



Prentice Hall Mathematics
©2004
Grades 6 – 12

**Program
Efficacy
Studies**

1998-2003

**Clinical research reports supporting
the efficacy of the
Prentice Hall Mathematics program**

Prentice Hall Mathematics Program Efficacy Studies

Table of Contents

Research Overview	3
Summary of Results	5
Program Efficacy Studies	
2003 Algebra 1	6
2003 Grades 6-8 Prentice Hall Mathematics	13
2002 Algebra 1	20
2000 Pre-Algebra	28
2000 Course 3 Middle School Mathematics	32
1999 Course 1 Middle Grades Mathematics & Algebra	37
1998 Course 2 Middle Grades Mathematics	41

Prentice Hall Mathematics Research Overview

Pearson Prentice Hall is proud of the fact that for over half a century we have used a variety of types of research as a base on which to build our mathematics programs. Over the years, we have worked collaboratively with our authors to make continual program improvements based on empirical, scientific research.

Three Phases of Research

Prentice Hall Mathematics is based on research that describes how students learn mathematics well, and provides classroom-based evidence of program efficacy. The three phases of research described below were integrated into the development of *Prentice Hall Mathematics*. The goal of establishing such rigorous research methods is to ensure that the program developed enables all students to learn the mathematics skills and concepts they need for academic success and for everyday life.

Prentice Hall's research is cyclical and ongoing, and provides evidence of a program's overall effectiveness based principally on students' test scores. Previous mathematics programs by Prentice Hall, Scott Foresman, and Addison-Wesely provided a strong basis for success. What we learned about the effectiveness of our previous programs informed the instructional design of our new program.

Pearson Prentice Hall is analyzing students' pretest and posttest scores with national districts and school-level data of users of the program. This process allows for continual monitoring of performance and learning from results on an ongoing basis.

(1) Exploratory Needs Assessment

Along with periodic surveys concerning curriculum issues and challenges, we conducted specific product development research, which included discussions with teachers and advisory panels, focus groups, and quantitative surveys. We explored the specific needs of students, teachers, and other educators regarding each book we developed in *Prentice Hall Mathematics*.

In conjunction with Prentice Hall authors, secondary research was done to explore educational research about learning. This research was incorporated into our instructional strategy and pedagogy to make a more effective mathematics program.

(2) Formative: Prototype Development and Field Testing

During this phase of research, we worked to develop prototype materials for each course in *Prentice Hall Mathematics*. Then we tested the materials, including field testing with students and teachers, and qualitative and quantitative evaluations of different kinds. We received solid feedback about our lesson structure in our early prototype testing. Results were channeled back into the program development for improvement.

(3) Summative: Validation Research

Finally, we conducted and continue to conduct longer-term research based on scientific, experimental designs under actual classroom conditions. This research identifies what works and what can be improved in revisions. We also continue to monitor the program in the market. We talk to our users about what works, and then we begin the cycle over again. This phase involves longitudinal, control-group research.

We also conduct independent, third-party research using quasi-experimental and experimental research designs. A national effect size study to validate the efficacy of *Prentice Hall Mathematics* comparing users versus matched non-users was recently completed. A longitudinal, experimental study will begin in Spring 2004. Specific program features such as the Instant Check System will be studied as well as fidelity of implementation and program efficacy. *Prentice Hall Mathematics* users will be compared to matched non-users and tested using national, standardized examinations. This process will inform the need for revision, help monitor student success, and identify how well our program works.

It is important to note that Pearson Prentice Hall uses this research to subsequently inform the development of the next program. Hence our three phases of scientific research form a cycle that is truly ongoing.

Summary of Results

Program Efficacy Studies

The following seven studies represent the third phase of our ongoing scientific research. The studies were conducted from 1998-2003 and address student performance on *Prentice Hall Mathematics* ©2004 and the programs from which *Prentice Hall Mathematics* ©2004 was built. Treatment and comparison groups were analyzed to validate the efficacy of Prentice Hall's mathematics programs against other mathematics programs. These clinical studies analyzed demographics, program implementation, pre- and post-test scores, and teacher effectiveness among other variables to determine the impact of Prentice Hall programs on student performance.

In each of the seven studies, findings identified significant statistical gains in students' test scores after one semester or one year of implementation of the respective Prentice Hall mathematics program. Further, confidence levels as high as **95%** were achieved in critical diagnostic skills areas. Specifically, significant gains were achieved in the following:

Algebra I:

- Variables, Expressions, Formulas
- Solving Linear Equations, Inequalities
- Graphing Linear Equations
- Functions and Graphs
- Quadratic Equations and Functions
- Geometry

Middle School Math:

- Patterns, Functions, Algebra
- Problem Solving and Reasoning
- Number and Number Relations
- Computation and Estimation
- Measurement
- Geometry and Spatial Sense
- Data, Statistics and Probability

Students using the Prentice Hall mathematics programs consistently performed as well or better than those using comparison programs.

Foundational Research

The extensive foundational research that was conducted for the *Prentice Hall Mathematics* ©2004 program is detailed in Prentice Hall's ***Research into Practice*** booklet. For more information or to obtain a copy, contact your local sales representative or visit phschool.com/MathResearch.

2003 Algebra 1 Program Efficacy Study

Abstract

This study investigated the effects of the Prentice Hall *Algebra 1* ©2004 textbook program on student performance. Four classes of high school algebra 1 students were assigned to a treatment group, and were instructed with the Prentice Hall *Algebra 1* program for a full school year. As a basis for comparison, four comparable classes of high school algebra 1 students in the same schools were assigned to a control group and were instructed using their current textbook programs for a full school year. All of the students were tested at the start of the school year and at the end of the school year with the same standardized test, the TerraNova® Algebra exam. Both the treatment and control groups showed significant overall learning improvement over the course of the 2002-03 school year, as well as in each of the algebra diagnostic skills areas. Effect Size data support the findings of improvement among these students.

Objective

The main objective of this study was to determine whether students who were enrolled in classes using Prentice Hall *Algebra 1* ©2004 significantly increased their algebra knowledge and skills after using the program for a full school year.

Methodology

This study followed a quasi-experimental research design. The treatment group used the Prentice Hall *Algebra 1* ©2004 program, and the control group used the full algebra 1 textbook program that had previously been adopted for use in the school. Both groups were tested in September 2002, prior to the program's introduction, and then again in May 2003, at the end of the school year. Therefore, for both the treatment group and the control group, there was a pre-test score and a post-test score. Only students who completed both the pre- and post-tests were included in this analysis.

A total of four schools and eight 9th grade algebra 1 classes comprising 106 students participated in the study. In a few cases, 10th and 11th grade students were integrated into the 9th grade classes.

	Treatment Group Prentice Hall <i>Algebra 1</i> Textbook Program	Control Group Other Algebra 1 Textbook Programs
Pre-Test (administered at beginning of school year)	4 classes 64 students	4 classes 42 students
Post-Test (administered at end of school year)	4 classes 64 students	4 classes 42 students

Both treatment and control classes were in the same school building. Treatment and control classes were selected by the teachers to be similar in student ability levels. In three of the four schools, one teacher taught both treatment and control classes. In one school where this was not possible, two teachers with similar backgrounds were selected.*

*An additional statistical analysis was completed to rule out any effects of classes taught by the same teacher versus classes taught by two different teachers. The data set was analyzed with and without the school in which two different teachers taught the treatment and control classes. The final results were statistically equivalent. Therefore, we have included the school in which two different teachers taught the treatment and control classes in this report of final results.

Demographic profiles of the four schools participating in this study are shown below.

	School 1 - IL	School 2 - IN	School 3 - IN	School 4 - TX
Enrollment	180	850	1,000	426
Ethnicity	Caucasian 99%, African-American 1%,	Caucasian 95%, Hispanic 5%, Asian 1%	Caucasian 100%	Caucasian 66%, Hispanic 27%, Asian 4%, African American 3%
Locale	Rural, Inside Metropolitan Area	Rural, Inside Metropolitan Area	Small Town	Rural, Outside Metropolitan Area
Poverty Level	Med-High Income (6.0 - 15.9% Poverty)	Med-High Income (6.0 - 15.9% Poverty)	Med-High Income (6.0 - 15.9% Poverty)	Med-Low Income (16.0 - 29.9% Poverty)

*Source: Market Data Retrieval

The intervention being studied comprised of a mainstream textbook with which teachers were generally familiar. Therefore, training was limited to on-site in-service conducted by Prentice Hall Consultants designed to ensure that teachers understood the treatment program and could instruct with it as its designers intended. Teachers (and administrators) were briefed and provided with materials that explained the study and reinforced the need to implement it as designed, without, for example, contamination of the program materials or approaches between their treatment and control classes. Prentice Hall Consultants also completed two in-school observations of each class (both treatment and control) during the full year treatment period: once in Fall 2002, and once in Spring 2003. These observations were designed to ensure that proper study implementation was adhered to.

The measures used in the analysis that follows are:

- Raw scores received on the TerraNova® Algebra exam, which measures overall student performance (*See Technical Post Script on page 10 for a full description of this measure*); and
- The Objectives Performance Index (OPI) of the TerraNova® Algebra exam, which measures student performance on several algebra diagnostic skills areas. (*See Technical Post Script on page 10 for a full description of this measure*).

All student tests were scored by the publisher of the TerraNova® Algebra exam, CTB/McGraw Hill. (*see Appendix on page 12 for company profile*). Statistical analyses and conclusions were completed by an independent statistician with experience in educational research methodologies and analyses (*see Appendix on page 12 for profile*).

