RESEARCH REPORT

A Comparison Group Study of the Effects of *Premier Science*, a Standards-based Middle School Product, on Student Achievement

Denis Newman
Empirical Education Inc.

Marco Muñoz
Jefferson County School District

Andrew Jaciw
Stanford University

April 27, 2004

Empirical Education Inc.
www.empiricaleducation.com
425 Sherman Avenue, Suite 210
Palo Alto, CA 94306
(650) 328-1734

Acknowledgements

Acknowledgement: This research was sponsored by Frey Scientific (a division of School Specialty) and Empirical Education Inc. in cooperation with the Jefferson County School District (Louisville, KY).

About Empirical Education Inc.

Empirical Education Inc. was founded to help K–12 school districts, publishers, and the educational R&D community assess new or proposed instructional programs through scientifically based pilot implementations. The company draws on the expertise of world-class researchers and methodologists assuring that the research is objective and takes advantage of current best practice in rigorous experimental design and statistical analysis. The company's findings let educators quantify the value of programs and help them partner with providers to implement those most effective for their students.

Reference this report: Newman, D., Muñoz, M., & Jaciw, A. (2004, April). A comparison group study of the effects of Premier Science, a standards-based middle school product, on student achievement. Palo Alto, CA: Empirical Education Inc. Retrieved 2020 from https://www.empiricaleducation.com/past_research/

Executive Summary

Two issues guided our research for Jefferson County School District: their interest in the impact of the various science adoption choices made by middle schools, and the large gap between African American students and White students in science achievement. Our study, conducted during the 2003-2004 school year with 1,289 students in nine schools, was designed to compare *Premier Science*, a new product published by Frey Scientific, to other products used in the district in raising test scores and in helping to close the achievement gap between student groups.

We found that students at the lower end of the reading scale scored higher on the state science test, the Kentucky Core Content Test (KCCT), if they were in classes taught by teachers using *Premier Science* than in classes taught by teachers who used more traditional methods of instruction. These findings allow the district to conclude that continued use of *Premier Science* is warranted as a path toward their strategic goals of raising science achievement while closing the achievement gap.

Intervention. *Premier Science* features explicit correlations with the science curriculum standards published by the American Association for the Advancement of Science and the National Research Council. It also promises high correlation with the Kentucky science standards in its emphasis on inquiry activities and conceptual understandings. *Premier Science* consists of 12 curriculum units designed for middle school grades covering the life, physical, and earth sciences. Each unit contains a series of inquiries supported by kits of student materials, multimedia presentations, and an extensive teacher guide. Units follow the "Five Es" structure with components designed to Engage, Explore, Explain, Extend, and Evaluate, a format based on the learning cycles common to many inquiry programs.

Research design. Because the schools had already chosen their science adoptions, we used a quasi-experimental research design to compare the performance of 7th grade science students attending schools using *Premier Science* to the performance of those at similar schools that used other products. We focused exclusively on 7th grade students because the KCCT was given only at that grade in middle school. We had information on student clustering in classrooms, but not on classroom clusters for teachers; nor did we have resources for observations of classrooms, interviews of teachers, or analysis of classroom-by-classroom differences in implementation of the product. Nonetheless, because we had test score and demographic data for individual students, we were able to address the question of whether the product had differential effects on student groups.

Participants. Implementation assessment ratings provided by the district's science specialist allowed us to narrow the group of *Premier Science* users to three schools. We formed the comparison group by identifying six district schools that closely matched the demographics and test scores of the user schools. The formation of two groups—*Premier Science* and comparison—allowed a comparison of student performance but did not meaningfully control for many factors that make the two groups of schools different. The schools' previous self-selection into one group or the other systematically confounded teacher and/or school leadership preference for a kind of teaching and the choice of a compatible instructional product. Teachers may have differed in their own science education, their teaching skills or orientation, and even their enthusiasm for science—all of which provide alternative explanations for any differences we found.

