIMSS-RE
Factual Knowledge	30%	32%
Reasoning and Analysis	35%	31%
Conceptual Understanding	35%	36%

Note: The TIMSS-RE percents do not add to 100% due to rounding.

During pilot testing, the 39-item TIMSS-RE was found to be too long for the allotted 45 minutes, and nine items were removed so the exam would fit reasonably in the time permitted. This was guided by examination of each item's discrimination and difficulty parameters. The final version of the test is referred herein as the TIMSS-F.

The reduction in exam length produced the following content-area breakdown. The content distribution of the TIMSS-F became an exact match to the original TIMSS content distribution.

Table D3. Content Breakdown of TIMSS and TIMSS-F

	TIMSS	TIMSS-F
Physical Science	40%	40%
Life Science	30%	30%
Earth Science	30%	30%

Additionally, the reduction in examination length altered the distribution of items across the cognitive domains slightly. Although more disparate in distribution than the original TIMSS, the TIMSS-F item distribution is still a reasonable approximation, with 7% more items in the

Factual Knowledge domain and 8% fewer items in the Conceptual Understanding domain. The Reasoning and Analysis domain increased in item representation by 2%.

Table D4. Cognitive Breakdown of TIMSS and TIMSS-F

	TIMSS	TIMSS-F
Factual Knowledge	30%	37%
Reasoning and Analysis	35%	37%
Conceptual Understanding	35%	27%

Note: The TIMSS-RE percents do not add to 100% due to rounding.

The reduction of the examination length produced a small change in the IRT item difficulty and discrimination indices. The mean item difficulty of the TIMSS-F was .29, with a range from -1.2 to +1.3. The mean item discrimination was .88, with a range of .3 to 1.7. Although the ability of the revised examination to discriminate between students of low and high ability was essentially unchanged, the revision did produce an examination that is somewhat more difficult. If one assumes that student ability is normally distributed, the shift in examination difficulty is a little less than one-third of a standard deviation, suggesting that about 11% (34/3) more students will find the examination above average in difficulty.

Multilog was used to compute the IRT difficulty and discrimination parameter estimates for the TIMSS-F. Table 5 shows the IRT parameters for the TIMSS-F at pre (using only the 30 items selected) and post-test, as well as those obtained for the original 2003 TIMSS.

Table D5. IRT Parameters for Pre- and Post-Test TIMSS-F and TIMSS

Item	Post-TIMSS-F		Pre-TIMSS-F		2003 TIMSS	
	a-est	b-est	a-est	b-est	a-est	b-est
CR_D1_32704_1	0.72	0.87	0.98	0.42	0.82	0.34
MC_D11_12030_2	0.07	6.88	0.50	1.24	0.68	0.23
CR_A11_32375_3	0.97	0.77	0.98	0.57	0.63	0.37
CR_C5_22160_4	0.83	-0.30	0.85	0.63	0.67	0.40
CR_B10_32206_5	1.25	1.48	1.04	0.36	1.11	0.72
MC_C2_32301_6	0.70	0.32	1.12	0.63	1.55	0.58
MC_D8_12028_7	0.74	-0.53	0.96	-0.15	0.83	-0.05
CR_D13_32532_8	0.35	3.98	0.52	-0.46	0.76	-0.49
CR_A5_32063_9	0.96	2.35	1.42	0.92	0.77	0.91
MC_C10_22040_10	0.42	-2.40	0.81	-0.10	0.69	-0.54
MC_D4_12038_11	0.83	-1.23	0.61	0.22	1.17	0.07
MC_C8_32150_12	1.71	-1.48	0.47	-0.29	0.59	-0.24
MC_C7_32652_13	0.87	-0.08	0.63	1.28	0.83	0.09
CR_D3_22191_14	1.36	0.38	0.93	0.19	0.66	-0.76
CR_D9_32626_15	2.08	-0.51	1.61	0.04	1.27	-0.01
MC_A12_32422_16	1.09	0.44	1.22	0.07	1.30	-0.13
MC_C13_32008_17	0.96	-0.36	1.01	0.96	0.95	-0.12
MC_D12_12029_18	0.76	-0.07	0.81	0.29	0.69	0.39
MC_C1_32055_19	1.39	-1.63	0.96	-1.17	0.99	-1.27