Analysis of Results

At the outset, it was important to determine whether there was a statistically significant difference between the treatment group and the control group on the overall pre-test score. For this, a *t-test* on the difference between the treatment and control mean pre-test scores was used. The measure used in this analysis was the raw score of the TerraNova® Algebra exam. Analysis of the mean scores shows there were **no significant differences at the starting point of the study** (see table below).

Subject	Treatment Mean Pre-test Raw Score (base)	Control Mean Pre-test Raw Score (base)	Absolute difference	<i>t-value</i>	Sig (<i>t-value</i>)
Algebra 1	8.31 (64)	8.90 (42)	0.59	1.08	0.282

Next, we examined whether or not there was a statistically significant difference between the pre-test and post-test scores for students in the treatment group. For this, a *t-test* on the difference between the mean pre-test and post-test scores was used. The measure used in this analysis was the raw score of the TerraNova® Algebra exam. Analysis of the mean raw scores shows that **students using the Prentice Hall Algebra 1 program showed significant improvement in overall test scores from the pre-test to the post-test** (see table below).*

Subject	Treatment Mean Pre-test Raw Score (base)	Treatment Mean Post-test Raw Score (base)	Absolute difference	<i>t-value</i>	Sig (<i>t-value</i>)¹
Algebra 1	8.31 (64)	14.81 (64)	6.50	7.76	0.000

For comparison, we also looked at whether or not there was a statistically significant difference between the pre-test and post-test scores for students in the control group. For this, a *t-test* on the difference between the mean pre-test and post-test scores was used. Analysis of the mean scores shows that **students using their current textbook programs also showed significant improvement in overall test scores from the pre-test to the post-test** (see table below).*

Subject	Control Mean Pre-test Raw Score (base)	Control Mean Post-test Raw Score (base)	Absolute difference	<i>t-value</i>	Sig (<i>t-value</i>)¹
Algebra 1	8.90 (42)	15.83 (42)	6.93	7.16	0.000

In addition, we wanted to find out whether there was a significant difference between the treatment and control groups on the post-test, after adjusting for the pre-test level, by using an Analysis of Covariance (ANCOVA). This analysis shows that **the overall mean post-test scores for the treatment and control groups were statistically equivalent** (see table below).*

Subject	Treatment Mean Post-test Raw Score (base)	Control Mean Post-test Raw Score (base)	Absolute difference	F-value	Sig (F) ¹
Algebra 1	14.81 (64)	15.83 (42)	1.02	0.29	0.589

We also performed an Effect Size analysis (see *Technical Post Script on page 10 for a full description of Effect Size and its interpretation*). **The Effect Size analysis confirms that there was significant learning improvement for both the treatment and control groups from pre to post** (see table below).

Analysis Performed:	Cohen's d^2
Treatment Pre to Post	1.31
Control Pre to Post	1.47

¹Shaded values are statistically significant at the 95% level of confidence or higher

²Shaded values are educationally significant (0.25 or higher)¹ (Tallmadge, G. (1977). *The joint dissemination review panel idea book*. Washington, DC: National Institute of Education and the U.S. Office of Education).

*No judgment is made here as to whether these should or should not be considered "meaningful" improvements, a distinction sometimes applied to observed real world changes in school/district performance on standardized tests.

We also used an ANCOVA to determine whether individual teachers had an impact on student performance and if there was a teacher interaction whether this interaction favored one of the groups or neither. The measure used in this analysis was the raw score of the TerraNova[®] Algebra exam.

This analysis shows that there were **significant post-test differences for students between the individual teachers participating in this study, independent of the textbook program being used**. However, there were **no significant differences for individual teachers in their treatment versus control class performance** (see table below).

Source of Possible Variation:	F-value	Sig (F) ¹
Individual Teacher Effect	67.59	0.000
Teacher by Section Interaction	0.27	0.849

Performance on algebra instructional objectives for the treatment and control groups was also analyzed. The measure used in this analysis was the OPI of the TerraNova® Algebra exam (see *Technical Post Script below for a full description of this measure*).

Consistent with the overall improvement in test scores, both treatment and control students showed significant gains at the 95% confidence level in all diagnostic skills areas for Algebra (as defined by CTB/McGraw Hill):

- Variables, Expressions, Formulas
- Solving Linear Equations, Inequalities
- Graphing Linear Equations
- Functions and Graphs
- Quadratic Equations and Functions
- Geometry

Conclusions

Both the treatment group using the Prentice Hall *Algebra 1* textbook program and the control group using their current algebra 1 textbook programs showed significant learning improvement over the course of a full school year. Both groups accomplished significant gains in their overall raw scores, and in all algebra diagnostic skill areas. When comparing the results for the treatment group and the control group, learning improvement occurred at equal levels.

Individual teacher interactions were demonstrated across both the treatment and control classes they taught, a possible indicator that teacher skill and experience played a large role in student outcomes. These interactions did not favor the treatment group or the control group.

¹Shaded values are significant at the 95% level of confidence or higher

Technical Post Script

For the purpose of completeness, we include here a brief discussion of the statistical tests and measures that were used. All of the statistical tests are in the classical statistical domain, and are broadly used across all disciplines including psychometrics. The program used in the computational steps of this project was SPSS version 11.5.1.

The *t*-test that was used states: *The mean of one population is equal to the mean of another population*. The hypothesis is that the means are equal; the alternative hypothesis is that they are unequal we use a two-sided test. The means and the variance are calculated and the *t*-statistic is computed. The significance of the *t*-statistic then computed. If the significance of the *t*-statistic is less than .05 the hypothesis is rejected; otherwise, the alternative hypothesis is accepted.

ANCOVA: The dependent variable (in this case the post-test score) is equal for both populations (treatment and control) after adjustment for the covariate (in this case the pre-test score). A number of assumptions are invoked in the ANCOVA: population variances about the regression lines; the regression curve is a straight line; the pre- and post-test values are approximately normally distributed. The *F*-value is the statistic that is computed from a series of

computations of sums of squares and sums of products. The distribution of this F-value is known and the significance may be computed. If the significance of the F-value is less than .05, the hypothesis for equal means of the post-test is rejected; otherwise, it is accepted.

Raw Score: “The number-correct or ‘raw’ score is the number of items answered correctly by a student on any given test section.”¹ (The TerraNova[®] Algebra test includes a total of 30 questions; therefore, the “raw” score for each student equals the number correct out of 30.)

¹Quoted directly from: *Beyond the Numbers, A Guide to Interpreting and Using the Results of Standardized Achievement Tests*, page 18, CTB/McGraw-Hill 2003.

Objectives Performance Index: “All versions of TerraNova[®] and TerraNova[®], the Second Edition yield criterion-referenced scores reported in terms of an OPI. An OPI is reported for each of the instructional objectives measured by the particular version of TerraNova[®] and TerraNova[®], the Second Edition. Each objective is measured by at least four items. CTB has established a standard of four items as the minimum number needed to produce reliable information regarding objective mastery. The OPI makes test results both understandable and useful for the teacher in planning effective learning strategies and activities. The OPI is an estimate of the number of items that a student would be expected to answer correctly if there had been 100 similar items for that objective. The OPI scale runs from ‘0’ to ‘100’.”²

²Quoted directly from: *Beyond the Numbers, A Guide to Interpreting and Using the Results of Standardized Achievement Tests*, page 12, CTB/McGraw-Hill 2003.

Effect Size: “Effect size (ES) is a name given to a family of indices that measure the magnitude of a treatment effect.”³ Effect size is “simply a way of quantifying the effectiveness of a particular intervention, relative to some comparison intervention. It is easy to calculate, readily understood and can be applied to any measured outcome in Education or Social Science.”⁴ Cohen’s *d* is commonly used to calculate effect size. Cohen’s *d* is a “standardized” measure found by dividing the difference in the group mean scores by the standard deviation (adjusted for sample size differences known as the “pooled” estimate of the standard deviation). Whereas “statistical significance” is based on a specific probability [“p-value”], it can be affected much more by sample size than effect size’s interpretation. In educational research, an effect size of 0.25 or more is commonly considered to be “educationally significant.”⁵

³Quoted directly from: University of Colorado at Colorado Springs website, link: <http://web.uccs.edu/lbecker/Psy590/es.htm>.

⁴Quoted directly from: Curriculum, Evaluation and Management Centre (CEM) website, link: <http://cem.dur.ac.uk/ebeuk/research/effectsize/ESguide.htm>.

⁵Tallmadge, G. (1977). *The joint dissemination review panel idea book*. Washington, DC: National Institute of Education and the U.S. Office of Education.

Appendix

William M. Bailey, Statistician, WMB & Associates, Orlando, FL

William M. Bailey is an independent statistician and market analyst with 15+ years of experience in the design, execution, interpretation and reporting of qualitative and quantitative research studies both in the education and consumer fields.

His experience with the U.S. Department of Education includes:

- Statistical analysis on several studies tracking pre- and post -course design among a broad-based selection of community colleges and universities
- Serving as an on-call statistical and methodological consultant to the Research to Practice Division
- Serving as a reviewer of proposals to the Office of Special Education & Programs (OSEP)

Mr. Bailey also serves as a beta test site for SPSS, working with development of their core statistical package since version 7 and is now involved with version 12. He also is assisting in the design and testing of the new complex sampling module.