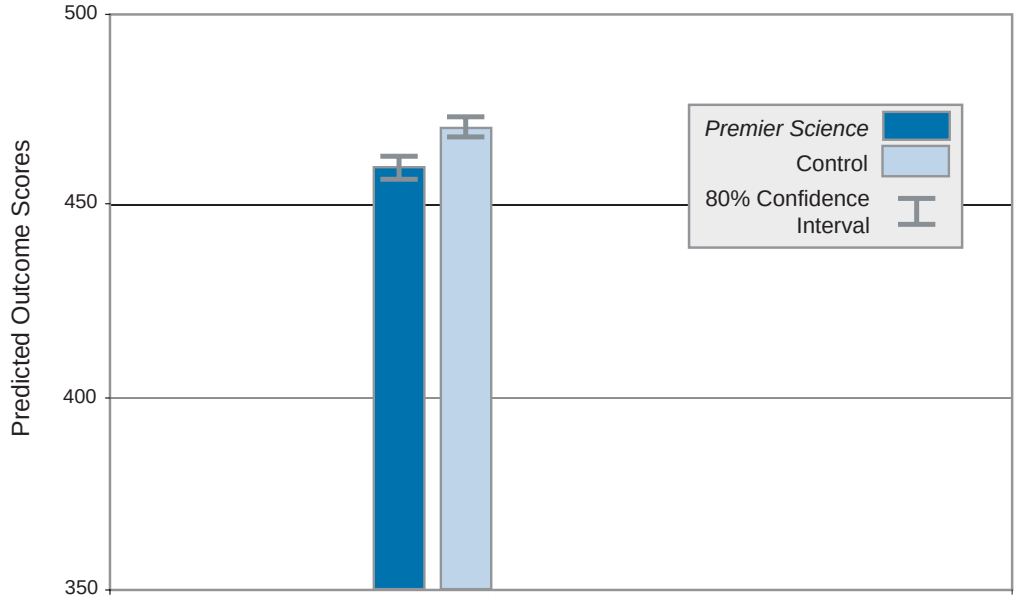
Statistical Analysis. We recognize that statistical analyses cannot make up for constraints in the original research design, especially for the potential bias that arises from not performing a randomized experiment. We utilized as much of the given information as possible to provide useful findings. We used analyses of covariance to control for the effects of possible confounders and to increase the precision of

the estimated treatment effect. We also modeled random effects to account for intra-class correlation at the class level. (Because teachers were not identified, the intra-class correlation due to clustering at the teacher level could not be controlled for.)

Application of these methods yielded two sets of data, one for users of *Premier Science* and one for the comparison group. Each set included the student's school attended and class period for science, socio-economic status (based upon Free/Reduced-price Lunch program participation), sex, race, score on the Stanford Diagnostic Reading Test (SDRT) taken at the beginning of the 2002-2003 school year, and score on the state science test (KCCT) given to all 7th graders. The KCCT science test was the outcome measure or dependent variable. For the SDRT and KCCT we used the "scale score" wherein raw data are converted to a scale that makes them easier to compare to other measures. From these data, we developed a statistical model to identify the set of variables that account for most of the variance in the outcome, selecting those that were important theoretically or related to critical policy decisions.

Results. The statistical modeling identified the covariates that accounted for most of the variation in the outcome and gave some control over selection bias that could arise due to the covariates being distributed differently in the two conditions. As noted above, student scores on the KCCT science test constituted the outcome variable of interest. The strongest influence on, or predictor of, science achievement was found to be the initial SDRT score. We also examined several other variables such as race, socio-economic status, and sex in alternative models but found them to be unconfounded with treatment. Their presence did not affect the estimate of the treatment impact or the interaction of interest – between the prior score and treatment – so they were excluded from further analyses.

Our analyses revealed that students in the *Premier Science* group had a small advantage in science achievement overall. They also show that this effect was substantially stronger for students at the lower end of reading ability. The bar graph represents the impact of *Premier Science* for the median student scoring in the bottom quartile of the SDRT pretest. The bar graph includes the 80% confidence interval as a marker at the top of the bars. Because the markers do not overlap, we have reasonable confidence that *Premier Science* would make a difference for this student.



Median Student of the Bottom Quartile of the Pretest

The difference in science test scores for the median student in the bottom quartile amounts to 10.6 points. This is a small difference, but one that is unlikely to have occurred by chance.

Since an initial concern of the school district was the gap in science achievement between African American students and other students, we also addressed this issue. Overall, the African American students are represented more heavily in the lower quartiles of reading ability. As noted above, our analysis showed that the reading pretest score was the best predictor of science achievement. Thus the apparent advantage that *Premier Science* gave African American students can be explained as an artifact of differences in reading ability.

Conclusion. We found a small but significant impact from being in the *Premier Science* classes for the students at the lower end of the reading scale. For the students at the higher end, there was no difference between *Premier Science* and comparison groups.