Item	Post-TIMSS-F		Pre-TIMSS-F		2003 TIMSS	
	a-est	b-est	a-est	b-est	a-est	b-est
CR_C6_32562_20	1.36	0.76	0.83	0.45	0.75	-0.13
CR_A10_22152_21	1.72	-1.03	0.81	-0.15	1.06	-0.13
MC_B5_12006_22	0.83	0.06	0.94	0.01	0.91	-0.12
MC_B6_22074_23	0.51	0.69	1.37	0.19	1.17	0.11
CR_C9_22035_24	0.11	1.69	0.35	-0.20	0.35	-0.21
MC_C12_32607_25	1.00	0.00	0.70	-0.10	0.95	-0.12
CR_C11_32707_26	1.89	1.66	1.72	0.71	1.56	0.95
MC_A1_12025_27	0.92	-0.48	0.57	1.07	0.77	1.14
CR_B9_22286_28	1.06	1.92	0.67	0.63	0.85	0.98
CR_B8_32131_29	1.78	-1.21	0.88	-0.33	0.95	-0.57
CR_D10_32711_30	1.22	0.87	0.80	0.63	0.88	0.62

*CR = constructed response; MC = multiple choice

To compare the technical qualities of the items selected for the TIMSS-F with the 2003 TIMSS, parameter estimates for the three versions of the TIMSS were compared via Pearson correlation. As shown in Table 6, although the correlations were not as high when the post-test data were considered, they were found to be acceptable to demonstrate the association between the TIMSS-F and the TIMSS. This reduction in observed correlation was likely due to the educational intervention that occurred between the pre- and the post-tests.

Table D6. Pearson Correlations

	Discrimination estimates	Difficulty estimates
Post and Pre TIMSS-F	0.44	0.40
Post-TIMSS-F and TIMSS	0.40	0.33
Pre-TIMSS-F and TIMSS	0.67	0.74

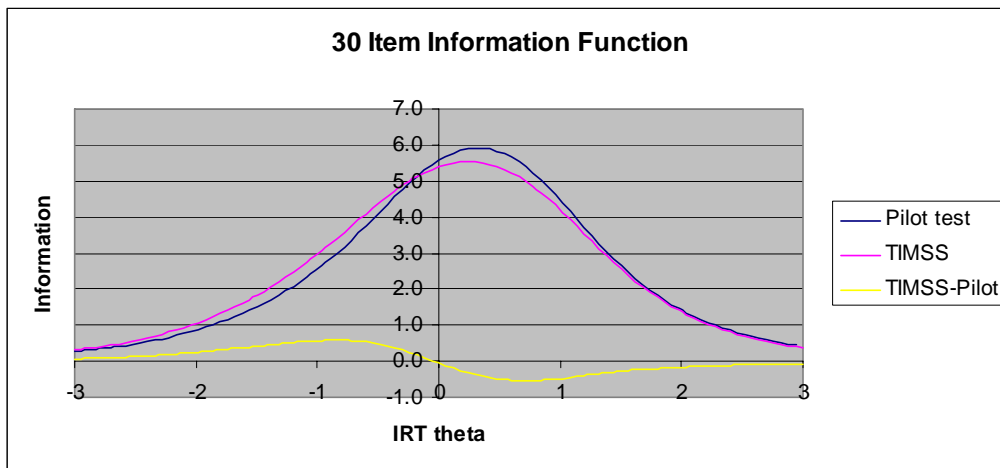
Note: All correlations are statistically significant at $p=.05$.