CTB/McGraw Hill, Monterey, CA

CTB/McGraw Hill, publisher of the TerraNova[®], is a division of the McGraw Hill Companies. It was founded in 1926 and provides a number of standardized achievement tests for both children and adults. The company scores over 20 million test documents each year.⁸

⁸ Source: www.CTB.com

2003 Grades 6-8 Mathematics Program Efficacy Study

Abstract

This study investigated the effects of the Prentice Hall *Mathematics, Course 2* ©2004 textbook program on student performance. Four classes of seventh grade mathematics students were assigned to a treatment group, and were instructed with the Prentice Hall *Mathematics, Course 2* program for a full school year. As a basis for comparison, four comparable classes of seventh grade mathematics students in the same schools were assigned to a control group and were instructed using their current textbook programs for a full school year. All of the students were tested at the start of the school year and at the end of the school year with the same standardized test, the TerraNova® Complete Battery Plus exam. The treatment group showed significant overall learning improvement over the course of the 2002-03 school year, while the control did not show significant learning improvement overall during this same time period. The treatment group achieved significant gains in six mathematics diagnostic skills areas, while attaining a higher level of achievement than the control group in Geometry and Spatial Sense and Data, Statistics, and Probability. The control group achieved significant gains in four mathematics diagnostic skills areas. Effect Size data support the findings among each group.

Objective

The main objective of this study was to determine whether students who were enrolled in classes using Prentice Hall *Mathematics, Course 2* ©2004 significantly increased their mathematics knowledge and skills after using the program for a full school year.

Methodology

This study followed a quasi-experimental research design. The treatment group used pre-production chapters and selected ancillary materials of the Prentice Hall *Mathematics, Course 2* ©2004 program for the first half of the school year (August – December 2002), and the full Prentice Hall program for the second half of the school year (January – May 2003). The control group used the full mathematics textbook program that had previously been adopted for use in the school. Both groups were tested in September 2002, prior to the program's introduction, and then again in May 2003, at the end of the school year. Therefore, for both the treatment group and the control group, there was a pre-test score and a post-test score. Only students who completed both the pre- and post-tests were included in this analysis.

A total of four schools and eight seventh grade mathematics classes comprising 126 students participated in the study.

	Treatment Group Prentice Hall <i>Mathematics</i> Textbook Program	Control Group Other Mathematics Textbook Programs
Pre-Test (administered at beginning of school year)	4 classes 65 students	4 classes 61 students
Post-Test (administered at end of school year)	4 classes 65 students	4 classes 61 students

Both treatment and control classes were in the same school building. Treatment and control classes were selected by the teachers to be similar in student ability levels. In each of the four schools, one teacher taught both treatment and control classes. Demographic profiles of the four schools participating in this study are shown below.*

	School 1 - IL	School 2 - IN	School 3 - NC	School 4 - NC
Enrollment	623	377	366	567
Ethnicity	Caucasian 77%, Hispanic 23%	Caucasian 98%, Hispanic 1%, Asian 1%	Caucasian 51%, African American 28%, Native American 21%	African American 74%, Caucasian 23%, Hispanic 1%, Asian 1%
Locale	Urban Fringe of Mid-Size City	Rural, Inside Metropolitan Area	Rural, Inside Metropolitan Area	Urban Mid-Size City
Poverty Level	Med-High Income (6.0 - 15.9% Poverty)	Med-High Income (6.0 - 15.9% Poverty)	Med-Low Income (16.0 - 29.9% Poverty)	Med-Low Income (16.0 - 29.9% Poverty)

*Source: Market Data Retrieval

The intervention being studied comprised of a mainstream textbook with which teachers were generally familiar. Therefore, training was limited to on-site in-service conducted by Prentice Hall Consultants designed to ensure that teachers understood the treatment program and could instruct with it as its designers intended. Teachers (and administrators) were briefed and provided with materials that explained the study and reinforced the need to implement it as designed, without, for example, contamination of the program materials or approaches between their treatment and control classes. Prentice Hall Consultants also completed in-school observations of each class (both treatment and control) during the full year treatment period. These observations were designed to ensure that proper study implementation was adhered to.

The measures used in the analysis that follows are:

- Raw scores received on the TerraNova® Complete Battery Plus exam, which measures overall student performance (*See Technical Post Script on page 18 for a full description of this measure*); and
- The Objectives Performance Index (OPI) of the TerraNova® Complete Battery Plus exam, which measures student performance on several mathematics diagnostic skills areas. (*See Technical Post Script on page 18 for a full description of this measure*).

All student tests were scored by the publisher of the TerraNova® Complete Battery Plus exam, CTB/McGraw Hill. (*see Appendix on page 19 for company profile*). Statistical analyses and

conclusions were completed by an independent statistician with experience in educational research methodologies and analyses (see Appendix on page 19 for profile).

Analysis of Results

At the outset, it was important to determine whether there was a statistically significant difference between the treatment group and the control group on the overall pre-test score. For this, a *t*-test on the difference between the treatment and control mean pre-test scores was used. The measure used in this analysis was the raw score of the TerraNova® Complete Battery Plus exam. Analysis of the mean scores shows there were **no significant differences at the starting point of the study** (see table below).

Subject	Treatment Mean Pre-test Raw Score (base)	Control Mean Pre-test Raw Score (base)	Absolute difference	<i>t</i>-value	Sig (<i>t</i>-value)
Mathematics	29.80 (65)	27.56 (61)	2.24	1.426	0.1563

Next, we examined whether or not there was a statistically significant difference between the pre-test and post-test scores for students in the treatment group. For this, a *t*-test on the difference between the mean pre-test and post-test scores was used. The measure used in this analysis was the raw score of the TerraNova® Complete Battery Plus exam. Analysis of the mean raw scores shows that **students using the Prentice Hall *Mathematics, Course 2* program showed significant improvement in overall test scores from the pre-test to the post-test** (significant at 99% confidence - see table below).*

Subject	Treatment Mean Pre-test Raw Score (base)	Treatment Mean Post-test Raw Score (base)	Absolute difference	<i>t</i>-value	Sig (<i>t</i>-value)¹
Mathematics	29.80 (65)	32.57 (65)	2.77	2.783	0.0071

For comparison, we also looked at whether or not there was a statistically significant difference between the pre-test and post-test scores for students in the control group. For this, a *t*-test on the difference between the mean pre-test and post-test scores was used. Analysis shows that the gain in mean scores is just outside the threshold of significance at the 95% confidence level (94%). Therefore, **students using their current textbook programs did not show significant improvement in overall test scores from the pre-test to the post-test**.

Subject	Control Mean Pre-test Raw Score (base)	Control Mean Post-test Raw Score (base)	Absolute difference	<i>t</i>-value	Sig (<i>t</i>-value)¹
Mathematics	27.56 (61)	29.49 (61)	1.93	1.997	0.0504

We also performed an Effect Size analysis (see *Technical Post Script on page 18 for a full description of Effect Size and its interpretation*). **The Effect Size analysis confirms that there was significant learning improvement for the treatment group, but not for the control group from pre to post** (see table below).

Analysis Performed:	Cohen's d^2
Control Pre to Post	0.204

In addition, we used an ANCOVA to determine whether individual teachers had an impact on student performance, and, if there was a teacher interaction, whether this interaction favored one of the groups, or neither. The measure used in this analysis was the raw score of the TerraNova® Complete Battery Plus exam.

¹Shaded values are statistically significant at the 95% level of confidence or higher.

²Shaded values are educationally significant (0.25 or higher)¹ (Tallmadge, G. (1977). *The joint dissemination review panel idea book*. Washington, DC: National Institute of Education and the U.S. Office of Education).

*No judgment is made here as to whether these should or should not be considered "meaningful" improvements, a distinction sometimes applied to observed real world changes in school/district performance on standardized tests.

This analysis shows that there were **significant post-test differences for students between the individual teachers participating in this study, independent of the textbook program being used**. However, there were **no significant differences for individual teachers in their treatment versus control class performance** (see table below).

Source of Possible Variation:	F-value	Sig (F) ¹
Individual Teacher Effect	22.121	0.000
Teacher by Section Interaction	0.767	0.515

Performance on mathematics instructional objectives for the treatment and control groups was also analyzed. The measure used in this analysis was the OPI (Objectives Performance Index) of the TerraNova® Complete Battery Plus exam (see *Technical Post Script on page 18 for a full description of this measure*).

Consistent with their significant overall improvement in test scores, treatment students showed significant gains in six out of seven diagnostic skills areas for mathematics (as defined by CTB/McGraw Hill). Control students showed significant gains in only four out of the seven skills.

CTB/McGraw Hill TerraNova® Mathematics Instructional Objectives (OPI):	Treatment Mean Point Gain¹ (pre to post)	Control Mean Point Gain¹ (pre to post)
Numbers and Number Relations	+3.58	+3.27
Computation and Estimation	+6.06	+5.29
Measurement	+5.03	+3.41
Geometry and Spatial Sense	+5.20	+3.64
Data, Statistics and Probability	+4.69	+3.61
Patterns, Functions, Algebra	+5.80	+4.41
Problem Solving and Reasoning	+6.08	+6.81
TOTAL GAIN	+36.44	+30.44

Conclusions

The treatment group using the Prentice Hall *Mathematics, Course 2* textbook program showed significant overall learning improvement over the course of a full school year, achieving gains in six mathematics diagnostic skills areas. The control group, using other textbook programs, did not achieve significant learning improvement overall during the same period, but did show gains in four mathematics skills areas.

Analysis of objectives performance scores suggests that the treatment group achieved greater strength than the control group in two skill areas: Geometry and Spatial Sense and Data, Statistics, and Probability. Achievement in these areas, along with a strong performance in the other five mathematics diagnostic skills, contributed to the treatment group's overall success.

Individual teacher interactions were demonstrated across both the treatment and control classes they taught, a possible indicator that teacher skill and experience played a large role in student outcomes. These interactions did not favor the treatment group or the control group.

¹Shaded values are significant at the 95% level of confidence or higher.