We can speculate that the textbooks used in the comparison condition accompanied a more text-based mode of instruction with greater emphasis on reading. *Premier Science*, on the other hand, is largely activity-based, thus encouraging learning through exploration and discussion. For students who are not good readers, such an approach may be more engaging. These findings are encouraging for advocates of hands-on inquiry in middle school science. Without specific observations or reports on classroom activities and approaches, however, we cannot be sure what kind of instructional practices actually occurred in classrooms that resulted in improved achievement on the state science test for students using *Premier Science*. Nonetheless, our study provides positive evidence for the use of *Premier Science* as an alternative to traditional science education products for 7th graders with low reading ability in this district.

A Comparison Group Study of the Effects of *Premier Science*, a Standards-based Middle School Product, on Student Achievement

Table of Contents

Introduction

Jefferson County School District conducted its middle school science adoption during the 2001-2002 school year, with each of the 24 middle schools making its own decision. While most of the schools chose to adopt a textbook from one of the major publishers, six schools chose a product that was new on the market and promised to be highly correlated with the Kentucky science standards in its emphasis on inquiry activities and conceptual understandings. The mixture of adoptions within the same district gave us an opportunity to compare the science achievement in schools that used *Premier Science* with similar schools that adopted a different science product. This is a report of our findings, based on the statewide science assessment conducted in the spring of 2003, after almost a year of using the new science materials. While the research lacks the controls of an experiment, the comparison may be useful for district decision-makers.

The district was interested in the impact of the various science adoption choices made by the middle schools. It was also concerned with the large gap between the African American students and White students in science achievement. In its March 2003 Comprehensive Improvement Plan, the district drew attention to this gap, evident in the Commonwealth Accountability Testing System (CATS) assessment, which tests science achievement in 4th, 7th, and 11th grade. For 7th grade (the assessment level examined in our study), the percentage of White students scoring at a novice level was 26% compared to 61% for African American students. The improvement plan stipulates that "… by 2004, the achievement gap in science will be reduced by decreasing the % difference between white and African American novice students and by decreasing the overall % of novice students." Thus it was important to find out both whether *Premier Science* was effective compared to other solutions in raising test scores and also to learn whether it helped to close the achievement gap between student groups.

Premier Science

Premier Science, published by Frey Scientific, consists of 12 curriculum units designed for 6th through 8th grade covering the life, physical, and earth sciences. Each unit, which can be purchased separately, contains a series of inquiries supported by kits of student materials, multimedia presentations, and an extensive teacher guide. Units follow the "Five Es" structure with components designed to Engage, Explore, Explain, Extend, and Evaluate, a format with its origins in the learning cycles common to many inquiry programs (Bybee, 1997).

The developers of *Premier Science* provide explicit correlations between their materials and the curriculum standards for school science published by the American Association for the Advancement of Science (1994) and the National Research Council (1995). These documents provide a wealth of information, both on the content that should be covered in middle school, and about the scientific processes students should understand and engage in. Our earlier report reviews the scientific evidence underlying these approaches to inquiry science (Newman, 2003).

Evidence of the Effectiveness of *Premier Science*

The No Child Left Behind Act directs school districts to base their adoption decisions on scientific research about the effectiveness of an instructional program or of the approach it incorporates. The US Department of Education (2003) recommends that districts look for highly rigorous research that provides strong evidence of effectiveness. To meet this "strong evidence" level, the research should use well-designed randomized experiments in which, for example, classrooms in a district are randomly assigned to use the new product or to continue using their existing materials and methods. The randomization

assures that no characteristics of the classes or teachers (such as their motivation to use a new product, demographics, etc.) are represented more in one group than in the other. Short of such an experiment, educators can look for "possible evidence" from studies, often called a quasi-experiments, that use comparison groups in which the effect of randomization is approximated by carefully matching students and teachers that are using the new product with students and teachers that are not using it (Shadish, Cook, and Campbell, 2002). Where the product being adopted has not been subject to a specific experimental test, educators are encouraged to look for evidence that products using a similar approach have been shown to be effective.

The approach taken by *Premier Science* has a long history, going back to curricula that were newly developed during the Sputnik era of the 60s and 70s. During that period of development and testing new science materials, hundreds of research studies were conducted to examine aspects of these curricula and to measure their effectiveness. In a paper titled "The effects of new science curricula on student performance" (Shymansky, Kyle, and Alport,1983), researchers conducted a "meta-analysis" as a way to summarize the effects found across all relevant studies. A meta-analysis is a research technique that combines the statistical results of experimental or other quantitative comparison group studies to determine an overall effect. The researchers thoroughly review all the published (and, where available, unpublished) reports to determine a set of studies that used appropriate research design and provided sufficient detail to include. They found, based on 105 studies, that the then-new science curricula of the 60s and 70s were more effective than the current traditional textbook programs of the 80s across a number of different measures including quantifiable achievement, student perceptions, process skills, problem solving, and other related skills such as performance on tasks modeled after Piagetian interviews. Seven years later, after some major improvements in meta-analysis procedures (Hedges and Olkin, 1985), researchers reanalyzed the set of studies using more conservative techniques and criteria (Shymansky, Hedges, and Woodworth, 1990). Their reanalysis continued to show advantages in achievement and student attitudes, as well as in some—but not all—of the related measures.