The item selection procedure previously described involved the careful selection of items by both difficulty and discrimination. Such a process represents an informal use of the information functions of IRT for item selection. However, using the information functions would permit the simultaneous consideration of both difficulty and discrimination, instead of choosing items first by difficulty and then by discrimination.

Information is the IRT analog of classical reliability. Just as an examination made of reliable items can be found to be a reliable tool for assessment, the information provided by individual items contributes to the information provided by the examination. For the purpose of classroom assessment, the information function of the examination is flatter, indicating a broader range of item difficulties as is appropriate for the broader range of abilities generally found in a classroom.

To contrast the information provided by the original TIMSS to the information provided by the TIMSS-F, the information function of each exam was computed and plotted on the same graph. In addition, the difference of the two information functions was also added to that graph.

Figure D1. 30 Item Information Function



- Notes: 1. IRT theta corresponds to student ability, presented on a scale of z-scores.
2. Pilot refers to the TIMSS-F.

This graph demonstrates that the selection process used to create the TIMSS-F resulted in an assessment with information characteristics highly similar to those of the original TIMSS. The differences in the information provided by the assessments are minor. In addition, the range over which information is adequately presented is over 1.5 standard deviations of ability. By extrapolation, the classical reliability of the TIMSS-F, at least at the total score level, should be also similar to the TIMSS.

Reliability and Validity of the TIMSS-F

Internal consistency

To examine the classical reliability of the TIMSS-F, correlational analyses were performed and Cronbach's alpha was computed using pre- and post-test data for the total score, the cognitive domain scales, and the content domain scales. (Cronbach's alpha represents the average of all possible split-half reliabilities.)

Table D7. Cronbach’s Alpha: Pre- and Post-Test

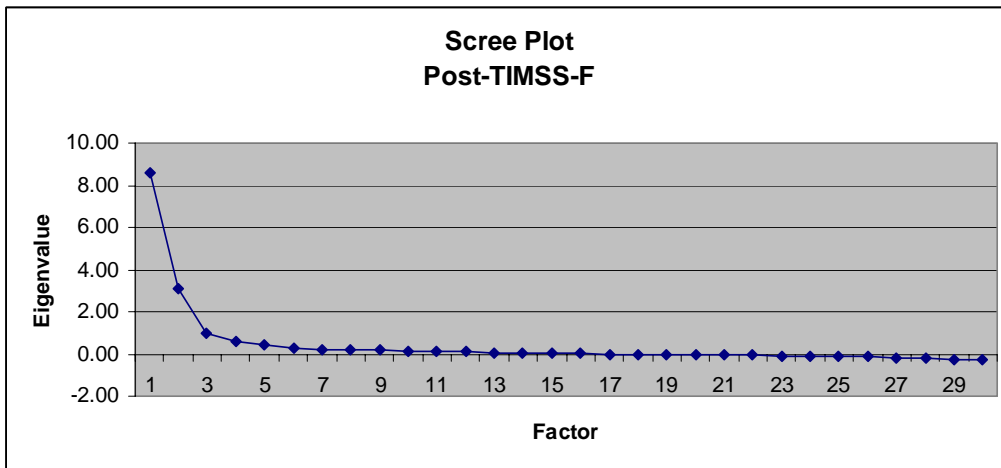
	Alpha		
	Pre-test	Post-test	Change
Entire Exam	.77	.80	.03
Life Science Content Domain	.52	.56	.04
Earth Science Content Domain	.54	.46	-.08
Physical Science Content Domain	.55	.68	.13
Conceptual Understanding Cognitive Domain	.64	.59	-.05
Factual Knowledge Cognitive Domain	.44	.45	.01
Reasoning and Analysis Cognitive Domain	.43	.70	.27

The observed values of alpha for the total score are sufficiently high for the intended use of the examination as an assessment tool. The domain scores are lower in internal consistency due to the reduced number of items on each scale. In addition, the reliability of the domain scores is not sufficient for use with diagnostic feedback.