For the purpose of completeness, we include here a brief discussion of the statistical tests and measures that were used. All of the statistical tests are in the classical statistical domain and are broadly used across all disciplines including psychometrics. The program used in the computational steps of this project was SPSS version 11.5.1.

The ***t-test*** that was used states: *The mean of one population is equal to the mean of another population.* The hypothesis is that the means are equal; the alternative hypothesis is that they are unequal. We use a two-sided test. The means and the variance are calculated and the t-statistic is computed. The significance of the t-statistic then computed. If the significance of the t-statistic is less than .05 the hypothesis is rejected; otherwise, the alternative hypothesis is accepted.

Technical Postscript

ANCOVA: The dependent variable (in this case the post-test score) is equal for both populations (treatment and control) after adjustment for the covariate (in this case the pre-test score). A number of assumptions are invoked in the ANCOVA: population variances about the regression lines; that the regression curve is a straight line; the pre- and post-test values are approximately normally distributed. The F-value is the statistic that is computed from a series of computations of sums of squares and sums of products. The distribution of this F-value is known and the significance may be computed. If the significance of the F-value is less than .05, the hypothesis for equal means of the post-test is rejected; otherwise, it is accepted.

Raw Score: “The number-correct or ‘raw’ score is the number of items answered correctly by a student on any given test section.”¹ (The mathematics section of the TerraNova® Complete Battery Plus, Level 17 includes a total of 57 questions; therefore, the “raw” score for each student equals the number correct out of 57).

¹Quoted directly from: *Beyond the Numbers, A Guide to Interpreting and Using the Results of Standardized Achievement Tests*, page 18, CTB/McGraw-Hill 2003.

Objectives Performance Index: “All versions of TerraNova® and TerraNova®, the Second Edition, yield criterion-referenced scores reported in terms of an OPI. An OPI is reported for each of the instructional objectives measured by the particular version of TerraNova® and TerraNova®, the Second Edition. Each objective is measured by at least four items. CTB has established a standard of four items as the minimum number needed to produce reliable information regarding objective mastery.... The OPI makes test results both understandable and useful for the teacher in planning effective learning strategies and activities. The OPI is an estimate of the number of items that a student would be expected to answer correctly if there had been 100 similar items for that objective. The OPI scale runs from ‘0’ to ‘100’.”²

²Quoted directly from: *Beyond the Numbers, A Guide to Interpreting and Using the Results of Standardized Achievement Tests*, page 12, CTB/McGraw-Hill 2003.

Effect Size: “Effect size (ES) is a name given to a family of indices that measure the magnitude of a treatment effect.”³ Effect size is “simply a way of quantifying the effectiveness of a particular intervention, relative to some comparison intervention. It is easy to calculate, readily understood and can be applied to any measured outcome in Education or Social Science.”⁴ Cohen’s *d* is commonly used to calculate effect size. Cohen’s *d* is a “standardized” measure found by dividing the difference in the group mean scores by the standard deviation (adjusted for sample size differences known as the “pooled” estimate of the standard deviation). Whereas “statistical significance” is based on a specific probability [“p-value”], it can be affected much more by sample size than effect size’s interpretation. In educational research, an effect size of 0.25 or more is commonly considered to be “educationally significant.”⁵

³Quoted directly from: University of Colorado at Colorado Springs website, link: <http://web.uccs.edu/lbecker/Psy590/es.htm>.

⁴Quoted directly from: Curriculum, Evaluation and Management Centre (CEM) website, link: <http://cem.dur.ac.uk/ebeuk/research/effectsize/ESguide.htm>.

⁵Tallmadge, G. (1977). *The joint dissemination review panel idea book*. Washington, DC: National Institute of Education and the U.S. Office of Education.

Appendix

William M. Bailey, Statistician, WMB & Associates, Orlando, FL

William M. Bailey is an independent statistician and market analyst with 15+ years of experience in the design, execution, interpretation and reporting of qualitative and quantitative research studies both in the education and consumer fields.

His experience with the U.S. Department of Education includes:

- Statistical analysis on several studies tracking pre and post course design among a broad-based selection of community colleges and universities
- Serving as an on-call statistical and methodological consultant to the Research to Practice Division
- Serving as a reviewer of proposals to the Office of Special Education & Programs (OSEP)

Mr. Bailey also serves as a beta test site for SPSS, working with development of their core statistical package since version 7 and is now involved with version 12. He also is assisting in the design and testing of the new complex sampling module.

CTB/McGraw Hill, Monterey, CA

CTB/McGraw Hill, publisher of the TerraNova[®], is a division of the McGraw Hill Companies. It was founded in 1926 and provides a number of standardized achievement tests for both children and adults. The company scores over 20 million test documents each year.⁶

⁶Source: www.CTB.com

2002 Algebra 1 Program Efficacy Study

Abstract

This study investigated the effects of algebra 1 textbook programs on student performance. High school algebra 1 students were assigned to either a treatment group (using Prentice Hall pre-publication chapters) or a control group (using their current textbook program). Students were tested at the start of the semester with a nationally normed standardized test (the TerraNova® Algebra exam). At the end of the semester-long treatment period, students were retested with the same standardized test. Both the treatment students using the Prentice Hall *Algebra 1* ©2004 pre-publication chapters and the control students using other algebra 1 textbook programs showed significant learning improvement over the course of one semester.

Objective

The main objective of this project was to determine whether students who were enrolled in classes using Prentice Hall *Algebra 1* ©2004 significantly increased their algebra knowledge and skills after using the pre-publication chapters for one semester. The measures that were used were the NCE (Normal Curve Equivalent) and the OPI (Objectives Performance Index) of the TerraNova® Algebra exam.

Methodology

The study was designed so that there was a treatment group to whom the Prentice Hall program was administered and a control group to whom the program was not administered. The treatment group used pre-production chapters of the Prentice Hall *Algebra 1* ©2004 program. The control group used the full algebra 1 textbook program that had been previously adopted for use in the school.

Both groups were tested in January 2002, prior to the program's introduction, and then again in May 2002, at the end of the school year. Therefore, for both the treatment group and the control group there was a pre-test score and a post-test score. Only students who completed both the pre- and post-tests were included in this analysis.

A total of eight teachers and sixteen ninth grade algebra 1 classes comprising 240 students participated in the study. In a few cases, tenth and eleventh grade students were integrated into the ninth grade classes.

	Treatment Group Prentice Hall <i>Algebra 1</i> Textbook Program	Control Group Other Algebra 1 Textbook Programs
Pre-Test (administered during 1 st week of semester)	8 classes 128 students	8 classes 112 students
Post-Test (administered at end of school year)	8 classes 128 students	8 classes 112 students

In each case, one teacher taught both treatment and control classes. The intervention being studied comprised of a mainstream textbook with which teachers were generally familiar. Therefore, training was limited to on-site in-service designed to ensure that teachers understood the treatment program and could instruct with it as its designers intended. Additionally, teachers (and administrators) were repeatedly briefed, and provided with materials that explained the study and reinforced the need to implement it as designed without, for example, contamination of the program materials or approaches between their treatment and control classes. Discussions with and short surveys of the teachers helped establish their understanding of the study's parameters.*

Both treatment and control classes were in the same school building. Treatment and control classes were selected by the teachers to be similar in student ability levels. Of the eight schools participating in this research project, four schools were in rural settings, three were in suburban settings, and one was in an urban setting. The schools had a range of sizes. Study participants were from five states: Florida, Illinois, Indiana, New Jersey, and North Carolina.

Statistical controls and tests were used to examine the following issues with regard to program effectiveness:

- 1) Whether the pre-test scores for the treatment group and the control group showed significant differences at the starting point of the study (despite being pre-selected as evenly matched);
- 2) Whether overall algebra knowledge and skills increased/decreased, or stayed the same from the pre-test to the post-test among students using the Prentice Hall *Algebra 1* pre-publication chapters (the treatment group) and among students using their incumbent textbook program (the control group); and
- 3) Whether there were differential effects between teachers or control group textbook programs used.

Analysis was conducted only for students who remained in the study from pre- through post-testing. A total of 29 students were excluded, primarily because they were absent on the day of the post-test (that is, about 1-2 students per class). There appears to be no evidence of differential mortality between treatment and control groups in terms of student ability levels that might skew study results.

Step 1a. To remove any spurious student data, an outlier analysis was completed to eliminate student scores that fell outside the normal curve distribution. The measure used in this analysis was the overall score, the NCE of the TerraNova® Algebra exam. An outlier was defined as any student with an NCE score falling 2 or more standard deviations from the mean. A total of 31 students had a pre-test and/or post-test score that qualified as an outlier and were thus removed from the data set.

Step 1b. Once the outlying scores were removed from the data set, an analysis of whether or not there was a statistically significant difference between the treatment group and the control group on the overall pre-test score was performed. For this, a t-test on the difference between the treatment and control mean pre-test scores was used. The measure used in this analysis was the NCE of the TerraNova® CTBS Complete Battery Plus.

*It has been well documented that teacher variation in experience, style and effectiveness can have a profound impact on student outcomes. Few, if any, instruments have been shown to be reliable in measuring teacher variability, such that it would be possible to select a priori “matched” teachers for treatment and control groups, or to effectively “balance” the groups on this factor after the study. Therefore, to neutralize this factor between the treatment and control groups, a single teacher taught both groups, with a clear understanding of the necessity to stay true to the instruction in each program.

The hypothesis of the t-tests shown below is that the pre-test means of the treatment and control groups are equal; the alternative hypothesis is that they are not equal, for which a two-tail test is appropriate. If the significance of the t-value is less than or equal to .10, then the hypothesis is rejected at the 90% confidence level and the alternate hypothesis is accepted.