Thus *Premier Science* passes the first test of being based on approaches that have been shown to work in the past. However, as a new product, it has not been subjected directly to the kind of rigorous effectiveness study that the US Department of Education recommends. The Department's recent guidance to school district decision-makers suggests that they undertake comparison group studies when possible after implementing a product to at least gain a "sense of whether the product is having effects that are markedly different from what the evidence predicts." (US Department of Education, 2003). Combined with the product's explicit alignment with the Kentucky state standards, research from the latter part of the 20th century provides possible evidence of effectiveness for products like *Premier Science*.

Our study of the outcomes from the first year of implementation was an attempt to test the hypothesis that follows from the 1990 study by Shymansky et al. suggesting that the product would outperform traditional textbooks. In addition, it was important for the district to know whether the product had differential effects for White and African American students.

Method

Our research took the form of a comparison group study. The idea was to compare the performance of 7th grade science students attending schools where *Premier Science* was used to the performance of those at schools with similar demographics that used other products. Since the schools had already chosen

which product to adopt, we did not have an opportunity to control which classes used the new product and which did not. Also, we did not have resources for observations of classrooms, interviews of teachers, or analysis of classroom-by-classroom differences in implementation of the product. We had information on student clustering in classrooms but not on the classroom clusters for teachers (7th grade teachers often taught more than one class period of science). Nonetheless, because we had test score and demographic data for individual students, we were able to address the question of whether the product had differential effects on differing categories of students.

Composing the Comparison Groups

The district's science specialist provided the names of schools that had selected *Premier Science* for adoption. The district's Planning and Program Evaluation department compiled the data on the students from those schools, as well as from six schools that matched those in terms of their demographics and previous test scores. We focused exclusively on the 7th grade students because the state science test, the Kentucky Core Content Test (KCCT), was given in middle school only at that grade.

The science specialist also provided a general assessment of the level of implementation of the product in the 7th grade in each of the six schools. This assessment was based on a five-point scale applied to reports from each school's science coordinator. Each school's 7th grade science product was rated on the following scale:

0 Implementation not evident or visible; poor

1 Some implementation

2 Good implementation

3 Very good implementation

4 Exemplary implementation

Since we were interested in the impact of the materials when used as intended, these implementation assessment ratings allowed us to narrow our focus in defining the group to be considered users. We eliminated two schools; one had ordered only a small subset of the *Premier Science* materials and the other did not order the product in time for use in the current year. At a third school, only one of the teachers was using *Premier Science*, and she was coded at level 1. Therefore the three remaining schools constituted our group of school users.

We formed the comparison group by identifying six schools that closely matched the demographics and test scores of the user schools. The matching procedure was at the school level and started with poverty level (percentage of students in the Free/Reduced-price Lunch program) for the schools. We refined the match through similarity on single-parent households, race, gender, and scores on the Stanford Diagnostic Reading Test (SDRT). The SDRT (Kramer, Conoley, and Murphy, 1992) is a standardized diagnostic test that the district administered to the 7th grade students during the first few weeks of the school year. Except for the SDRT, all data used for matching were from the 2001-2002 school year. Students from the two sets of schools formed the *Premier Science* and comparison groups. Students in special education were excluded from our analysis, as were 12 students, equally divided between *Premier Science* and comparison groups, who were missing scores on the SDRT. Table 1 shows the characteristics of the two groups.

Table 1: How the *Premier Science* and comparison groups broke down in terms of important variables

	Number of Students	Number of Classes	Free/Reduced Lunch Percent	African American Percent	SD Reading		KY Science	
					Mean	SD	Mean	SD
Premier Science	625	28	61.0	34.6	652	36.2	492	28.3
Comparison	664	35	64.4	44.7	659	45.8	490	37.90

For each student, we knew the class period in which he or she took science and, therefore, the class clusters. We did not know which classes were taught by the same teacher. We estimate that approximately 16 teachers were responsible for teaching these classes.