Factor Analysis

To further demonstrate the validity of the TIMSS-F, a factor analysis using unweighted least squares and varimax rotation was performed on the post-test data. Square multiple correlations were used to estimate the prior communalities. The first unrotated factor accounted for 63% of the total variability; the second, 23%, suggesting the TIMSS-F is strongly unidimensional. The scree plot, as seen below, suggests a two-factor solution.

Figure D2. Scree Plot: Post TIMSS-F



Additional exploratory analyses suggested the existence of other minor factors that underlie these data, but the amount of additional variability explained by these factors is small, and additional confirmatory data would be required to accept the existence of these factors in the population.

The overwhelming evidence of unidimensionality demonstrated in the magnitude of the first unrotated factor further indicates the internal consistency (and by inference, the validity) of the TIMSS-F.

Concurrent Validity

The concurrent validity of the TIMSS-F was established by correlating the pre- and post-TIMSS-F data with pre- and post-ITBS data. The correlation between the total TIMSS-F scores and the ITBS scores was approximately .7 in both the pre-test and the post-test. The content domains of the TIMSS-F correlated with the ITBS are reduced levels because of the lower number of items on each content scale.

Table D8. Concurrent Validity of TIMSS-F

Concurrent Validity of TIMSS-F Using ITBS Scaled Score		
	Pre-test	Post-test
Total TIMSS-F	0.65	0.70
Life Science	0.47	0.62
Earth Science	0.50	0.54
Physical Science	0.53	0.58
N	264	250
p<.0001 for all correlations		

Scoring of the TIMSS-F

Each multiple-choice question is worth one point. Multiple-choice questions provide students with four or five response options, of which only one is correct. These questions can be used to assess any of the behaviors in the cognitive domains. However, because they do not allow for students' explanations or supporting statements, multiple-choice questions may be less suitable for assessing students' ability to make more complex interpretations or evaluations. Therefore, constructed-response questions are included.

Constructed-response questions are worth one or two points, depending on the nature of the task and the skills required to complete it. For this type of test item students are required to construct a written response, rather than select a response from a set of options. Constructed-response questions are particularly well suited for assessing aspects of knowledge and skills that require students to explain phenomena or interpret data based on their background knowledge and experience. Scoring guides (rubrics) for each constructed response question describe the essential features of appropriate and complete responses. The guides focus on evidence of the type of behavior the question assesses. They describe evidence of partially correct and completely correct responses. In addition, sample student responses at each level of understanding provide important guidance to those who will be rating the students' responses. In scoring students' responses to constructed-response questions, the focus is solely on students' achievement with respect to the topic being assessed, not on their ability to write well. However, students need to communicate in a manner that will be clear to those scoring their responses.

All tests were scored externally by Claremont Graduate University (CGU). CGU trained scorers in the use of TIMSS scoring guides for constructed-response (CR) items before scoring occurred. Given that the established CR scoring rubrics from the TIMSS were used, CGU measured interrater reliability by calculating percent agreement across six randomly selected constructed response questions (21% of the total CR questions) across two classes. Interrater reliability was determined through a three-step process. First, two researchers and two graduate students (with experience in coding) scored one randomly selected constructed-response question. All four individuals examined the scoring rubric and 10 student responses. The team discussed the rubric, identified the codes, discussed the logic behind the question, and then proceeded to (as a group) code the 10 responses together. Once coders were familiar with the format of the assessment and scoring rubric, the second step involved having the two coders independently code the remaining five items using a sample of 10 student responses (50 student responses total). The percent agreement across coders and across trials was .40, .80, .90, .90, .90 respectively. The main reason for the low agreement on the first trial involved the difficulty coding the incorrect responses. Responses could not be simply coded as incorrect, given that were levels associated with most incorrect responses. The third and final step involved discussing each student response where disagreements occurred and coming to a solution regarding that item. Once the disagreements were clarified, scorers demonstrated high interrater reliability.