Since the significance of the t-value is greater than .10 (that is, 0.155, as shown below), we accept the hypothesis; that is, that the pre-test means of the treatment and control groups are equal at the 90% level of confidence, showing **no significant differences at the starting point of the study.**

Subject	Treatment pre-test NCE (base)	Control pre-test NCE (base)	Absolute difference	t-value	Sig (t-value) ¹
Algebra 1	42.86 (128)	40.41 (112)	2.45	1.43	0.155

Step 2a. This step consisted of an analysis of whether or not there was a statistically significant difference between the pre-test and post-test scores within the treatment group and within the control group. For each group, a t-test on the difference between pre-test and post-test scores was used. The NCE was the measure used in this analysis.

The hypothesis of the t-tests shown below is that the pre-test and post-test means are equal; the alternative hypothesis is that they are not equal. If the significance of the t-value is less than or equal to .10, then the hypothesis is rejected at the 90% confidence level and the alternate hypothesis is accepted.

For the treatment group, the significance of the t-value is less than .10 (that is, 0.016, as shown below). Therefore, we accept the alternative hypothesis.

- **Thus, students using the Prentice Hall *Algebra 1* pre-publication chapters showed significant improvement in test scores from the pre-test to the post-test.***

For the control group, the significance of the t-value is also less than .10 (that is, 0.006, as shown below). Therefore, we accept the alternative hypothesis.

- **Thus, students using other textbook programs also showed significant improvement in test scores from the pre-test to the post-test.***

Subject	Treatment pre-test NCE (base)	Treatment post-test NCE (base)	Absolute difference	t-value	Sig (t-value) ¹
Algebra 1	42.86 (128)	46.35 (128)	3.49	2.44	0.016

Subject	Control pre-test NCE (base)	Control post-test NCE (base)	Absolute difference	t-value	Sig (t-value) ¹
Algebra 1	40.41 (112)	43.79 (112)	3.38	2.79	0.006

¹Shaded values are significant at the 90% level of confidence or higher.

*No judgment is made here as to whether these should or should not be considered “meaningful” improvements, a distinction sometimes applied to observed real world changes in school/district performance on standardized tests.

Step 2b. An Analysis of Covariance (ANCOVA) was also performed to test the difference between the treatment and control groups on the post-test, adjusting for the pre-test level. The NCE was the measure used in this analysis.

The hypothesis of the ANCOVA is that the post-test means of treatment and control classes are equal after the adjustment for the pre-test means. The alternative hypothesis is that they are not equal. If the significance of the F-value is less than .10, then the hypothesis is rejected at the 90% confidence level and the alternative hypothesis is accepted.

For the post-test means, the significance of the F-value is greater than .10 (that is, 0.470, as shown below). Therefore, we accept the original hypothesis.

- **Thus, the treatment and control groups showed parity on the post-test scores.**

Subject	Treatment post-test NCE (base)	Control post-test NCE (base)	Absolute difference	F-value	Sig (F) ¹
Algebra 1	46.35 (128)	43.79 (112)	2.56	0.52	0.470

Step 3. In addition to the above analyses, an ANCOVA was performed to determine whether individual teachers or different competitive textbook programs used in the control classes had an impact on student performance.

a) Teachers

The hypothesis of the ANCOVA is that the post-test means of the teacher classes are equal after the adjustment for the pre-test means. The alternative hypothesis is that they are not equal. If the significance of the F-value is less than .10, then the hypothesis is rejected at the 90% confidence level and the alternative hypothesis is accepted.

For the individual teacher main effect, the significance of the F-value is less than .10 (that is, 0.003, as shown below). Therefore, we accept the alternative hypothesis.

- **Thus, there were significant post-test differences for students between the individual teachers participating in this study, independent of the textbook program being used.**

There is a more detailed hypothesis which tests for the differences between the means of the teachers by section (treatment and control). For the teacher by section interaction, the significance of the F-value is also less than .10 (that is, 0.035, as shown below). Therefore, we accept the alternative hypothesis.

- **Thus, there were significant differences for individual teachers in their treatment versus control class performance. However, the pattern of difference was not uniform, suggesting other, unmeasured factors were in effect, perhaps including program implementation issues.**

Source of Possible Variation:	Sum of Squares	Degrees of Freedom	Mean Sum of Squares	F-value	Sig (F) ¹
Individual Teacher Effect	2833.52	7	404.79	3.23	0.003
Teacher by Section Interaction	1932.29	7	276.04	2.21	0.035

¹Shaded values are significant at the 90% level of confidence or higher

b) Competitive Programs

The hypothesis of the ANCOVA is that the post-test means of the control classes are equal after the adjustment for the pre-test means. The alternative hypothesis is that they are not equal. If the significance of the F-value is less than .10, then the hypothesis is rejected at the 90% confidence level and the alternative hypothesis is accepted.

For the competitive program main effect, the significance of the F-value is greater than .10 (that is, 0.619, as shown below). Therefore, we accept the original hypothesis.

- **Thus, the different competitive textbook programs used in the control classes did not have a differential impact on student performance.**

Source of Possible Variation:	Sum of Squares	Degrees of Freedom	Mean Sum of Squares	F-value	Sig (F) ¹
Program Effect	31.10	1	31.10	0.25	0.619

Conclusions

Both the treatment group (using the Prentice Hall *Algebra 1* pre-publication chapters) and the control group (using other algebra 1 textbook programs) showed significant learning improvement over the course one semester of study.

Different textbook programs used in the control classes did not have an impact on overall student performance. However, there were significant teacher effects found.

- Individual teacher effects were demonstrated across both the treatment and control classes they taught, a possible indicator that teacher skill and experience played a large role in student outcomes. These effects did not favor treatment or control.
- Program implementation issues may explain some of the teacher effects found between the treatment and control classes. Although there was not a uniform pattern of teacher effects found in the data between treatment and control classes, results of a survey completed by each teacher at the conclusion of the study suggest that use of the pre-publication form of the Prentice Hall program may account for some of these differences. Teachers had only a partial Prentice Hall Teacher's Edition and a few ancillary materials for the treatment class, while having a full textbook program for the control class. Some teachers expressed discomfort with having only partial supplementary materials with the treatment class, which sometimes resulted in limited usage of these items.
- A reliable measure of teacher variability in skill level and experience is needed for studies of this kind to help isolate program effects from teacher effects.

Technical Post Script

For the purpose of completeness, we include here a brief discussion of the statistical tests and measures that were used. All of the statistical tests are in the classical statistical domain, and are broadly used across all disciplines including psychometrics. The programs that were used in the computational steps of this project were SPSS versions 6/9.

The ***t-test*** that was used states: *The mean of one population is equal to the mean of another population when the variance is unknown.* The hypothesis is that the means are equal; and the alternative hypothesis is that they are unequal; for which we use a two sided test. The means and the variance are calculated and the t-statistic is computed. The significance of the t-statistic then computed. If the significance of the t-statistic is less than .05 the hypothesis is rejected; otherwise, the alternative hypothesis is accepted.

ANCOVA: The dependent variable (in this case the post-test score) is equal for both populations (treatment and control) after adjustment for the covariate (in this case the pre-test score). A number of assumptions are invoked in the ANCOVA: population variances about the regression lines; the regression curve is a straight line; and the pre- and post-test values are approximately normally distributed. The F-value is the statistic that is computed from a series of computations of sums of squares and sums of products. The distribution of this F-value is known and the significance may be computed. If the significance of the F-value is less than .05, the hypothesis for equal means of the post-test is rejected; otherwise, it is accepted.

Normal Curve Equivalent: “Comparison of Scores across Tests. The NCE scale, ranging from 1 to 99, coincides with the national percentile scale (NP) at 1, 50, and 99. NCEs have many of the same characteristics as percentile ranks, but have the additional advantage of being based on an equal-interval scale. The difference between two successive scores on the scale has the same meaning throughout the scale. This property allows you to make meaningful comparisons among different achievement test batteries and among different tests within the same battery. You can compare NCEs obtained by different groups of students on the same test or test battery by averaging the test scores for the groups.”¹

¹Quoted directly from: *Teacher's Guide to TerraNova*, page 138, CTB/McGraw-Hill 1999.

Objectives Performance Index: “The OPI makes test results both understandable and useful for the teacher in planning effective learning strategies and activities. The OPI is an estimate of the number of items that a student would be expected to answer correctly if there had been 100 similar items for that objective....The OPI scale runs from ‘0’ for total lack of mastery to ‘100’ for complete mastery. For CTB achievement tests, OPI scores between 0 and 49 are regarded as the Non-Mastery level. Scores between 50 and 74 are regarded as indications of Partial Mastery. Scores of 75 and above are regarded as the Mastery level.”²

²Quoted directly from: *Beyond the Numbers, A Guide to Interpreting and Using the Results of Standardized Achievement Tests*, page 11, CTB/McGraw-Hill 1997.

All tests were scored by CTB/McGraw-Hill, the publisher of TerraNova. Statistical analyses and conclusions were performed by an independent firm, Pulse Analytics Inc., Ridgewood, New Jersey.

Appendix: School District Profiles

School District 1 FL

Enrollment: 62,011

Ethnicity: Asian 1%, African-American 16%, Hispanic 8%, Caucasian 75%

Poverty Level: 17% (Med-High)

College Bound Students: 73%

School District 2 IL

Enrollment: 181

Ethnicity: African-American 1%, Caucasian 99%

Poverty Level: 16% (Med-High)

College Bound Students: 80%

School District 3 IN

Enrollment: 2,389

Ethnicity: Hispanic 6%, Caucasian 93%

Poverty Level: 12% (Low-Med)

College Bound Students: 62%

School District 4 IN

Enrollment: 10,500

Ethnicity: African-American 20%, Hispanic 1%, Caucasian 79%

Poverty Level: 23% (Med-High)

College Bound Students: 58%

School District 5 NJ

Enrollment: 2,289

Ethnicity: Asian 2%, Hispanic 3%, Caucasian 95%

Poverty Level: 1% (Low)

College Bound Students: not available

School District 6 NJ

Enrollment: 4,708

Ethnicity: Asian 5%, African-American 1%, Hispanic 6%, Caucasian 87%

Poverty Level: 2% (Low)

College Bound Students: not available

School District 7 NC

Enrollment: 16,500

Ethnicity: Asian 0%, African-American 31%, Hispanic 4%, Caucasian 63%, Native American 1%

Poverty Level: 22% (Med-High)

College Bound Students: 81%

School District 8 NC

Enrollment: 9,285

Ethnicity: Asian 0%, African-American 24%, Hispanic 1%, Caucasian 75%

Poverty Level: 11% (Low-Med)

College Bound Students: 78%

* Source: Market Data Retrieval

2000 Pre-Algebra Program Efficacy Study

Abstract

This study investigated the effects of pre-algebra textbook programs at the eighth grade level. Eighth grade pre-algebra students were assigned to either a study group (Prentice Hall text) or a control group (current text). Students were tested at the start of the academic year with a nationally normed standardized test (the TerraNova[®] CTBS Complete Battery Plus). At the end of the full year treatment period, students were re-tested with the same standardized test. The eighth grade *study* students utilizing the Prentice Hall *Pre-Algebra* program showed significant learning improvement, while the control students using other pre-algebra textbook programs did not.

Objectives

The main objective of this project was to determine whether students who were enrolled in classes using Prentice Hall *Pre-Algebra: Tools for a Changing World* significantly increased their pre-algebra knowledge and skills after using the program. An additional objective was to measure whether students using their incumbent (non-Prentice Hall) textbook program significantly increased their pre-algebra knowledge and skills during the same time period. The measure that was used was the NCE (Normal Curve Equivalent) of the TerraNova[®] CTBS Complete Battery Plus.

Methodology

The study was designed so that there was a study group to whom the Prentice Hall program was administered and a control group to whom the program was not administered. The control group used whatever pre-algebra program had been adopted for use in the school.

Both groups were tested in September 2000, prior to the program's introduction, and then again in May 2001, at the end of the school year. Therefore, for both the study group and the control group there was a pre-test score and a post-test score. Only students who completed both the pre- and post-tests were included in this analysis.

A total of eight eighth grade pre-algebra classes comprising 120 students participated in the study.

	Study Group Prentice Hall <i>Pre-Algebra</i> Textbook Program	Control Group Other Textbook Programs
Pre-Test (administered during 1 st week of school)	4 classes 65 students	4 classes 55 students
Post-Test (administered at end of school year)	4 classes 65 students	4 classes 55 students

In each case, one teacher taught both study and control classes. Both study and control classes were in the same school building. Study and control classes were selected to be similar in student ability levels. Of the four schools participating in this research project, two schools were in urban settings, one was in a suburban setting, and one was in a rural setting. These schools had a range of sizes. Study participants were from three states: Colorado, Illinois, and Washington.

Statistical tests were used to examine the following issues with regard to program effectiveness:

- 1) whether the pre-test scores for the study group and the control group showed significant differences at the starting point of the study;
- 2) whether pre-algebra knowledge and skills increased, decreased, or stayed the same from the pre-test to the post-test among students using Prentice Hall *Pre-Algebra* (the study group); and whether pre-algebra knowledge and skills increased, decreased, or stayed the same from the pre-test to the post-test among students using their incumbent textbook program (the control group).

Step 1. This step consisted of an analysis of whether or not there was a statistically significant difference between the study group and the control group on the pre-test score. For this, a *t-test* on the difference between the study and control mean pre-test scores was used. The measure used in this analysis was the NCE of the TerraNova® CTBS Complete Battery Plus.

The hypothesis of the *t-tests* shown below is that the pretest means of the study and control groups are equal; the alternative hypothesis is that they are not equal, for which a two tail test is appropriate. If the significance of the t-value is less than or equal to .05, then the hypothesis is rejected at the 95% (5% significance) level and the alternate hypothesis is accepted.

Since the significance of the t-value is greater than .05 (that is, 0.949, as shown below), we accept the hypothesis; that is, that the pre-test means of the study and control groups are equal at the 95% level of confidence, showing no significant differences between the study and control groups at the starting point of the study.

Subject, Grade	Study pre-test NCE (base)	Control pre-test NCE (base)	Absolute difference	t-value	Sig (t-value)
Pre-Algebra, 8 th Grade	43.94 (65)	44.13 (55)	0.19	0.06	0.949

Step 2. This step consisted of an analysis of whether or not there was a statistically significant difference between the pre-test and post-test scores within the study group and within the control group. For each group, a t-test on the difference between pre-test and post-test scores was used. The NCE was the measure used in this analysis.

The hypothesis of the t-tests shown below is that the pre-test and post-test means are equal; the alternative hypothesis is that they are not equal. If the significance of the t-value is less than or equal to .05, then the hypothesis is rejected at the 95% (5% significance) level and the alternate hypothesis is accepted.

For the study group, the significance of the t-value is less than .05 (that is, 0.006, as shown below). Therefore we accept the alternative hypothesis: the pre-test and post-test means of the study group are not equal at the 95% level of confidence.

- Thus, students using the Prentice Hall *Pre-Algebra* program showed significant improvement in test scores from the pre-test to the post-test.

In contrast, for the control group, the significance of the t-value is greater than .05 (that is, 0.149, as shown below). Therefore we accept the original hypothesis: the pre-test and post-test means of the control group are equal at the 95% level of confidence.

- Thus, students using other textbook programs did not show significant improvement in test scores from the pre-test to the post-test.

Subject, Grade	Study pre-test NCE (base)	Study post-test NCE (base)	Absolute difference	t-value	Sig (t-value)
Pre-Algebra, 8 th Grade	43.94 (65)	47.49 (65)	3.55	2.83	0.006

Subject, Grade	Control pre-test NCE (base)	Control post-test NCE (base)	Absolute difference	t-value	Sig (t-value)
Pre-Algebra, 8 th Grade	44.13 (55)	46.13 (55)	2.00	1.46	0.149

Conclusions

The study group, using the Prentice Hall *Pre-Algebra: Tools for a Changing World* program, showed significant learning improvement over the course of a full school year; the control group, using other textbook programs, did not show significant learning improvement over this same period.

Technical Post Script

For the purpose of completeness, we include here a brief discussion of the statistical tests and measures that were used. All of the statistical tests are in the classical statistical domain, and are broadly used across all disciplines including psychometrics. The programs that were used in the computational steps of this project were SPSS versions 6/9.

The *t-test* that was used states: *The mean of one population is equal to the mean of another population when the variance is unknown.* The hypothesis is that the means are equal; and the alternative hypothesis is that they are unequal; for which we use a two-sided test. The means and the variance are calculated and the *t*-statistic is computed. The significance of the *t*-statistic then computed. If the significance of the *t*-statistic is less than .05 the hypothesis is rejected; otherwise, the alternative hypothesis is accepted.

Normal Curve Equivalents: "Comparison of Scores across Tests. The NCE scale, ranging from 1 to 99, coincides with the national percentile scale (NP) at 1, 50, and 99. NCEs have many of the same characteristics as percentile ranks, but have the additional advantage of being based on an equal-interval scale. The difference between two successive scores on the scale has the same meaning throughout the scale. This property allows you to make meaningful comparisons among different achievement test batteries and among different tests within the same battery. You can compare NCEs obtained by different groups of students on the same test or test battery by averaging the test scores for the groups."¹

¹Quoted directly from: Teacher's Guide to TerraNova, page 138, CTB/McGraw-Hill 1999.

All tests were scored by CTB/McGraw-Hill, the publisher of TerraNova®. Statistical analyses and conclusions were performed by an independent firm, Pulse Analytics Inc., Ridgewood, New Jersey.

2000 Course 3 Middle School Mathematics Program Efficacy Study

Abstract

This study investigated the effects of mathematics textbook programs on student performance. Eighth grade math students were assigned to either a study group (using the Scott Foresman-Addison Wesley *Middle School Math, Course 3* ©1999 textbook program) or control group (using their current mathematics textbook program). Students were tested at the start of the academic year with a nationally normed standardized test (the TerraNova® CTBS Complete Battery Plus, Level 18, Form A). At the end of the full-year treatment period, students were re-tested with the same standardized test. Both the study students utilizing the Scott Foresman-Addison Wesley *Middle School Math, Course 3* ©1999 textbook program and the control students using other mathematics textbook programs showed significant learning improvement over the course of the school year.

Objective

The main objective of this project was to determine whether students who were enrolled in classes using the Scott Foresman-Addison Wesley *Middle School Math, Course 3* ©1999 program significantly increased their mathematics knowledge and skills after using the pre-publication chapters for a full school year. An additional objective was to measure whether students using their incumbent textbook program significantly increased their mathematics knowledge and skills during the same time period. The measures that were used were the NCE (Normal Curve Equivalent) and the OPI (Objectives Performance Index) of the TerraNova® CTBS Complete Battery Plus.

Methodology

The study was designed so that there was a study group to whom the Scott Foresman-Addison Wesley program was administered and a control group to whom the program was not administered. The control group used whatever mathematics program had previously been adopted for use in the school.

Both groups were tested in September 2000, prior to the program's introduction, and then again in May 2001, at the end of the school year. Therefore, for both the study group and the control group there was a pre-test score and a post-test score. Only students who completed both the pre- and post-tests were included in this analysis.

A total of ten eighth grade math classes comprising 185 students participated in the study:

	Study Group SF-AW Middle School Math Textbook Program	Control Group Other Textbook Programs
Pre-Test (administered during 1 st week of school)	5 classes 100 students	5 classes 85 students
Post-Test (administered at end of school year)	5 classes 100 students	5 classes 85 students

In all cases, one teacher taught both study and control classes. Both study and control classes were in the same school building. Study and control classes were selected to be similar in student ability levels. Of the five schools participating in this research project, three schools were in suburban settings, and two were in urban settings. These schools had a range of sizes. Study participants were from three states: Colorado, New Jersey, and Washington.

Statistical tests were used to examine the following issues with regard to program effectiveness:

- 1) Whether the pre-test scores for the study group and the control group showed significant differences at the starting point of the study;
- 2) Whether overall mathematics knowledge and skills increased, decreased, or stayed the same from the pre-test to the post-test among students using the Scott Foresman-Addison Wesley program (the study group) and among students using their incumbent textbook program (the control group); and
- 3) Whether the students in the study and control groups showed significant learning improvement in key mathematics diagnostic areas.

Step 1. This step consisted of an analysis of whether or not there was a statistically significant difference between the study group and the control group on the pre-test score. For this, a *t-test* on the difference between the study and control mean pre-test scores was used. The measure used in this analysis was the NCE of the TerraNova® CTBS Complete Battery Plus.

The hypothesis of the *t-tests* shown below is that the pre test means of the study and control groups are equal; the alternative hypothesis is that they are not equal, for which a two tail test is appropriate. If the significance of the *t-value* is less than or equal to .05 then the hypothesis is rejected at the 95% (5% significance) level and the alternate hypothesis is accepted.

Since the significance of the *t-value* is greater than .05 (that is, 0.306, as shown below), we accept the hypothesis; the pre-test means of the study and control groups are equal at the 95% level of confidence, showing no significant differences at the starting point of the study.

Subject, Grade	Study Pre-test NCE (base)	Control Pre-test NCE (base)	Absolute difference	<i>t-value</i>	Sig (<i>t-value</i>)
Math, 8th Grade	47.06 (100)	44.95 (85)	2.11	1.03	0.306

Step 2. This step consisted of an analysis of whether or not there was a statistically significant difference between the pre-test and post-test scores within the study group and within the control group. For each group, a *t-test* on the difference between pre-test and post-test scores was used. The NCE was the measure used in this analysis.

The hypothesis of the *t-tests* shown below is that the pre-test and post-test means are equal; the alternative hypothesis is that they are not equal. If the significance of the *t-value* is less than or equal to .05 then the hypothesis is rejected at the 95% (5% significance) level and the alternate hypothesis is accepted.

For the study group, the significance of the *t-value* is less than .05 (that is, 0.001, as shown below). Therefore, we accept the alternative hypothesis: the pre-test and post-test means of the study group are not equal at the 95% level of confidence.

- Thus, students using the Scott Foresman-Addison Wesley *Middle School Math* program showed significant improvement in overall test scores from the pre-test to the post-test.

For the control group, the significance of the *t-value* is also less than .05 (that is, 0.003 as shown on page 35).

- Thus, students using other textbook programs also showed significant improvement in overall test scores from the pre-test to the post-test.

Subject, Grade	Study Pre-test NCE (base)	Study Post-test NCE (base)	Absolute difference	t-value	Sig (t-value) ¹
Math, 8 th Grade	47.06 (100)	50.97 (100)	3.91	3.30	0.001

Subject, Grade	Control Pre-test NCE (base)	Control Post-test NCE (base)	Absolute difference	t-value	Sig (t-value) ¹
Math, 8 th Grade	44.95 (85)	48.42 (85)	3.47	3.07	0.003

Step 3. In addition to NCE scores, performance on mathematics instructional objectives for the study and control groups was also analyzed. The measure used in this analysis was the OPI of the TerraNova® CTBS Complete Battery Plus.

Consistent with the overall improvement in test scores, both students using Scott Foresman-Addison Wesley *Middle School Math, Course 3* ©1999 (the study group) and students using other textbook programs (the control group) showed significant gains in all key mathematics instructional areas over the course of a full school year. For both groups, the highest gains were achieved in the areas of: Patterns, Functions, Algebra; Problem Solving and Reasoning; and Number & Number Relations.

8 th Grade Mathematics Instructional Objectives:	Study	Control
	Mean Point Gain ¹ (pre to post)	Mean Point Gain ¹ (pre to post)
Number & Number Relations	+10.12	+9.28
Computation and Numerical Estimation	+8.97	+9.02
Measurement	+7.77	+7.31
Geometry & Spatial Sense	+9.27	+6.45
Data Analysis, Statistics & Probability	+8.32	+8.11
Patterns, Functions, Algebra	+12.00	+10.13
Problem Solving & Reasoning	+9.37	+10.53
TOTAL GAIN	+65.82	+60.83

¹Shaded values are significant at the 95% level of confidence or higher.

Conclusions

Students who used the Scott Foresman-Addison Wesley *Middle School Math, Course 3* ©1999 program showed significant overall learning improvement over the course of a full school year. Students using other textbook programs also showed significant learning improvement over this same period.

Consistent with the overall findings, students in both the study and control groups showed significant improvement on all key mathematics instructional objectives. The largest gains were accomplished in the areas of Patterns, Functions, Algebra; Problem Solving and Reasoning; and Number & Number Relations.

Technical Post Script

For the purpose of completeness, we include here a brief discussion of the statistical tests and measures that were used. All of the statistical tests are in the classical statistical domain and are broadly used across all disciplines including psychometrics. The programs that were used in the computational steps of this project were SPSS versions 6/9.

T-test: *The mean of one population is equal to the mean of another population when the variance is unknown.* The hypothesis is that the means are equal; and the alternative hypothesis is that they are unequal; for which we use a two-sided test. The means and the variance are calculated and the t-statistic is computed. The significance of the t-statistic then computed. If the significance of the t-statistic is less than .05 the hypothesis is rejected; otherwise, the alternative hypothesis is accepted.

Normal Curve Equivalent - Comparison of Scores across Tests: The normal curve equivalent (NCE) scale, ranging from 1 to 99, coincides with the national percentile scale (NP) at 1, 50, and 99. NCEs have many of the same characteristics as percentile ranks, but have the additional advantage of being based on an equal-interval scale. The difference between two successive scores on the scale has the same meaning throughout the scale. This property allows you to make meaningful comparisons among different achievement test batteries and among different tests within the same battery. You can compare NCEs obtained by different groups of students on the same test or test battery by averaging the test scores for the groups.”¹

¹Quoted directly from: Teacher’s Guide to TerraNova®, page 138, CTB/McGraw-Hill 1999.

Objectives Performance Index: “The OPI is an estimate of the number of items that a student would be expected to answer correctly if there had been 100 similar items for that objective.... The OPI scale runs from ‘0’ for total lack of mastery to ‘100’ for complete mastery. For CTB achievement tests, OPI scores between 0 and 49 are regarded as the Non-Mastery level. Scores between 50 and 74 are regarded as indications of Partial Mastery. Scores of 75 and above are regarded as the Mastery level.”²

²Quoted directly from: *Beyond the Numbers, A Guide to Interpreting and Using the Results of Standardized Achievement Tests*, page 11, CTB/McGraw-Hill 1997.

All tests were scored by CTB/McGraw-Hill, the publisher of TerraNova®. Statistical analyses and conclusions were performed by an independent firm, Pulse Analytics Inc., Ridgewood, New Jersey.

1999 Course 1 Middle Grades Mathematics & Algebra Program Efficacy Study

Abstract

This study investigates the effects of mathematics programs across three different grades. Students in sixth grade mathematics, eighth grade algebra, and ninth grade algebra were assigned to either a control group (current text) or study group (Prentice Hall text). Students were tested at the start of the academic year with a nationally normed standardized test (the TerraNova[®] CTBS Basic Battery). At the end of the full-year treatment period, students were re-tested with the same standardized test. All students increased their mathematics abilities after the year of instruction. Most notably, students in classes utilizing the Prentice Hall mathematics programs (*Middle Grades Math, Course 1* and *Algebra: Tools for a Changing World*) did significantly better as compared to the students in the control classes.

Objective

The object of this project was to determine the whether the improvement of students who were enrolled in classes using Prentice Hall *Middle Grades Math, Course 1* and *Algebra: Tools for a Changing World* is above the expected level of performance of those who were not enrolled in classes using these programs. The measure that was used was the NCE (Normal Curve Equivalent) of the TerraNova[®] CTBS Basic Battery.

Methodology

The study was designed so that there was a study group to whom the Prentice Hall program was administered and a control group to whom the program was not administered. The control group used whatever math program had previously been adopted for use in the school.

Both groups, at each grade level, were tested in September 1998, prior to the program's introduction to the study group, and then again at the end of the program in May 1999. Therefore, for each grade level there was a study group and control group, and a pre-test and a post-test score. Only students who completed both the pre- and post-tests were included in this analysis.

Three grade levels of mathematics were tested: (1) Grade 6; (2) Grade 8 Algebra; (3) Grade 9 Algebra. Twelve teachers and 350 students participated in the study.

Grade Levels	Prentice Hall Program	No. of Groups*	No. of Students
6	<i>Middle Grades Math, Course 1</i>	4	165
8, 9	<i>Algebra: Tools for a Changing World</i>	1 at Grade 8; 3 at Grade 9	185

*Each group consisted of two different classes: a control class and a study class

Where possible, one teacher taught both study and control classes. In situations where the same teacher could not teach both study and control classes, another teacher of similar teaching style, background, and tenure was selected to teach the control class. Both study and control classes were in the same school building. Study and control classes were selected to be similar in student ability levels. Of the eight schools participating in this research project, three schools were in rural settings, two were in suburban settings and three were in urban settings. Schools had a range of sizes. Study participants were from six states: California, Colorado, Illinois, New Jersey, New York, and Wisconsin.

A number of statistical tests were used to examine the issue of success of the Prentice Hall program; namely, that the post-test score for the study group was above that of the control group.

Phase 1. This phase consisted of an analysis of whether there was a statistically significant difference between the study group and the control group on the pre-test score. For this, a t-test on the difference between the study and control pre-test score for each grade level was used. In particular, for those grades for which there was a statistically significant difference, it was very important to use the Analysis of Covariance (ANCOVA) so that these differences could be adjusted for the comparison of the post-test results.

Phase 2. An ANCOVA was performed, using the F-value, to test the difference between the study and the control groups on the post-test adjusting for the pre-test level. This was necessary because the post-test scores of the two groups had to be adjusted to the relative starting points on the pre-test (that is, eliminating the effect of a higher or lower starting level as contributing to the observed post-test score).

For both phases, a significance level of .05 was used. That is, if the significance level was below 5%, (shown as **) we reject the hypothesis of equality of means.

Results of Phase 1: Test for Pre-test differences

Grade	Study pre-test NCE (base)	Control pre-test NCE (base)	Absolute Difference	t-value	Sig (t-value)
6 MGM	53.57 (87)	46.24 (78)	7.33	2.83	.005**
8 Algebra	51.28 (28)	35.71 (28)	15.57	4.22	.000**
9 Algebra	49.35 (68)	48.16 (61)	1.19	.45	.641

The hypothesis of the t-tests shown above is that the pre-test means of the study and control groups are equal; the alternative hypothesis is that they are not equal, for which a two-tail test is appropriate. If the significance of the t-value is less than or equal to .05, then the hypothesis is rejected at the 95% (5% significance) level and the alternate hypothesis is accepted.

Results of Phase 2: Test for Post-test differences adjusting for Pre-test level

Grade	Study post-test NCE	Control post-test NCE	Absolute Difference	F-Value	Sig (F-value)
6 MGM	65.80	54.39	11.42	26.92	.000**
8 Algebra	69.25	43.57	25.68	37.71	.000**
9 Algebra	59.94	52.27	7.67	4.50	.013**

The hypothesis of the ANCOVA is that the post-test means of the study and control groups are equal after the adjustment for the pretest means. The alternative hypothesis is that they are not equal. If the significance of the F-value is less than .05, then the hypothesis is rejected at the 95% (5% significance) level and the alternative hypothesis is accepted.

Conclusion

We see that the study groups (using the Prentice Hall programs) tested significantly higher than the control groups (not enrolled in the Prentice Hall program).

Technical Post Script

For the purpose of completeness, we include here a brief discussion of the statistical tests that were used. All of the statistical tests are in the classical statistical domain, and are broadly used across all disciplines, including psychometrics. The programs that were used in the computational phase of this project were SPSS versions 6/9.

T-test: The mean of one population is equal to the mean of another population when the variance is unknown. The hypothesis is that the means are equal; and the alternative hypothesis is that they are unequal for which we use a two-sided test. The means and the variance are calculated and the t-statistic is computed. The significance of the t-statistic then computed. If the significance of the t-statistic is less than .05 the hypothesis is rejected; otherwise, the alternative hypothesis is accepted.

ANCOVA: The dependent variable (in this case the post-test score) is equal for both populations (study and control) after adjustment for the covariate (in this case the pre-test score). A number of assumptions are invoked in the ANCOVA: population variances about the regression lines; the regression curve is a straight line; and the pre- and post-test values are approximately normally distributed. The F-value is the statistic that is computed from a series of computations of sums of squares and sums of products. The distribution of this F-value is known and the significance may be computed. If the significance of the F-value is less than .05, the hypothesis for equal means of the post-test is rejected; otherwise, it is accepted.

Normal Curve Equivalent - Comparison of Scores across Tests: The normal curve equivalent (NCE) scale, ranging from 1 to 99, coincides with the national percentile scale at 1, 50, and 99. NCEs have many of the same characteristics as percentile ranks, but have the additional advantage of being based on an equal-interval scale. The difference between two successive scores on the scale has the same meaning throughout the scale. This property allows you to make meaningful comparisons among different achievement test batteries and among different tests within the same battery. You can compare NCEs obtained by different groups of students on the same test or test battery by averaging the test scores for the groups.”

Quoted directly from: Teacher’s Guide to TerraNova®, page 138, CTB/McGraw-Hill 1999.

All tests were scored by CTB/McGraw-Hill, the publisher of TerraNova®. Statistical analyses and conclusions were performed by an independent firm, Pulse Analytics Inc., Ridgewood, New Jersey.

1998 Course 2 Middle Grades Mathematics Program Efficacy Study

Abstract

This study investigated the effects of seventh grade mathematics programs. Teachers participating in the study taught one control group (current text) and one study group (Prentice Hall text—*Middle Grades Math* ©1999). Student achievement levels prior to the study were compared using prior year standardized test results. At the end of the treatment period, students were tested with the same chapter test. Overall, the Prentice Hall pilot classes performed significantly better than the control classes.

Methodology

This study of *Middle Grades Math, ©1999 (MGM3e)* was conducted during the months of April and May of 1998. A total of 14 seventh grade math teachers, representing three states; Tennessee, Virginia, and West Virginia tested one chapter of the *MGM3e*, Course 2. Teachers selected the chapter that corresponded with their regular curriculum. The following six chapters were chosen:

- Chapter 1: Interpreting Data and Statistics
- Chapter 3: Algebra: Integers and Equations
- Chapter 6: Using Proportions and Percents
- Chapter 7: Investigating Geometry
- Chapter 8: Geometry and Measurement
- Chapter 9: Using Probability

Teachers selected two classes of equal ability—a study group, which used *MGM3e*, and a control class, which used the current textbook. Teachers provided the previous year's standardized test scores (means) for both classes. The standardized tests served as a pretest to determine whether study and control classes were comparable prior to program implementation.

Teachers taught the study class from the *MGM3e*, and continued teaching the control class with their current textbook. Approximately 600 students among 28 classes (both study and control) were part of the field test; on average, the class size was 26, with a range of 17 to 30 students. For any given teacher, the study took between 2 and 4 weeks for completion, with an average of 2.9 weeks. Time-on-task by chapter is presented below:

Chapters	No. Teachers Using	Week Range	Average No. Weeks
Chapter 1: Interpreting Data and Statistics	1	3	3
Chapter 3: Algebra: Integers and Equations	2	2	2
Chapter 6: Using Proportions and Percents	3	3 to 4	3.25
Chapter 7: Investigating Geometry	2	3	3
Chapter 8: Geometry and Measurement	4	2 to 4	3
Chapter 9: Using Probability	3	2.5 to 3.5	2.8
Total/Average	14*	N/A	2.9

***Note:** one teacher taught from two chapters: seven and eight.

At the end of the study, teachers administered the same chapter test to both study and control classes and reported the scores.

Analysis of Results

STANDARDIZED TEST RESULTS

Table 1 presents standardized test results for the study classes and control classes. Each teacher had one study and one control class. Teachers obtained the prior year's (1997, sixth grade) standardized test scores on both their study class and control class. Only mean math scores for each type class are reported.

- As seen, both the study and control classes were comparable prior to the *MGM3e* field study in April/May of 1998; that is, there are no significant differences among test results for study and control classes.

Table 1.
Summary of Standardized Test Results

Test Type	Score Type	No. of Students		Average Math Scores	
		Study	Control	Study	Control
T-CAP	percentile	22	23	.72	.78
SAT-9	percentile	23	23	.47	.56
SAT-9	scale	26	21	679	668
VLP	scale	29	28	265	263
VLP	scale	24	24	267	266
VLP	scale	25	22	255	257

Test Key: T-CAP= Tennessee Comprehensive Assessment Program
SAT-9=Stanford Achievement Test (9th edition)
VLP= Virginia Literacy Passport

CHAPTER TEST RESULTS

Table 2 presents chapter test results. The chapter tests, six (6) in total, were administered only as post tests, to both study and control classes. Test results were analyzed for 616 students, among 28 classes: 14 study and 14 control. Note: one teacher taught and tested two chapters: seven (7) and eight (8). Test results are reported as average percent correct.

- As seen, overall, the study classes (.70) did significantly better than the control classes (.61). This is also true for five of the six chapters (all but chapter 3).

Table 2.
Summary of Chapter Test Results
By Chapter

Chapter Tested	No. Students		Average Percent Correct	
	Study N=331	Control N=327	Study %	Control %
Chapter 1: Interpreting Data & Statistics	22	23	86**	55
Chapter 3: Integers & Equations	43	41	65	67
Chapter 6: Proportions & Percents	64	57	63*	58
Chapter 7: Investigating Geometry	38	39	75**	62
Chapter 8: Geometry & Measurement	101	102	74*	69
Chapter 9: Using Probability	63	65	70**	61
Total Student Average (Weighted)			70**	61

**significant at the 99% confidence level

*significant at the 90% confidence level

Note: as indicated above, one teacher taught both chapters 7 and 8, so the total base (658) is greater than total number of students participating (616).