The formation of two groups—*Premier Science* and comparison—allows a comparison of student performance but does not meaningfully control for many factors that make the two groups of schools different. Because the schools initially self-selected into one group or the other, there is a systematic confounding of teacher and/or school leadership preference for a kind of teaching and the choice of a compatible instructional product. Therefore, aside from using the product, the teachers may have differed in their own science education, their teaching skills or orientation, and even their enthusiasm for science—all of which provide alternative explanations for any differences we found.

Statistical Analysis

We recognize that statistical analyses cannot make up for constraints in the original research design. In our analyses we utilized as much of the given information as possible to provide useful findings. We used analyses of covariance (ANCOVA) to control for the effects of possible confounders and to increase the precision of the estimated treatment effect. We also modeled random effects to account for intra-class correlation at the class level. However, because teachers were not identified, the intra-class correlation due to clustering at the teacher level could not be controlled for.

Application of these methods yielded two sets of data, one for users of *Premier Science* and one for the comparison group. Each set consisted of student records including the student's socio-economic status (based upon Free/Reduced-price Lunch program participation), sex, race, score on the SDRT taken at the beginning of the 2002-2003 school year, and score on the state science test (KCCT) given to all 7th graders. The KCCT science test was the outcome measure or dependent variable. For the SDRT and KCCT we used the "scale score," meaning that the raw data were converted to a scale that makes the results easier to compare to other measures. For each student we also knew which school they attended and their class period for science.

We tested several statistical models. The models represent theories concerning which variables impact the outcome. The intention is to try to include in the model as many variables as possible that are correlated with treatment and that affect the outcome. Such variables are called confounders. Inclusion of confounders in the model effectively allows us to look at the impact of the treatment variable holding the effects of the confounders constant. Ideally we would include all of the critical confounders in the model thereby yielding an more accurate estimate of the treatment impact. Normally it is unrealistic to expect to collect information on, and control for, all confounders. Unfortunately, not controlling for the impact of important confounders limits the interpretability of the estimated treatment effects.

Results

The statistical modeling identified the covariates that accounted for most of the variation in the outcome and gave an indication of the strength of their impact. The student scores on the KCCT science test was the outcome or "dependent" variable of interest. The strongest influence on, or predictor of, science achievement was the student score on the Stanford Diagnostic Reading Test. We examined several other variables such as race, socio-economic status, and sex in alternative models but found they accounted for a very limited amount of variance, and did not affect the estimate of the treatment impact once the effect of the SDRT was factored in.

Table 2: Multi-level mixed model for science test outcome results controlling for reading pretest

Descriptive statistics: science outcomes	Raw Group Means	Standard Deviation	Number of Students	Number of Classes	
Premier Science	492.245	28.289	625	28	
Comparison	490.423	37.851	664	35	
Mixed model: Fixed factors related to science outcomes	**Estimate of Coefficient**	**Standard Error**	**DF**	***t* value**	***p* value**
Intercept	488.10	2.969	4	164.37	<.0001
Pretest score (centered at the mean)	0.606	0.021	1211	28.96	<.0001
Condition (*Premier Science* = 1; comparison = 0)	5.680	4.229	1211	1.34	0.178
Condition by pretest interaction	-0.106	0.032	1211	-3.26	<.0001
Mixed model components	**Estimate of Variance Component**	**Standard Error**		***z* value**	***p* value**
Class mean achievement	49.211	13.670		3.60	0.0002
Within class variation	435.42	17.687		24.62	<.0001

We found that students in the *Premier Science* group had a small advantage in science achievement overall. In Table 2, the line for condition shows a 5.68-point advantage for the average student if placed in the *Premier Science* rather than the comparison group. The p value of .178 indicates a 17.8% chance that an impact with an absolute value this large (or larger) may have occurred simply by chance. Even more interestingly, our statistical model showed that this effect was substantially stronger for students at the lower end of reading ability. This is shown in the very low p value for the condition by pretest interaction. One way to display the interaction between the condition and the students' reading ability is shown in the scatterplot in Figure 1.

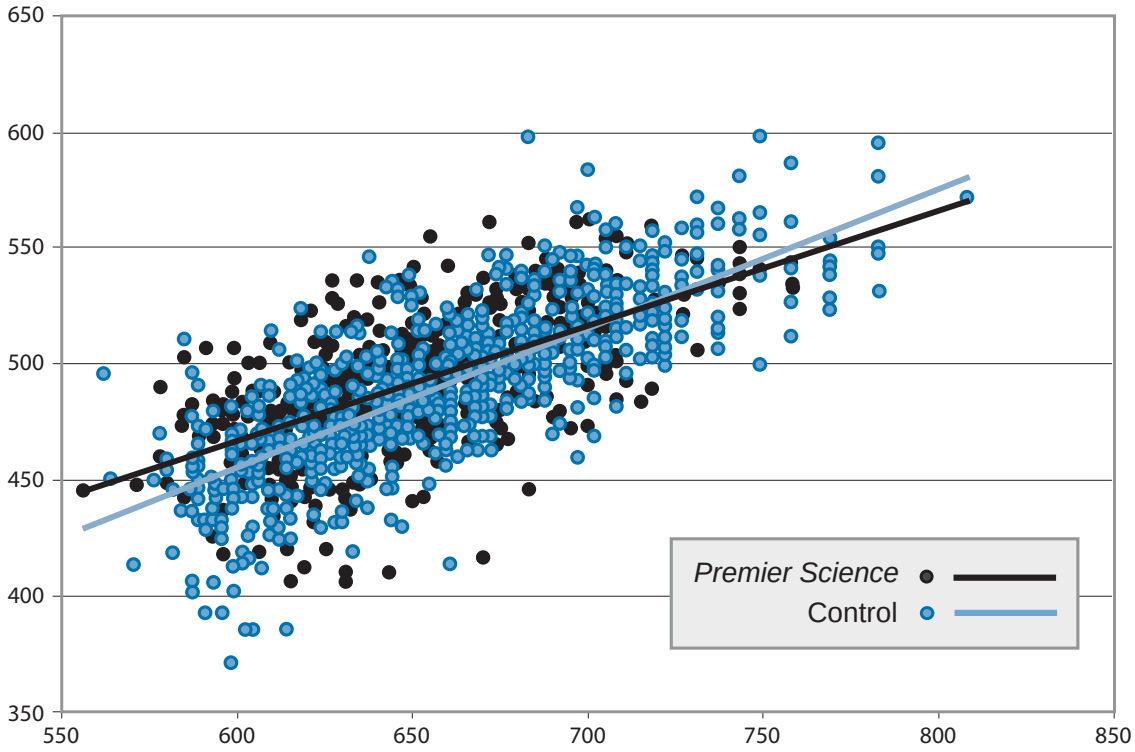


Figure 1: Scatterplot with interaction between condition and prior score in reading

Each student is plotted as either a dark (*Premier Science*) or light (comparison) dot. The position of the dot depends on the student's score on the reading test (x-axis) and the science outcome measure (y-axis). Although inspection of the scatter does not reveal an obvious pattern, our statistical model provides a more precise representation of how the *Premier Science* and comparison groups mapped onto the relationship between the earlier reading test and later science test. In general the line for the *Premier Science* group is higher than the comparison group, indicating an advantage for the students exposed to *Premier Science*. The fact that the two lines are not parallel represents the interaction between the condition and the student's prior score. Students at the lower end of the reading scale scored higher if they were in classes taught by teachers using *Premier Science* than in classes taught by teachers who used more traditional methods of instruction.

The next questions we addressed were the magnitude of this difference and whether it could have occurred by chance. Figure 2 graphs the difference between the heights of the two lines measured in units of the science test score. The difference is greater at the lower end of the reading scale and diminishes to less than zero toward the right of the graph. We also indicate the prediction for the median student for each quartile of the reading test. The shaded bands represent how likely the difference indicated by the black line could have happened just by chance. These confidence intervals are an alternative way of expressing what is often called statistical significance. The band with the darkest shading surrounding the black line is the "50-50" area, where the difference is considered equally likely to lie within the band as not. As we move out to the lighter bands, the likelihood increases that the true difference exists within the bands. The outer band represents conventional significance for which there is only a 5% chance that the true value of the difference lies outside the band. We can be quite confident that, at least for the students in the lower part of the reading scale, there was a measurable difference.

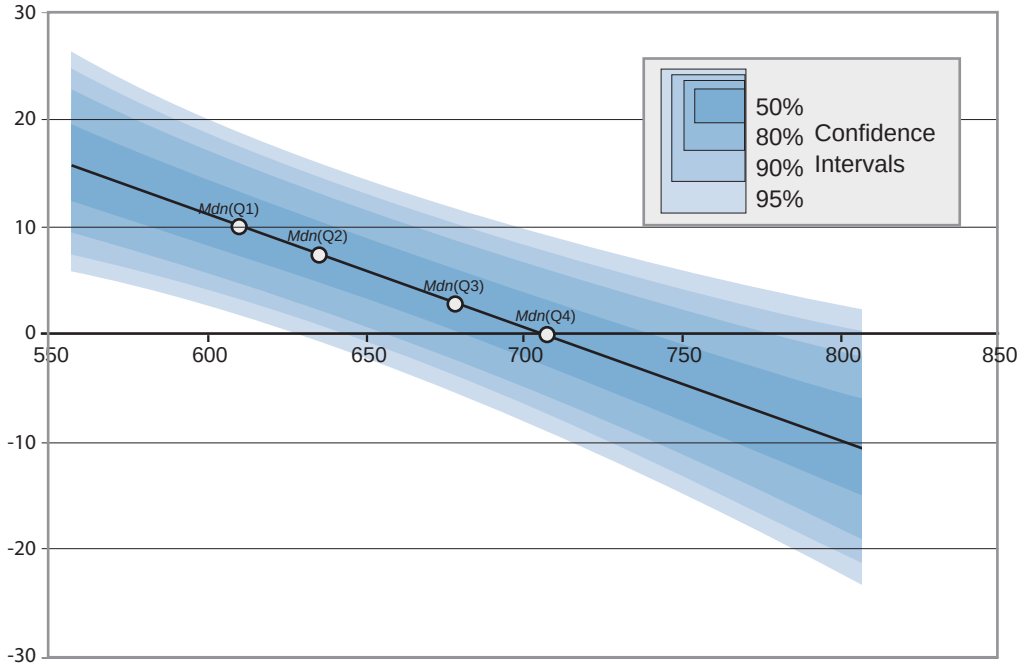


Figure 2: Science outcome—difference between *Premier Science* and comparison with values for the median student at each quartile of the pretest

Figure 3 shows some of the same information as presented in Figure 2 but in bar graph form. The bar graph represents the impact of *Premier Science* for the median student in the bottom quartile of the pretest. The bar graph includes the 80% confidence interval as a marker at the top of the bars. This marker is an alternative representation of the 80% band in Figure 2. Because the markers do not overlap, we have reasonable confidence that *Premier Science* would make a difference for this student.

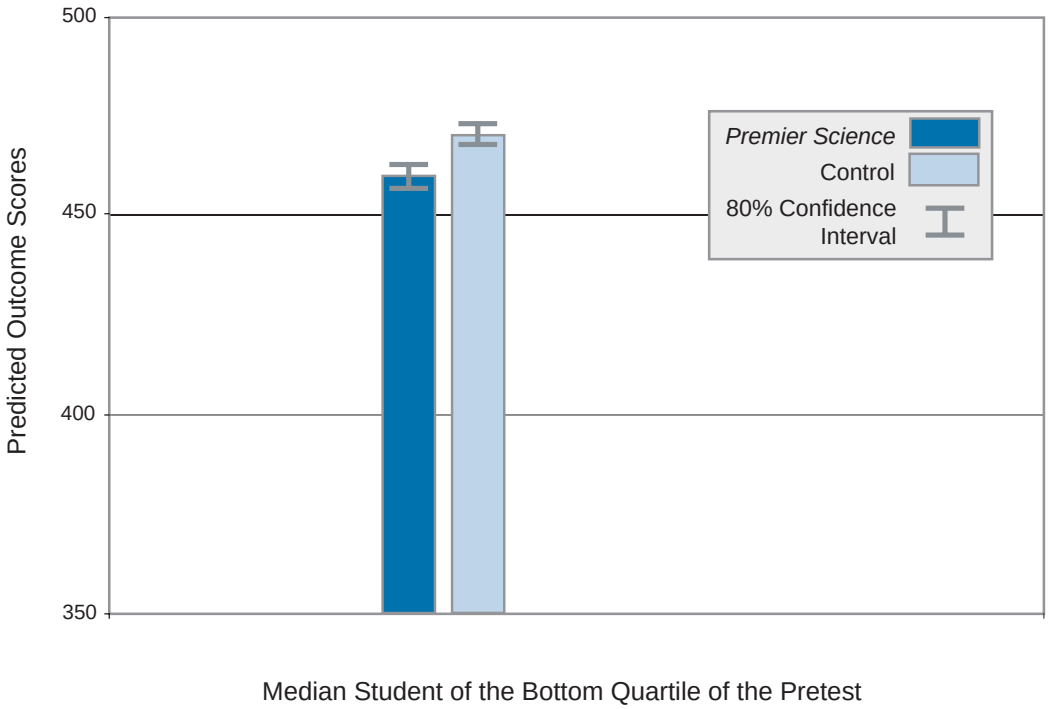


Figure 3: Science outcome—difference between *Premier Science* and comparison groups for the median student in the bottom quartile

The difference in science test scores for the median student in the bottom quartile amounts to 10.6 points. Because this test is given only in 7th grade and does not provide growth expectations, it is not possible to relate this difference to, for example, a grade equivalent scale. The prediction for this student, if in the comparison group, is about 460 points on the science test, placing him or her within the upper range of the "novice" category, which extends from 325 (bottom of the scale) to 518. The average advantage gain from being in the *Premier Science* group for this student still leaves him or her in the novice category. We can get an idea of the size of the gain relative to the overall spread in scores. For the spread, we use the pooled standard deviations of posttest performance (33.56). Dividing the difference (10.6) by the pooled standard deviations gives .32.

Since an initial concern of the school district was the gap in science achievement between the African American students and other students, we also addressed this issue. As previously noted, we did not use the racial categories in our statistical model because including them did not provide a better theory of the data once the differences in reading score were included. Overall, the African American students are represented more heavily in the lower quartiles of reading ability, as can be seen in Table 2.

Table 3: Distribution of African American and non-African American students in the quartiles of reading scores

Number of Students	First Quartile	Second Quartile	Third Quartile	Fourth Quartile	Total
Non African American	148	160	201	267	776
African American	175	162	121	55	513

Our analysis showed that the reading score was the best predictor of science achievement. Thus the apparent advantage that *Premier Science* gave African American students can be explained as an artifact of differences in reading ability.

Discussion

We found a small but significant impact from being in the *Premier Science* classes for the students at the lower end of the reading scale. For the students at the higher end, there was no difference between the *Premier Science* and comparison groups.

We can speculate that the textbooks used in the comparison condition accompanied a more text-based mode of instruction and greater emphasis on reading. *Premier Science*, on the other hand, is largely activity-based, thus encouraging learning through exploration and discussion. For students who are not good readers, such an approach may be more engaging. Without specific observations or reports on classroom activities and approaches, we cannot be sure what kind of instructional practices actually occurred in classrooms that resulted in improved achievement on the state science test for students using *Premier Science*. Still, the findings are encouraging for advocates of hands-on inquiry in middle school science.

A weakness that is inherent in any comparison group study, where we form the groups by matching users of an intervention with similar non-users, is that we have no control over many factors that could be important other than the use or non-use of the intervention. For example, teachers in schools that

chose to use *Premier Science* may be more inclined to use inquiry materials and more confident in their abilities as science teachers. These teachers might have outperformed the other schools even without using *Premier Science* materials. This is one of the main reasons that researchers prefer to use random assignment rather than matching (Cook, 2002). With random assignment we can be sure that these important characteristics are distributed between the users and non-users of the product.

Nonetheless, our results allow the school district to conclude that continued use of *Premier Science* is warranted as a path toward their strategic goals of raising science achievement while closing achievement gaps.

References

American Association for the Advancement of Science (1994) *Benchmarks for Science Literacy*. Oxford: Oxford University Press.

Bybee, R (1997) Achieving scientific literacy: From purposes to practices Boston: Heinemann

Cohen, J. (1988) Statistical power analysis for the behavioral sciences. (2nd ed.) Hillsdale, NJ: Erlbaum.

Cook, T.D. (2002) Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, 24 (3) 175-199.

Hedges, L.V. & Olkin, I (1985) *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Kramer, J. J., Conoley, J. C., & Murphy, L. L. (1992). *The eleventh mental measurements yearbook*. Lincoln, NE: University of Nebraska.

National Research Council (1995) *National Science Education Standards*, Washing, DC: National Academy Press.

Newman, D. (2003) *The Scientific Research Base for* Premier Science. (Empirical Education Reports) Palo Alto, CA: Empirical Education Inc.

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

Shymansky, J.A., Kyle, W.C., & Alport, J.M. (1983) *The effects of new science curricula on student performance*. Journal of Research in Science Teaching, 20, 387-404.

Shymansky, J.A. Hedges, L.V. & Woodworth, G. (1990) *A reassessment of the effects of inquiry-based science curricula of the 60s on student performance*. Journal of Research in Science Teaching, 27 (2) 127-144.

US Department of Education (2003) *Identifying and implementing educational Practices Supported by Rigorous Evidence: A User Friendly Guide*. Washington DC: Institute of Education Sciences. (Available at: http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf)