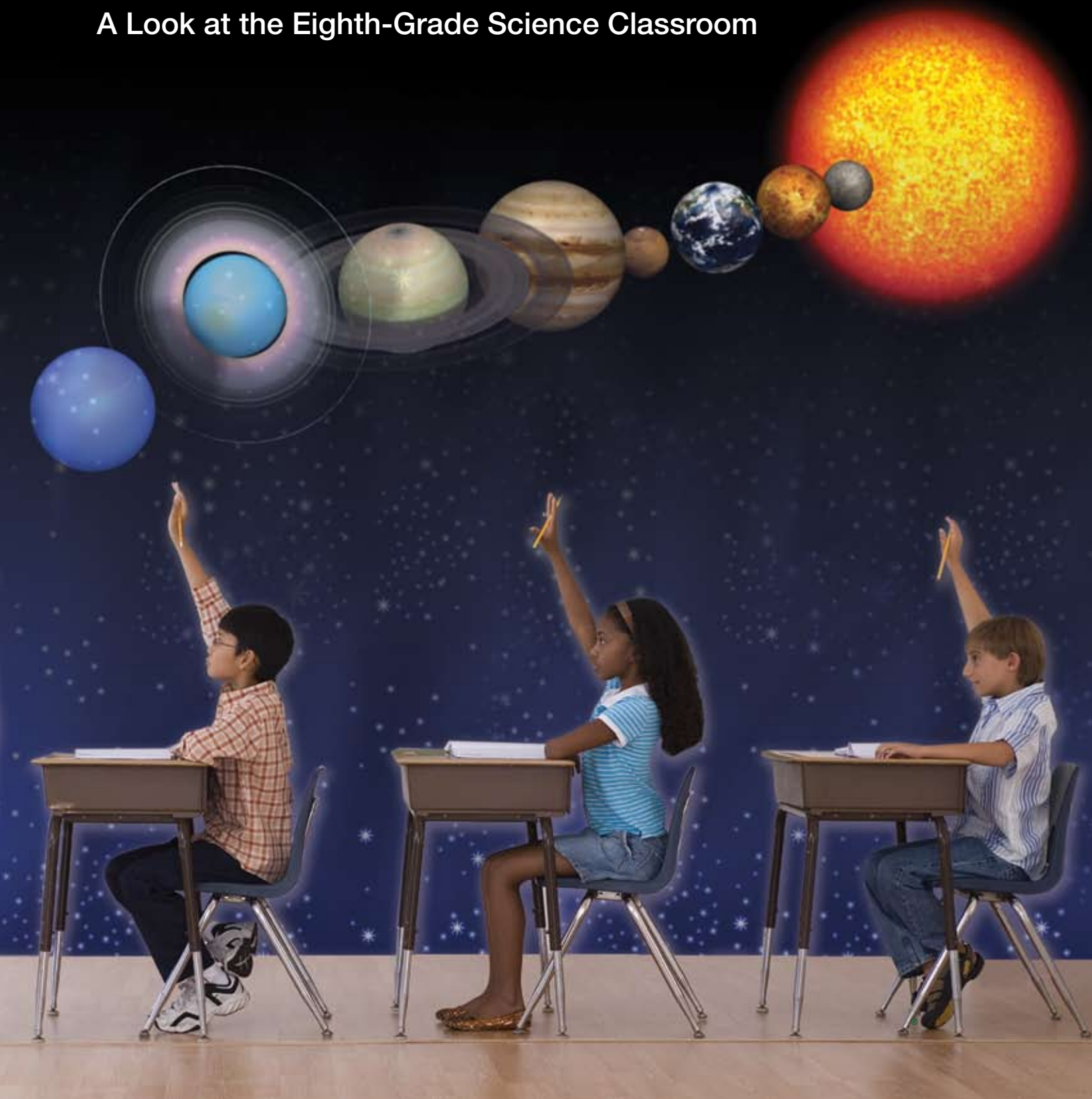
ETS®

Listening. Learning. Leading.®

Policy Information Report

Exploring What Works in Science Instruction:

A Look at the Eighth-Grade Science Classroom



This report was written by:

Henry Braun
Boston College

Richard Coley
Yue Jia
Catherine Trapani
Educational Testing Service

Additional copies of this report can be ordered for $15 (prepaid) from:

Policy Information Center
Mail Stop 19-R
Educational Testing Service
Rosedale Road
Princeton, NJ 08541-0001
609-734-5212
pic@ets.org

Copies can be downloaded free from: *www.ets.org/research/pic*

May 2009
Policy Evaluation and
Research Center
Policy Information Center
Educational Testing Service

Table of Contents

Preface

In a recent opinion piece in *The New York Times*, Brian Green, professor of physics at Columbia University and author of *The Elegant Universe* and *The Fabric of the Cosmos*, writes:

> *Science is the greatest of all adventure stories, one that's been unfolding for thousands of years as we have sought to understand ourselves and our surroundings. Science needs to be taught to the young and communicated to the mature in a manner that captures this drama. We must embark on a cultural shift that places science in its rightful place alongside music, art and literature as an indispensable part of what makes life worth living.*

In reality, however, the science achievement of U.S. students and the science literacy of the general population are mediocre. Such mediocrity does not bode well for our nation's ability to compete with the growing science and engineering talent that is emerging among the many nations with which we will be both competing and cooperating in the future.

Broad recognition of the challenges that our nation faces has led to the creation of several special commissions and resulting reports that address these concerns from both policy and pedagogical perspectives. In support of these efforts, Henry Braun, Richard Coley, Yue Jia, and Catherine Trapani have mined the data of the National Assessment of Educational Progress (NAEP) with sophisticated statistical models in an attempt to identify aspects of U.S. eighth-grade science classrooms that are associated with science achievement. Their report, *Exploring What Works in Science Instruction: A Look at the Eighth-Grade Science Classroom*, identifies instructional strategies and teacher characteristics that appear to make a difference in NAEP science scores. And because teacher practices that make a difference must make a difference for all Americans, these practices are examined for the population generally and for students grouped by their various racial/ethnic characteristics and income level. The results offer much food for thought for anyone concerned with science education.

Michael T. Nettles
Senior Vice President
Policy Evaluation and Research Center

Acknowledgments

The authors wish to acknowledge the helpful comments and feedback provided by the following reviewers: Courtney Bell, ETS; Richard Duschl, Pennsylvania State University; Kate McNeil, Boston College; and Andreas Oranje, ETS. Waverly Van Winkle, Darlene Rosero, and Scott Davis provided data analyses. Richard Pliskin was the editor, Marita Gray designed the cover, and Sally Acquaviva provided desktop publishing services. Errors of fact or interpretation are those of the authors.

The science achievement of U.S. students has been flat for a decade; in fact, in a recent international assessment, U.S. students ranked lower, on average, than their peers in 16 of 30 developed nations.[1] The National Assessment of Educational Progress (NAEP), in addition to serving as the nation's report card, collects information from students, teachers, and schools on instructional practices, teacher characteristics, and school factors, allowing us to peer into U.S. schools and view students and teachers as they interact in the context of schools and classrooms. This report takes advantage of these data and provides a view of eighth-grade science classrooms in 2005, using statistical models particularly suited to this type of data — hierarchical linear models (HLMs). Our intent is to identify teacher characteristics and instructional strategies that are associated with student science achievement. After briefly describing the state of science achievement, teacher characteristics, and instructional strategies across U.S. eighth-grade classrooms, the report presents the results of a three-phase analysis that is designed to illuminate the statistical relationships among these variables.

Science Achievement

Since 1996, the overall science scores of eighth graders have remained unchanged at 149 (on a scale of 0 to 300). Black students, however, showed an increase of 3 scale points since the 1996 assessment. Scores of no other racial/ethnic group improved. Although reduced, the White-Black achievement gap remained substantial. In 2005, White eighth graders scored 36 points higher than Black eighth graders and 31 points higher than Hispanic eighth graders. Male students continued to score higher than females.

With respect to the achievement levels set by the National Assessment Governing Board (NAGB), in 2005, 59 percent of eighth graders scored at or above the *Basic* level, 29 percent performed at or above the *Proficient* level, and 3 percent scored at or above the *Advanced* level.

Teacher Characteristics

Most eighth-grade students (83 percent) have a science teacher who possesses a regular/standard teaching certificate. About half have teachers with less than 10 years of experience teaching elementary or secondary school. The distribution of teacher experience in science teaching inclines toward the lower end of the range — that is, less experience.

About one-third of eighth-grade students have teachers with four years or less experience teaching science. Many students have science teachers with more overall teaching experience than science teaching experience, raising the possibility that science teaching was not the field of choice for those teachers or the subject for which they trained.

Instructional Strategies

As reported by students, several instructional activities stand out as more commonly practiced than others. Reading a science textbook, taking a science test or quiz, and having a teacher conduct a science demonstration were activities that more than half of eighth-grade students indicated occurred almost every day or once or twice a week. Other popular activities (40 percent or more of students reporting frequent activity) were hands-on exercises and investigations, talking about results from hands-on exercises, and working with other students on a science activity or project. Activities occurring much less often included presenting an oral science report, preparing a written science report, using library resources for science, and reading a book or magazine about science. In addition, 70 percent of students indicated that they did science projects that took a week or more.

Understanding Science Achievement

Using multilevel analyses, we found the following student and teacher characteristics to be associated with student science achievement:

- Student demographic characteristics had statistically significant associations with achievement. Black and Hispanic students scored considerably lower than White students, and males scored higher than females. English-language learners and students with disabilities scored much lower than other students.
- Students with many books in the home scored considerably higher than students with fewer books; and students who were absent frequently scored much lower than other students.
- Students whose teachers held a standard teaching certificate scored slightly higher than other students, and students whose teachers' years of total experience exceeded their years of science experience scored slightly lower than other students.

[1] W. Grigg, M. Lauko, and D. Brockway, *The Nation's Report Card: Science 2005* (NCES 2006 – 466), U.S. Department of Education, National Center for Education Statistics, Washington, D.C., 2006 and *Education Week*, "U.S. Students Fall Short in Math and Science," December 4, 2007.

We also used multilevel analyses to identify which pedagogical strategies made an incremental contribution to accounting for differences in students' science achievement after adjusting for student and teacher characteristics. Three subgroups of teacher pedagogy variables showed different patterns of association between the use of the pedagogy and science achievement.

The first group comprises instructional strategies for which increasing frequency was associated with higher average scores:

- Reading a science textbook
- Doing hands-on activities in science
- Writing long answers to science tests and assignments
- Students talking about measurements and results from hands-on activities
- Students working with others on a science activity or project

The second group comprises instructional activities for which increasing frequency was associated with lower average scores:

- Students giving an oral science report
- Students using library resources for science

The third group is characterized by the "Goldilocks" metaphor. Higher scores were associated with a moderate or intermediate frequency for these instructional strategies:

- Students taking a science test
- Teacher doing a science demonstration
- Students discussing science in the news
- Students reading a book or magazine about science
- Students preparing a written science report

Another phase of analysis was undertaken to determine whether the patterns revealed by multilevel analysis hold when the results are disaggregated by various student and school characteristics. Having identified factors that are related to achievement, can we detect differences among groups of students in their access to these "effective" pedagogies or strategies? We chose five instructional strategies, representative of the three subgroups, from among those studied in the multilevel analysis. For each one, we cross-classified students by racial/ethnic group and the percentage of students in the school eligible for the school lunch program, a measure of school disadvantage.

- **Reading a science textbook** – Across all racial/ethnic and school disadvantage groups, scores increase with the frequency of reading a science textbook. The percentage of students in the optimal categories is similar for each combination of race/ethnicity and school disadvantage. Thus, the analysis suggests that mean score differences among racial/ethnic groups cannot be accounted for by differences in exposure to this instructional strategy. These data suggest that it is reasonable to recommend that science teachers should make some use of science textbooks in their teaching.

- **Working with others on a science project** – Across all racial/ethnic and school disadvantage groups, scores increase with the frequency of working with others on a science project. Thus, the analysis indicates that mean score differences among racial/ethnic groups cannot be accounted for by differences in exposure to this instructional strategy. A reasonable inference from these data is that science teachers can make effective use of group work on science projects.

- **Students giving an oral science report** – Across all racial/ethnic and school disadvantage groups, scores decrease with the frequency of students giving an oral science report. The score deficit is most severe for students who reported giving an oral report more often than once or twice a month. Black and Hispanic students and students attending more disadvantaged schools were more likely to experience higher frequencies of this strategy. Since excessive use of this strategy is more likely to be seen in more disadvantaged schools, curtailing this practice may help to close the achievement gap.

- **Teacher doing a science demonstration** – Across all racial/ethnic and school disadvantage groups, scores are lowest in the "never or hardly ever" category and highest in the category of "one or two times a week." For all levels of school disadvantage, Black students are less likely to be exposed to the optimal use of this strategy. Thus, schools may make progress in closing the achievement gap by focusing on this type of strategy.

- **Discussing science in the news** – Across all racial/ethnic and school disadvantage groups, scores are highest in the middle response category, "less often than every day." Black and Hispanic students are less likely than other students to fall into that category, suggesting that the achievement gap may be due, in part, to differences in the frequency of exposure to this instructional strategy.

Many of the findings are in line with the predictions one would make based on the arguments found in the National Academy of Sciences report *Taking Science to School*. Other findings are somewhat puzzling and call for further investigation. In any case, more detailed information on how these practices are implemented in classrooms and information about the contexts in which they are employed should be gathered to provide further evidence with regard to the efficacy of these strategies.

"We live in a society — and a nation, and a world — exquisitely dependent on science and technology, in which hardly anyone knows anything about science and technology."

Carl Sagan[2]

Our nation's educational goals must include both preparing a larger proportion of our youth to enter science and engineering (S&E) fields and nurturing a general population with higher levels of scientific literacy. The need is especially critical in the near future as the baby boom generation moves into the retirement years; the increasing success of other nations in developing their human capital makes this replenishment and expansion so important. Yet, while the production of S&E professionals is central to U.S. economic growth and security, according to a recent report from the National Science Board, the future strength of the U.S. S&E workforce is imperiled by two long-term trends:

- Global competition for S&E talent is intensifying, such that the United States may not be able to rely on the international S&E labor market to fill unmet skill needs.
- The number of native-born S&E graduates entering the workforce is likely to decline unless the nation acts to improve success in educating S&E students from all demographic groups, especially those that have been underrepresented in S&E careers.[3]

These concerns have led to a number of reports addressing both policy and pedagogy. Among the latter, a recent publication of the National Academy of Sciences, *Taking Science to School: Learning and Teaching Science in Grades K – 8*, provides a comprehensive review of what is known about learning and teaching in science.[4] It argues that there are four main strands that describe students who are proficient in science:

- Know, use, and interpret scientific explanations of the natural world
- Generate and evaluate scientific evidence and explanations
- Understand the nature and development of scientific knowledge
- Participate productively in scientific practices and discourses

The report recommends that science instruction "should provide opportunities for students to engage in all four strands of science proficiency." The NAEP design framework for the questions to which students respond in describing their science classrooms predates the report. Nevertheless, many of the questions shed light on the kinds and frequencies of activities in which students are engaged when learning science. Accordingly, the results reported here constitute a significant contribution to the research agenda proposed in *Taking Science to School*.

We undertook this report to better understand the correlates of science achievement among U.S. students, at a critical period in their academic careers, in the context of the science classroom. The large, nationally representative sample of students amassed by NAEP represents a unique resource. Through exploration of the data from the NAEP 2005 science assessment, we hope to provide some insights into which classroom practices and teacher characteristics may be related to science achievement.[5]

Before we turn to more sophisticated analyses, in this section we briefly describe the overall results of the 2005 NAEP science assessment, and provide a national picture of the eighth-grade science classroom with respect to the teachers and their pedagogical strategies.

[2] Press conference held in September 1988 introducing the 1986 NAEP science assessment.

[3] National Science Board, *The Science and Engineering Workforce: Realizing America's Potential*, National Science Foundation, August 14, 2003. See also, Committee on Prospering in the Global Economy of the 21st Century: An Agenda for American Science and Technology, Committee on Science, Engineering, and Public Policy, *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future*, National Academy of Sciences, National Academy of Engineering, and Institute of Medicine of the National Academies, Washington, D.C.: National Academies Press, 2007.

[4] Richard A. Duschl, Heidi A. Schweingruber, and Andrew W. Shouse (eds.), *Taking Science to School: Learning and Teaching Science in Grades K – 8*, Washington, D.C.: National Academies Press, 2007.

[5] The principal tool used is a variant of multiple regression analysis, appropriate for the hierarchical structure of the data collected by NAEP. More information about Hierarchical Linear Modeling (HLM) is provided in the second section of this report.

Science Achievement

Below, we provide overall results for the three grade levels assessed. We present one view of science achievement by examining average scale scores. At grade 4, the average score in 2005 was higher than in earlier years. At grade 8, there was no overall improvement in the average score since 1996. At grade 12, the average score declined since 1996.

We present another view by examining the percentage of students who reach the achievement levels — *Basic, Proficient*, and *Advanced* — set by NAGB, based on recommendations from panels of educators and members of the public, to provide a context for interpreting student performance.[6]

- At grade 4 in 2005, 68 percent of students scored at or above the *Basic* achievement level. 29 percent performed at or above the *Proficient* level and 3 percent scored at or above the *Advanced* level.
- At grade 8 in 2005, 59 percent of students scored at or above the *Basic* level, 29 percent performed at or above the *Proficient* level, and 3 percent scored at or above the *Advanced* level.
- At grade 12 in 2005, 54 percent of the students scored at or above the *Basic* level, 18 percent scored at or above the *Proficient* level, and 2 percent scored at or above the *Advanced* level.

There was some good news regarding minority students' performance in grades 4 and 8. At grade 4, since 2000, average scores increased by 7 points for Black students and by 11 points for Hispanic students. White and Asian/Pacific Islander fourth-graders also improved since 1996, as did Hispanic and Black students. At grade 8, Black students were the only racial/ethnic group to make gains since 1996, and no racial/ethnic group has showed improvement since 2000.[7]

The Eighth-Grade Science Classroom

The design of NAEP allows us to peer into the nation's classrooms and learn about how science is being taught and gather information about the characteristics and qualifications of science teachers.[8] We selected grade 8 because this is when students begin to focus on science coursework, and the grade serves as a gateway to more advanced science courses in high school and ultimately in preparing for an S&E career. It should be noted that NAEP is based on a nationally representative sample of students, not of teachers. Thus, the information and data provided here pertain to the characteristics and practices of teachers of a nationally representative sample of students, and not a nationally representative sample of teachers. Consequently, the percentages reported should be interpreted as the percentage of students whose teachers possess that characteristic or use that practice.

[6] As provided by law, NCES, upon review of congressionally mandated evaluations of NAEP, has determined that achievement levels are to be used on a trial basis and should be interpreted with caution. However, NCES and NAGB have affirmed the usefulness of these performance standards for understanding trends in achievement, and these levels have been widely used by national and state officials. Descriptions of these levels for each grade can be found in Appendix A of the Science Framework for the 2005 NAEP at the NAGB website, http://www.nagb.org/pubs/pubs.html. This framework, used to guide the 1996, 2000, and 2005 assessments, requires assessment in three broad fields — and three elements of knowing and doing science: conceptual understanding, scientific investigation, and practical reasoning. This science framework also specifies that some questions and tasks should assess students' understanding of the nature of science and key organizing themes of science. The nature of science encompasses the historical development and habits of mind that characterize science and technology; themes of science are ideas that transcend the scientific disciplines and give scientists tools for investigating the natural world. Themes included in the framework are systems, models, and patterns of change.

[7] Grigg, Lauko, and Brockway, 2006.

[8] As part of the 2005 NAEP Science Assessment, information was collected from students, teachers, and schools on instructional practices and other school factors. The choice of factors included reflects the perspectives of practitioners, researchers, and policymakers. There may be other school conditions and practices that foster instruction and learning, but the NAEP items represent factors that have been widely discussed in the literature.

Teachers. NAEP data provide information on the years of experience of eighth-grade science teachers, both in terms of teaching elementary or secondary education and in terms of teaching science. Table 1 shows the distribution of experience.

Table 1

Percentage of Eighth-Grade Students, by Years of Experience of Their Science Teachers, 2005

Years	Years Teaching Elementary or Secondary School	Years Teaching Science
0 to 4	25%	32%
4 to 9	24	26
10 to 19	27	25
20 plus	24	17

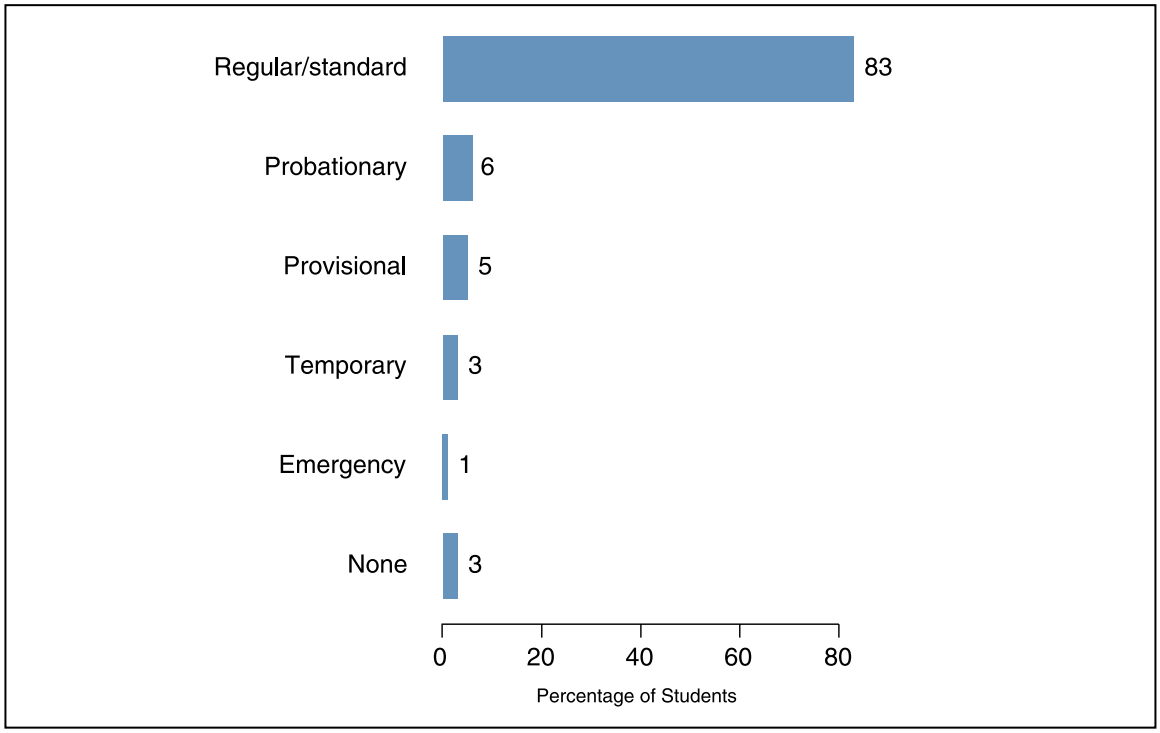
Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Slightly less than half of eighth-grade students have teachers with fewer than 10 years of experience teaching elementary or secondary school. On the other hand, nearly 60 percent of students have teachers with less than 10 years of experience teaching science. Thus, the distribution of teacher experience in teaching science is shifted toward the lower end of the range (i.e., less experience), relative to overall experience. This may mean that science was not the field of choice or preparation for some teachers. For example, elementary school teachers may have moved up to middle school and taken on responsibility for teaching science.

NAEP also provides information on the certification status of eighth-grade science teachers (Figure 1). Most students (83 percent) have a science teacher that possesses a regular/standard certificate. Note that this does not imply that they possess certification in science education.

Figure 1

Percentage of Eighth-Grade Students, by Type of Teaching Certificate Held by Their Teacher

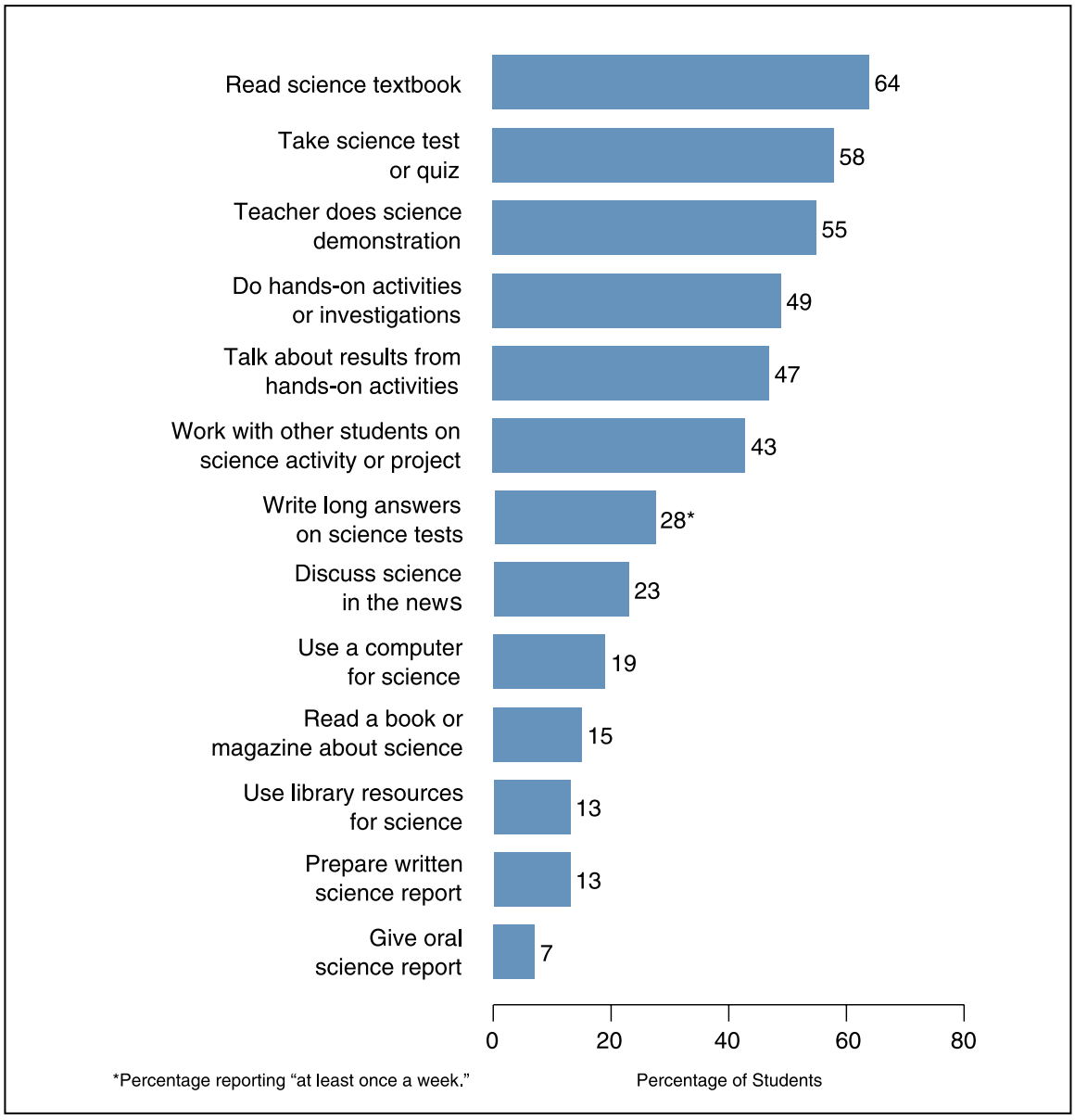


Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Instructional Strategies. This section provides an overview of science teachers' instructional strategies in teaching eighth-grade science, as reported by students. Figure 2 provides a summary of the results.[9] [10] Figure 2 combines the response categories —"almost every day" and "once or twice a week." Several instructional activities stand out as more commonly practiced than others. More than half of students reported having read a science textbook, taken a science test or quiz, and having a teacher conduct a science demonstration. Other popular activities (40 percent or more of students reporting frequent activity) were doing hands-on activities or investigations, talking about results from hands-on activities, and

Figure 2

Percentage of Eighth-Grade Students Reporting Participation in Various Classroom Activities Almost Every Day or Once or Twice a Week



Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

[9] More detailed data on the frequencies are available from the NAEP Data Explorer at http://nces.ed.gov/nationsreportcard/nde/.

[10] As a result of the exploratory variable selection analysis described in Appendix B, three instructional strategies that were reported in NAEP were removed from the HLM analysis. They are teacher uses computers, hands-on work with living things, and hands-on activities that are not listed in the NAEP survey. These strategies were excluded from Figure 2.

working with other students on a science activity or project. In addition, 70 percent of students indicated that they did science projects that took a week or more. Activities occurring much less often included giving an oral science report, preparing a written science report, using library resources for science, and reading a book or magazine about science.

This section of the report has presented some basic data from the eighth-grade 2005 NAEP Science Assessment. We summarized overall achievement on the assessment and provided descriptive statistics on teacher characteristics and instructional practices. These data are both illuminating and provocative. They are provocative because they beg the question of whether science achievement is associated with specific pedagogical strategies or teacher characteristics. If we can gain some understanding of the kinds of science instruction and teacher characteristics that are related to achievement, we may be able to help educators and policymakers craft more effective education and pedagogical policies. Unfortunately, no one study can provide a definitive answer to that question. In particular, NAEP has been primarily designed to provide a snapshot of schooling and achievement at a point in time, and so there are inherent limitations in its ability to answer questions related to the efficacy of one strategy or another. Moreover, the students' responses offer no information on how well the particular strategy was implemented or whether it was appropriate, given the time and place. Both quality and appropriateness are strongly related to effectiveness.

Nonetheless, each NAEP administration produces a rich database drawn from a nationally representative sample of students that, when probed with more sophisticated statistical tools, can yield useful insights into the relationships among student characteristics, teachers' pedagogical strategies, and student achievement. The balance of this report is devoted to revealing and interpreting these relationships. If strong relationships exist, then they can serve as the basis for further investigations to shed light on the causal mechanisms that underlie them.

There are limitations in the utility of findings based on looking at the relationships between an outcome of interest and single explanatory variables, a "one at a time" strategy; the linkages among the different explanatory variables can yield counter-intuitive and even misleading results. The usual remedy is to employ a statistical methodology called multiple regression that allows one to explore the relationship between an outcome — for example, science achievement — and all the potential explanatory variables simultaneously. Moreover, with multiple regression, it is possible to take account of differences among students (with respect to those characteristics that are associated with science achievement) before considering the possible contributions of instructional practices. This is a useful way of addressing questions of pedagogical efficacy. Our principal tool here is a variant of multiple regression analysis, appropriate for the hierarchical structure of the data collected by NAEP — Hierarchical Linear Modeling (HLM).

Students invited to take the NAEP assessment are identified according to a carefully implemented two-stage sampling scheme. First, a systematic random sample of schools is selected, and then a random sample of approximately 30 students is drawn from each school. The original sample for the NAEP Science 2005 grade 8 assessment contained records for 148,595 students drawn from more than 6,300 schools. Slightly more than 40,000 students were excluded from the analyses because of missing data. The most common reason for the missing data was the lack of a match between the student and the teacher, precluding complete analysis of the relationships among science scores, teacher characteristics, and instructional practices. The analysis sample comprised 107,933 students.[11] For further details and a comparison of the reduced sample to the sample used in NAEP reporting, see Appendix A.

We present the full set of student-level variables employed in the analysis in Table 2, in which they are organized for convenience into four categories. There are seven student demographic characteristics, eight student home environment characteristics, 22 teacher pedagogy characteristics, and 11 teacher characteristics.[12] Table 3 contains the six school characteristics that are also included in the analysis.[13]

Each student or school characteristic generates a number of variables, depending on the count of classifications corresponding to the characteristic. For example, there are five classifications for race/ethnicity: White, Black, Hispanic, Asian/Pacific Islander, and American Indian/Alaska Native. One of the classifications is chosen to serve as the base for all comparisons. Four indicator variables are then defined, one for each of the other classifications.[14] We then compile the values of all variables into a database that we use as the input to the regression program.

With the exception of parental education, the student demographic characteristics are derived from school records. Parental education and all student environment characteristics are from the student questionnaire, as are all teacher pedagogy characteristics other than the responses related to the teacher having adequate resources and the total time the teacher spends with his/her class on science instruction in a typical week. Those teacher responses and all teacher characteristics are taken from the teacher questionnaire. Finally, all school characteristics are derived either from the questionnaire completed by the principal or her designee, or from school records.

The two-stage sampling scheme described above generates a hierarchical data structure of students nested within schools. To properly analyze this type of data, multilevel or hierarchical regression models are generally employed. Unlike single-level regression models, multilevel models properly take account of both the statistical relationships among students in the same schools and those among students in different schools, as well as providing appropriate standard errors for judging the statistical significance of the results. In the first level of a typical two-level model, a separate regression equation is fit to the data from each school. In the second level, the sets of school-specific regression coefficients from the first level are represented as functions of school-level characteristics. For more details, consult Appendix B.

The principal goal of this stage of the study was to identify those pedagogical strategies that make an incremental contribution to accounting for differences among students in science achievement after adjusting for differences among students with respect to demographic characteristics, home environment, and teacher characteristics. With this goal in mind, we employed a stagewise exploratory regression strategy.

[11] Additional analyses were undertaken to determine if excluding the students with missing *Teacher Match Code* will have a significant impact on the relationship between science achievement and teacher pedagogies. The results show that there is no noticeable difference between estimates of regression coefficients from the two different samples. All the estimates using the full NAEP reporting sample are within one standard error of the estimates using the sample that includes students who could not be matched with their teacher.

[12] Items on the student questionnaire related to student attitudes towards science were not employed because of the likelihood that students' responses could be causally related to their teachers' instructional strategies. Including them in the model would then result in biased estimates of the focal parameters.

[13] The NAEP background questionnaires are available online. For the 2005 NAEP Science Grade 8 Teacher Questionnaire, go to nces.ed.gov/nationsreportcard/pdf/05BQteacherG8sci.pdf; for the student questionnaire, go to nces.ed.gov/nationsreportcard/pdf/05BQstudentG8science.pdf; for the school questionnaire, go to nces.ed.gov/nationsreportcard/pdf/05BQschoolG8.pdf.

[14] The indicator variable corresponding to a particular classification takes the value "1" if the individual falls in that classification, and takes the value "0" otherwise. The fitted regression coefficient attached to such an indicator variable represents the average difference in science scores between the students in the corresponding classification and the students in the base classification, holding all the other variables in the model equal.

Table 2

Student-level Variables Included in the HLM Analysis

Student Demographic Characteristics	Student Home Environment Characteristics	Teacher Pedagogical Strategy Characteristics	Teacher Characteristics
Gender	Newspaper at home	Student doing hands-on with electricity	Teacher holding above bachelor academic degree
Race/ethnicity	Magazines at home	Student doing hands-on with chemicals	Teacher holding standard teaching certificate
IEP disability status	Number of books at home	Student doing hands-on with rocks	Years of teaching experience greater than years of science teaching experience
Eligibility for free/reduced-price school lunch	Computer at home	Student doing hands-on with magnifying glass/microscope	Teacher's college focus being science and/or education
Status as an English-language learner (Limited English Proficient)	Encyclopedia at home	Student doing hands-on with barometer	Teacher being a leader for science education
Title 1 participation	Number of pages read in school and for homework	Student doing hands-on with simple machines	Teacher undergraduate study major in science
Parental education	Number of absences	Student doing science projects in school that take a week or more	Teacher undergraduate study minor in science
	Talking about things studied in school with family	Student writing long answers to science tests/assignments	Teacher undergraduate study major in education
		Student reading a science textbook	Teacher graduate study major in science
		Student reading a book or a magazine about science	Teacher graduate study major in education
		Student discussing science in the news	Teacher graduate study minor in education
		Student doing hands-on activities in science	
		Student talking about the measurements and results from hands-on activities or investigations	
		Student using a computer for science	
		Student working with others on a science activity/project	
		Student giving an oral science report	
		Student preparing a written science report	
		Student taking a science test	
		Student using library resources for science	
		Teacher doing a science demonstration	
		School resources for teachers	
		Total time teacher spending with his/her class on science instruction in a typical week	

Table 3

School-level Variables Included in the HLM Analysis

Percentage of minority students (Black and Hispanic)
Percentage of limited-English-proficient students
Percentage of students eligible for school lunch program
Region of the country
Percentage of students enrolled for special education
School type (private or public)

Recall that we grouped substantively related characteristics into the five categories. We entered variables into the student-level regression model sequentially by category, according to a predetermined order, based on both statistical considerations and interpretive goals. The order was: student demographic characteristics, student home environment characteristics, teacher pedagogical strategies, and teacher characteristics.

At each stage, we retained the set of variables corresponding to a particular characteristic only if the regression coefficients associated with the variables in the set generally exceed a predefined statistical threshold — for example, significant at the 0.05 level. At the next stage, we entered into the regression the variables remaining from the previous stage together with all the variables in the next category. We continued the process until the last category was entered. In the final model, the estimated regression coefficients of the teacher pedagogy variables, for example, represent the strength of the relationships between NAEP science scores and instructional strategies after taking account of measured differences among students who are not influenced by the teacher, as well as differences in the characteristics of their teachers. For more detail on the methodology, see Appendix B. We refer to this set of analyses as Phase 1.

In Phase 1, the burden of proof, so to speak, is on the pedagogical variables. That is, the strength of the estimated statistical relationship between those variables and NAEP science scores is influenced by the other variables in the model and their relationship to both the pedagogical variables and NAEP science scores. For example, suppose that students whose parents have higher levels of education are also more likely to be exposed to pedagogical strategies that are truly more effective. Suppose further that higher parental education is strongly positively associated with science achievement. Then it is quite possible that the final fitted model will underestimate the strength of the relationship between those pedagogical strategies and NAEP science scores. With these considerations in mind, it is appropriate to conduct a supplemental analysis that examines the relationship between science achievement and teacher pedagogies directly — that is, without introducing student demographics and home environments into the model. This is the purpose of the second set of analyses, which we refer to as Phase 2.

Finally, it is important to recognize that the patterns in the regression coefficients in the final fitted models represent a type of average over all students in the analysis set. Such averaging may not reflect the patterns that would be observed when the models are fit to subgroups of students defined by various student and school characteristics. The degree of consistency of the patterns in the relations among variables across many such subgroups aids in proper interpretation of the findings. Moreover, examining the data at this level can reveal if there is differential exposure to putatively supportive instructional strategies. In other words, to what extent are student characteristics associated with teachers' pedagogy? Do students have differential access to "effective" instructional strategies based on their characteristics? We refer to this set of supplemental analyses as Phase 3.

Cautions in Interpretation

Notwithstanding the many strengths of the NAEP design and the rigor of its implementation, there are some important limitations to NAEP-based analyses that should be borne in mind. First, the analyses rely heavily on responses provided by students. These responses only imperfectly reflect what actually occurred in the classroom because of the way students interpreted the questions and the accuracy of their recall. In addition, there is no information with respect to the quality and appropriateness of the strategy as it was implemented in the classroom.

Second, and equally important, is the fact that NAEP is an observational study rather than a true experiment. Teachers typically choose which strategies to adopt and when to employ them, rather than having the strategies randomly allocated to them by some external mechanism. The implication is that the relationships uncovered, should there be any, cannot be interpreted causally. That is, if it is determined that students exposed to strategy A score higher on average than those exposed to strategy B, other things being equal, one cannot conclude that the difference is due to the differential effectiveness of the strategies — as one could, in principle, were this an

experimental study. In the case of the NAEP sample, there may be some student characteristics that are associated both with a tendency to be exposed to strategy A and with higher science performance. Similarly, there may be teacher characteristics that are associated both with greater general efficacy and the choice of a particular strategy. These circumstances are instances of selection bias.

Selection bias is always a danger in observational studies. To mitigate its effects, investigators often collect ancillary data on the students and employ regression methods to adjust the outcomes for differences among students in these measured variables. It is hoped that the adjusted outcomes are less susceptible to selection bias. In this report, we have adjusted science scores for a host of student demographics and home environment characteristics. The relationships between these adjusted outcomes and teacher pedagogy, if any, are more credible than those found with unadjusted outcomes.

At the same time, as explained in the rationale for the Phase 2 analyses, statistical adjustment through regression may be a double-edged sword, because it can result in under- or overestimation of the strength of certain relationships. The direction and extent of the bias depends, in part, on the degree of self-selection that is not accounted for in the model. That is, if there are unmeasured student characteristics that are not captured by the variables explicitly included in the regression model, and if they are associated both with the particular pedagogy to which the student is exposed and with science performance, then the estimated effects of different pedagogies can be confounded with differences among the students with respect to those unmeasured characteristics.

In particular, NAEP is a cross-sectional study in which prior measures of academic achievement are not collected. A student's observed level of science achievement as measured by NAEP depends not only on her experiences in the current year but, to some degree, also on her educational experiences, in school and out, up to the time of the NAEP assessment. Adjustments that include prior test scores are more likely to account for important differences among students than are adjustments based on student and home characteristics alone.

Statistical adjustment is an appropriate and widely used approach to the analysis of observational data. However, rarely, if ever, can it fully compensate for the lack of randomization. Accordingly, the results must be treated with caution. As explained previously, we have undertaken two sets of supplementary analyses to enhance the credibility of the findings of the main study.

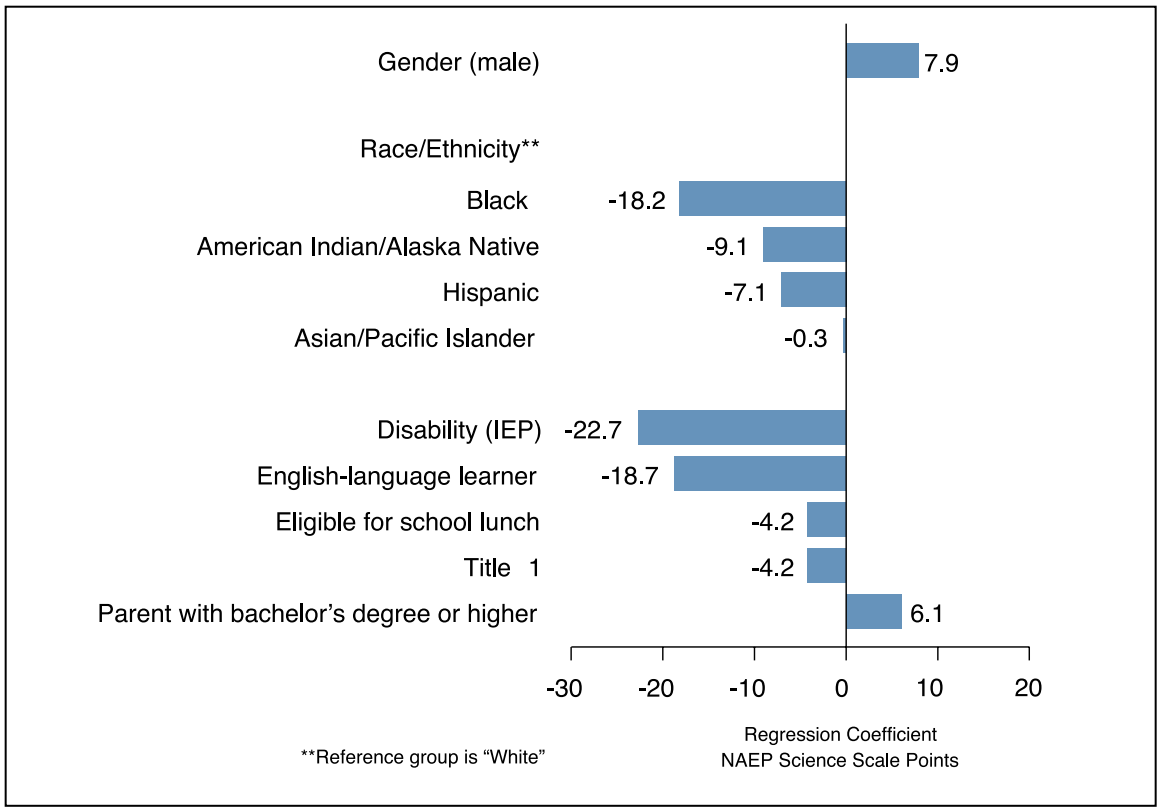
Phase 1

As explained earlier, the final regression model is the culmination of a sequence of exploratory analyses that systematically and sequentially examined the relationships between NAEP science scores and various categories of explanatory variables. The results for this model are contained in Figures 3 to 7, corresponding to the five categories of characteristics.[15] They have a common format, with the variables related to each retained characteristic displayed alongside the fitted regression coefficients. Bars to the right of the vertical zero line indicate that there is a positive statistical effect of the labeled category compared with the baseline category. Similarly, bars to the left of the vertical zero line indicate a negative statistical effect. The length of a bar is proportional to its magnitude, and the corresponding numerical value is adjacent to the bar.[16] We provide tables showing estimated standard errors and the corresponding p-values in Appendix C.

In view of the extremely large size of the NAEP sample, nearly all the estimated regression coefficients are highly statistically significant. Accordingly, practical significance assumes greater salience in interpreting the results. We suggest that estimated effects with a magnitude greater than two score points are notable. This is based both on the results we obtained for teacher characteristics and what changes in NAEP scores from one administration to another generally elicit comments from policymakers.

Figure 3

Estimated* NAEP Science Score Gains or Losses Associated With Student Demographic Characteristics, Grade 8, 2005



* Each estimate is adjusted for teacher characteristics, teacher pedagogical strategies, and for other student characteristics in the model.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

[15] Only the variables with associated regression coefficients that exceed a pre-defined statistical threshold (i.e., significant at the 0.05 level) are presented in the figures.

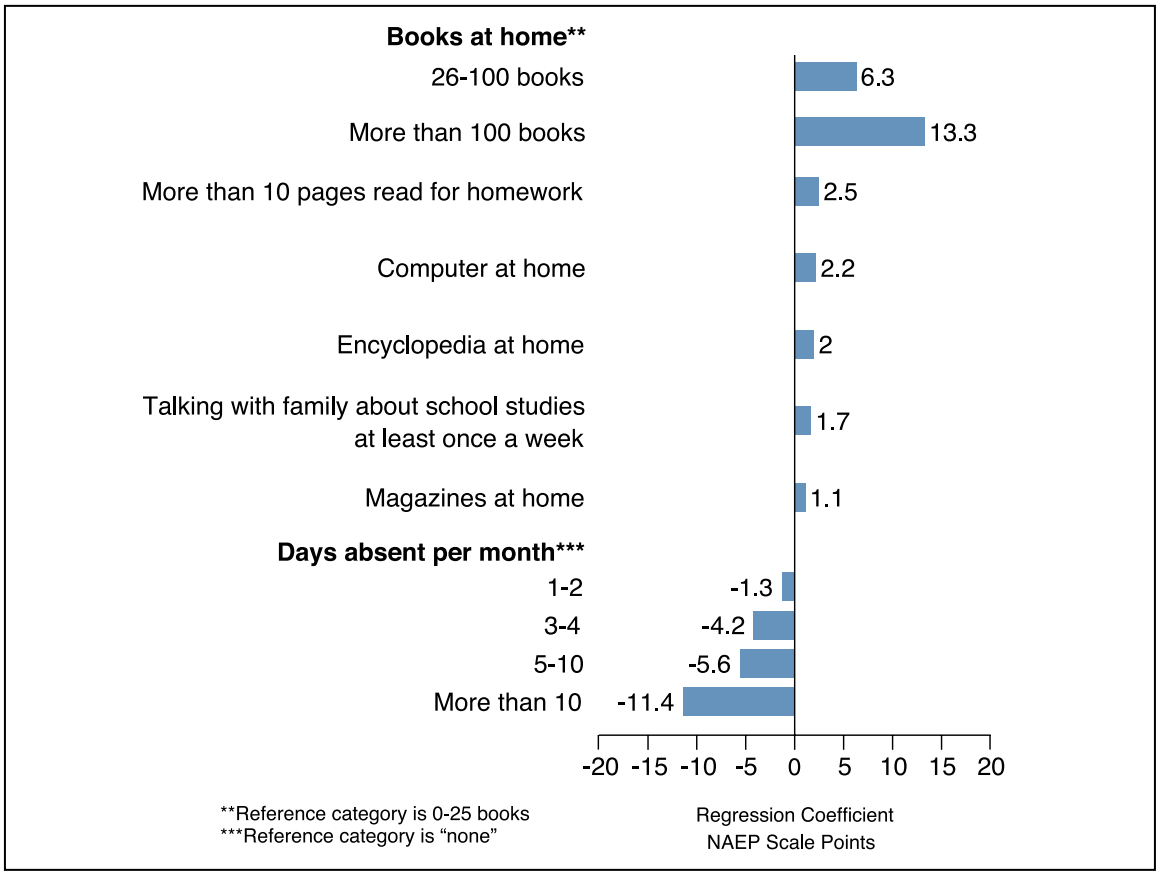
[16] In statistical parlance, the coefficients are partial regression coefficients since they have been adjusted for the other variables in the model.

Examination of Figure 3 reveals that the regression coefficients corresponding to student demographics other than the ethnic classification "Asian/Pacific Islander" are all large in magnitude and statistically significant. For example, regarding gender, the base category is female. Thus, the interpretation of the corresponding coefficient is that, on average, males score 7.9 points higher than females. For the race/ethnicity characteristic, White students are the base category. Thus, the fitted coefficient for Black students means that, other things being equal, on average Black students score 18.2 points lower than White students.[17] The "achievement gaps" associated with English-language learners and with students with disabilities are strikingly large, 18.7 and 22.7 points, respectively.

Figure 4 shows the results for variables related to student home environment characteristics. Some of the coefficients are very substantial. For example, compared with students with fewer than 25 books in the home, students with more than 100 books in the home score 13.3 points higher on average. Similarly, compared with students reporting not being absent, those reporting being absent 10 or more days per month score, on average, 11.4 points lower.

Figure 4

Estimated* NAEP Science Score Gains or Losses Associated With Student Home Environment Characteristics, Grade 8, 2005



* Each estimate is adjusted for teacher characteristics, teacher pedagogical strategies, and other student characteristics in the model.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

[17] The phrase "other things being equal" means that when comparing two levels of a characteristic (e.g., Whites and Blacks of the student characteristic race/ethnicity), the values of all the other variables in the model are held constant. For the sake of brevity, this phrase is not repeated but should be assumed by the reader.

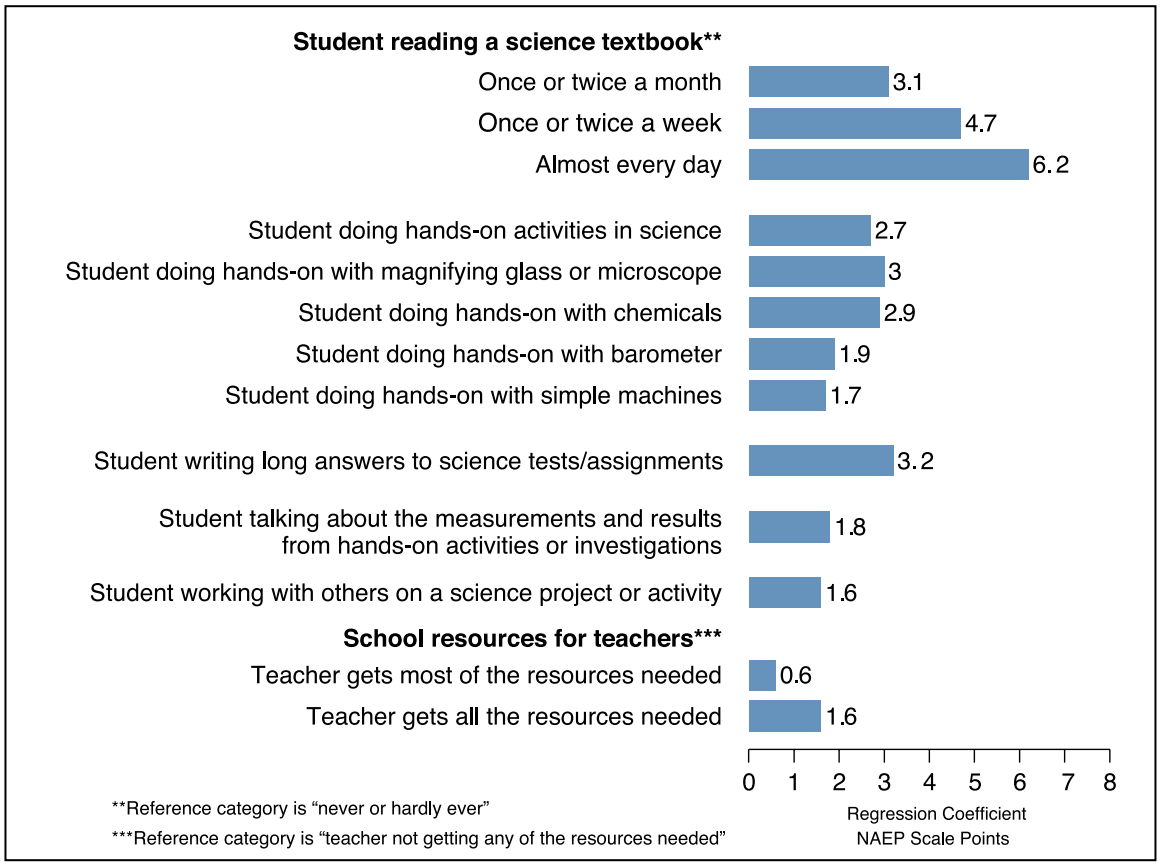
The data in Figures 5, 6, and 7, which display the results for variables associated with teacher pedagogies, are particularly interesting. We have organized the variables into three subgroups according to the pattern in the coefficients. The first group (Figure 5) comprises those strategies whose reported use is associated with higher average scores or, if there are more than two frequency levels, then increased frequency is associated with higher average scores. The second group (Figure 6) displays those strategies for which increasing frequency of reported use is associated with lower average scores. The third group (Figure 7) comprises those strategies with three or more frequency levels in which the highest average score is associated with an intermediate frequency — what we have termed a "Goldilocks" pattern.

As shown in Figure 5, students doing hands-on activities in science score 2.7 points higher than those who don't, on average. Interestingly, students responding in the affirmative to a more specific question — for example, "Do you do hands-on activities with simple machines?" — earn an additional premium of 1.7 points, on average. Thus, students who respond consistently to the hands-on questions — i.e., in the affirmative to the general question, to at least one specific question, and who talk about the results — score at least 6.2 points (2.7 + 1.7 + 1.8) higher than those students who report that their teachers do not employ hands-on activities.

It is also evident that reading a science textbook more frequently, being asked to write long answers on tests or assignments, and working with others on projects are each associated with higher average scores.[18] Finally, students whose teachers agree that they have all the resources they need score significantly higher than those students whose teachers indicate that they have none or few of the resources they need.

Figure 5

Estimated* NAEP Science Score Gains Associated With Teacher Pedagogical Strategies, Grade 8, 2005



* Each estimate is adjusted for teacher characteristics, student characteristics, and other teacher pedagogical strategies in the model.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

[18] Reading a textbook may refer to homework assignments, to the use of text materials directly in class, or both. Note that there is no information on the kind of textbook referred to.

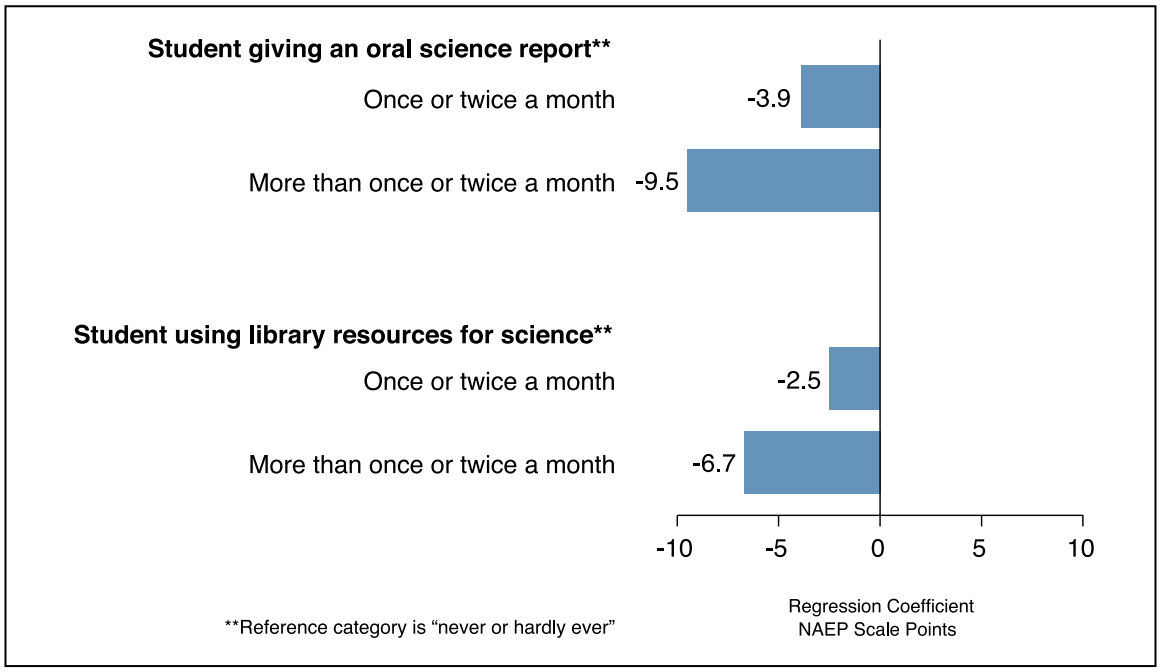
There are just two strategies in the second group where increased frequency is associated with lower scores (Figure 6): giving an oral science report and using library resources for science. In both cases, the score deficit associated with the highest frequency level is quite substantial. There is no obvious explanation for this finding. One can speculate that in many cases, these activities are not well coordinated with the curriculum and act more as a distraction than as a support for student learning.

Perhaps the most surprising results are found in the last group, shown in Figure 7, comprising five strategies where the highest average score is associated with an intermediate frequency — again — that might be termed a "Goldilocks" pattern. For example, compared with students who never take a science test (the baseline group), those who take a test once or twice a month score 8.3 points higher on average. Those who take a test once or twice a week score 6.1 points higher, while the average score of those who take a test almost every day is not statistically different from the average score of the baseline group. To take another example, students who prepare a written science report once or twice a month score, on average, one point higher than those who never or rarely write a science report (the baseline group). However, those who prepare a report more often than twice a month score on average 2.9 points lower than the baseline group. These patterns bear scrutiny, and we will examine them further below.

We retained only two teacher characteristics for the final model. As shown in Figure 8, students whose teachers hold a standard teaching certificate scored 1.5 points higher than those students with teachers having other or no credentials. Students whose teachers' total experience exceeds their science teaching experience scored 1.4 points lower than other students. One possible explanation for this finding is that science teaching was not the original field of choice for these teachers and so these teachers, perhaps, may not be as well prepared to teach science.

Figure 6

Estimated* NAEP Science Score Gains Associated With Teacher Pedagogical Strategies, Grade 8, 2005

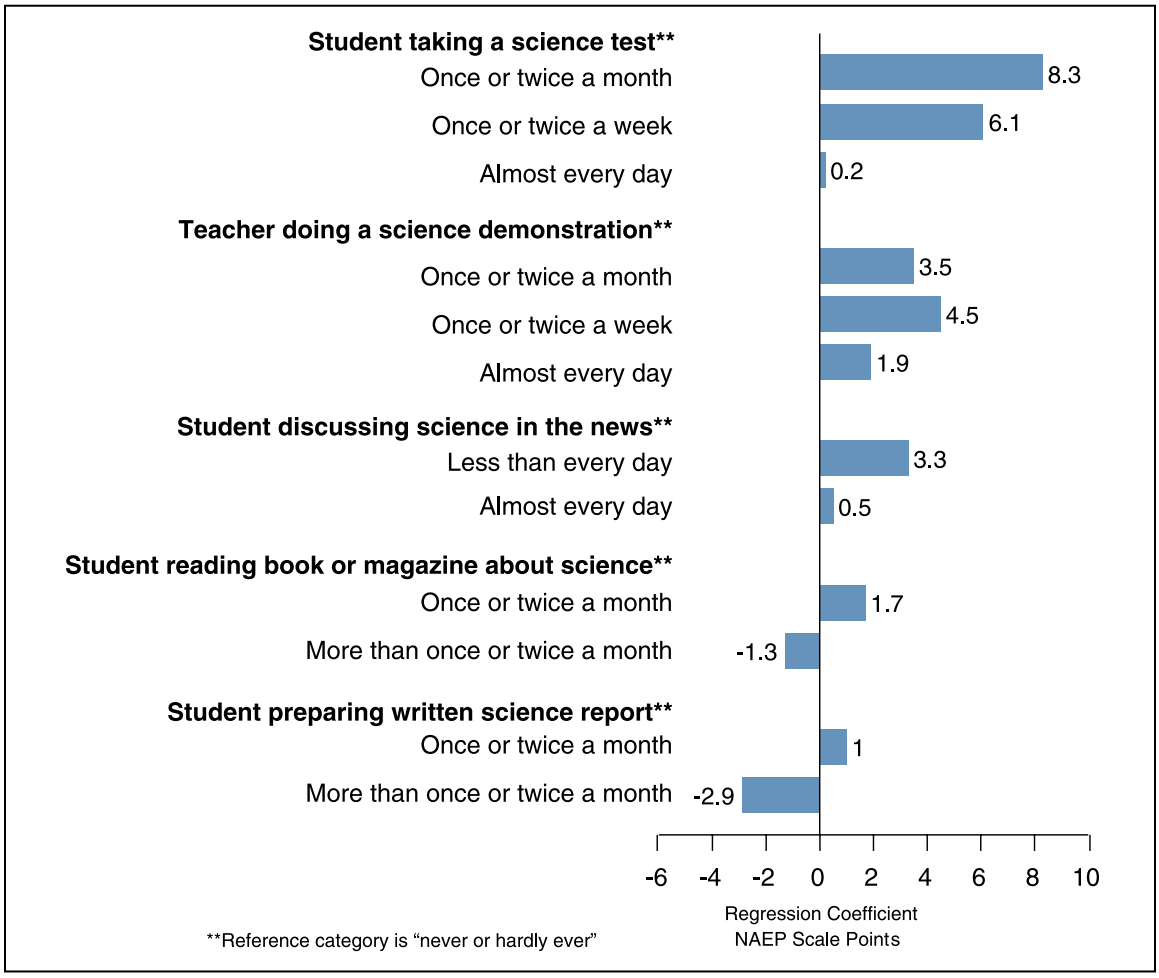


* Each estimate is adjusted for teacher characteristics, student characteristics, and other teacher pedagogical strategies in the model.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Figure 7

Estimated* NAEP Science Score Gains or Losses Associated With Teacher Pedagogical Strategies, Grade 8, 2005

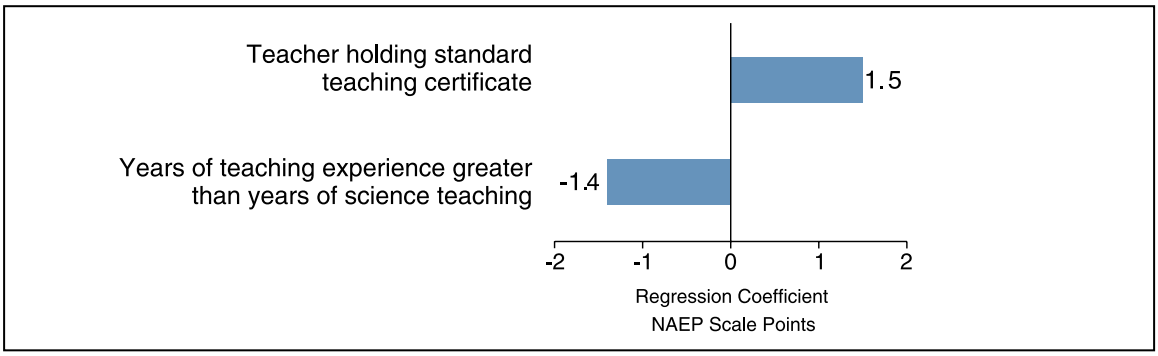


* Each estimate is adjusted for teacher characteristics, student characteristics, and other teacher pedagogical strategies in the model.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Figure 8

Estimated* NAEP Science Score Gains or Losses Associated With Teacher Pedagogical Strategies, Grade 8, 2005



* Each estimate is adjusted for student characteristics, teacher pedagogical strategies, and other teacher characteristics in the model.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Finally, we retained four school characteristics in the model as useful predictors of differences among schools in (adjusted) average science scores (Figure 9). In particular, schools enrolling larger proportions of Black and Hispanic students and larger proportions of students eligible for school lunch programs have lower average science scores. Regionally, only schools in the West have average scores significantly different from schools in the Northeast (the base category). Note that the difference in magnitude of the coefficient for the percentage of disadvantaged minorities compared with the others is the result of differences in scaling: The percentage variable assumes values from 0 to 100, while the other variables are coded as 0 or 1.

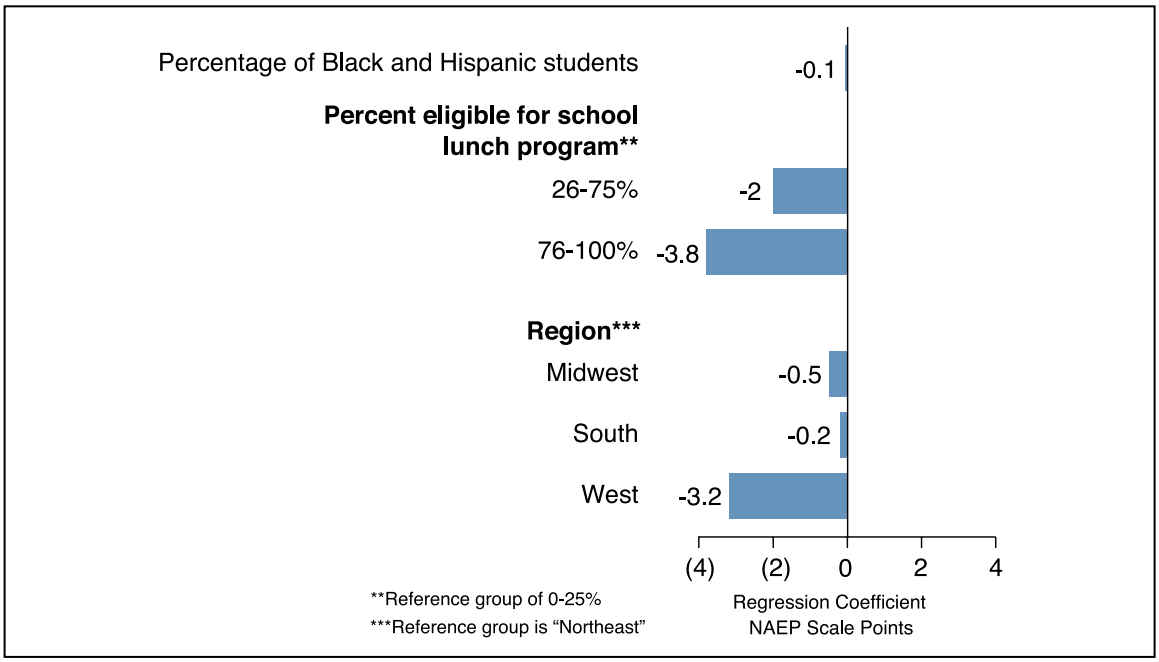
Phase 2

The Phase 1 analyses were intended to answer the question of which pedagogical strategies make an incremental contribution to accounting for differences among students in science achievement after differences among students regarding demographic characteristics and home environments, as well as teacher characteristics, have been taken into account. As indicated at the outset, these findings should be interpreted carefully. The apparent strength of the retained strategies' relationships to science achievement, as quantified by the magnitudes of the corresponding regression coefficients, may well underestimate their true strength. This would occur if students with those demographic characteristics and home environments that were positively correlated with science achievement were also more likely to be exposed to effective pedagogies. That might be the case, for example, if schools enrolling students with higher socioeconomic status also hired more experienced, better qualified science teachers who tended to employ these pedagogies.

In Phase 2, the models were fit using only teacher characteristics and teacher pedagogies without controlling for other variables. We examined only those teacher characteristics and teacher pedagogies that were retained in the final model of Phase 1. We restrict our attention to the model in which both categories of variables are included. Figure 10 displays the results for teacher characteristics, and Figures 11, 12, and 13 display the results for teacher pedagogies. Note that we have retained the ordering in Figures 5, 6, and 7.

Figure 9

Estimated* NAEP Science Score Losses Associated With School Characteristics, Grade 8, 2005



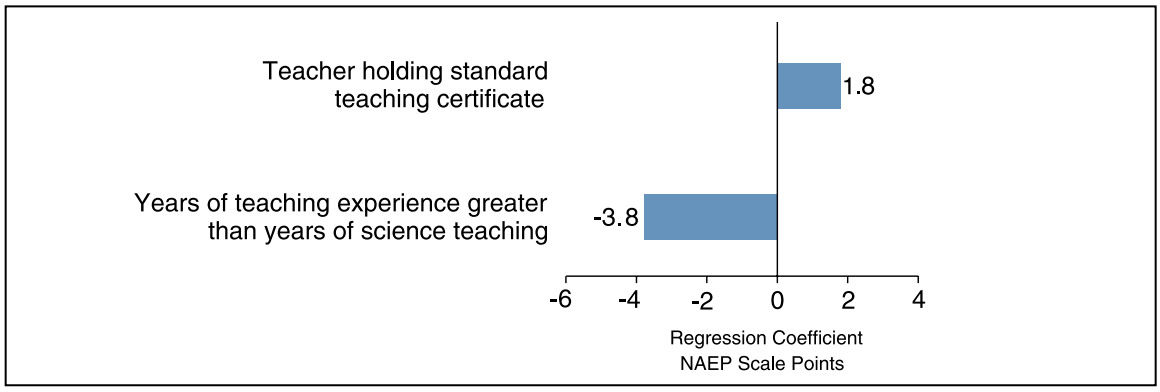
* Each estimate is adjusted for teacher characteristics, teacher pedagogical strategies, and student characteristics in the model.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Figure 10 shows that both teacher characteristics are statistically related to science performance. The regression coefficient for the teacher having a standard certificate is nearly the same as it was in Phase 1. However, the coefficient for the differential teaching experience is negative again, but substantially larger than it was in Phase 1.

Figure 10

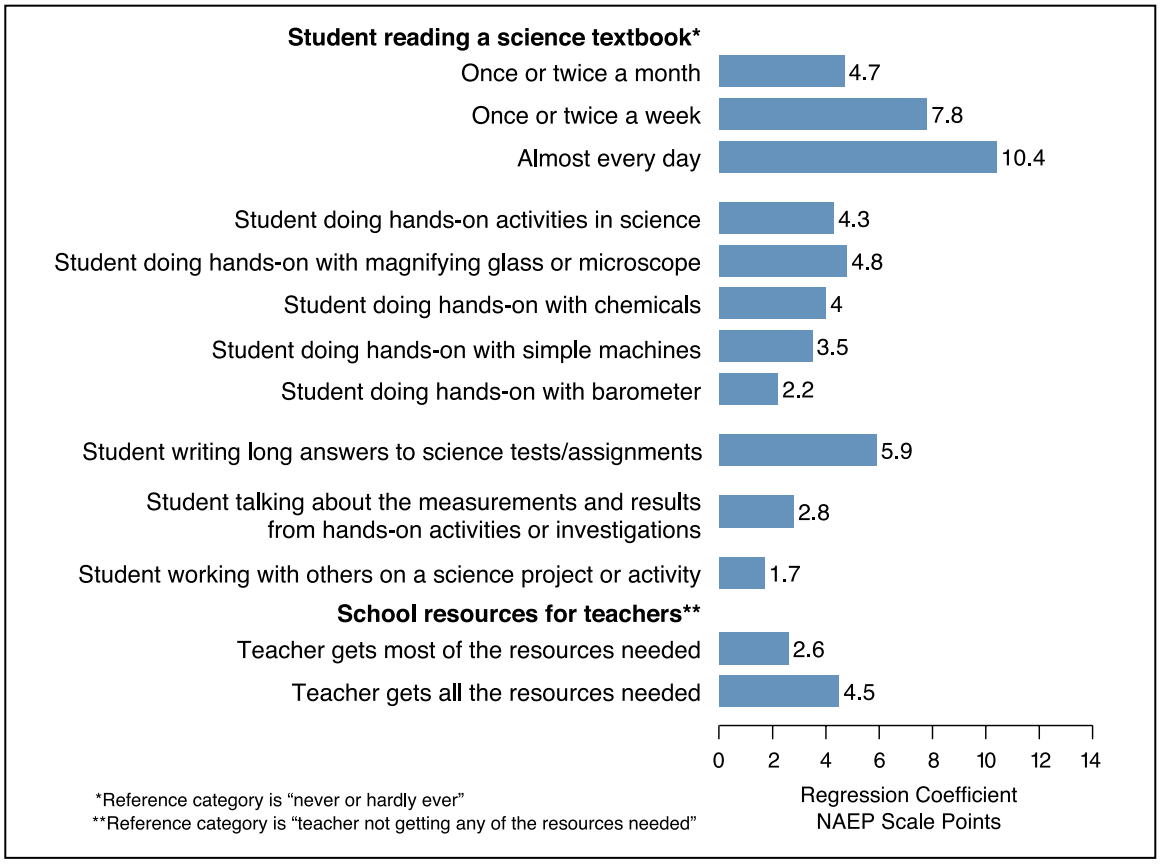
Estimated Unadjusted NAEP Science Score Gains and Losses Associated with Teacher Characteristics, Grade 8, 2005



Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Figure 11

Estimated Unadjusted NAEP Science Score Gains and Losses Associated with Teacher Pedagogical Strategies, Grade 8, 2005



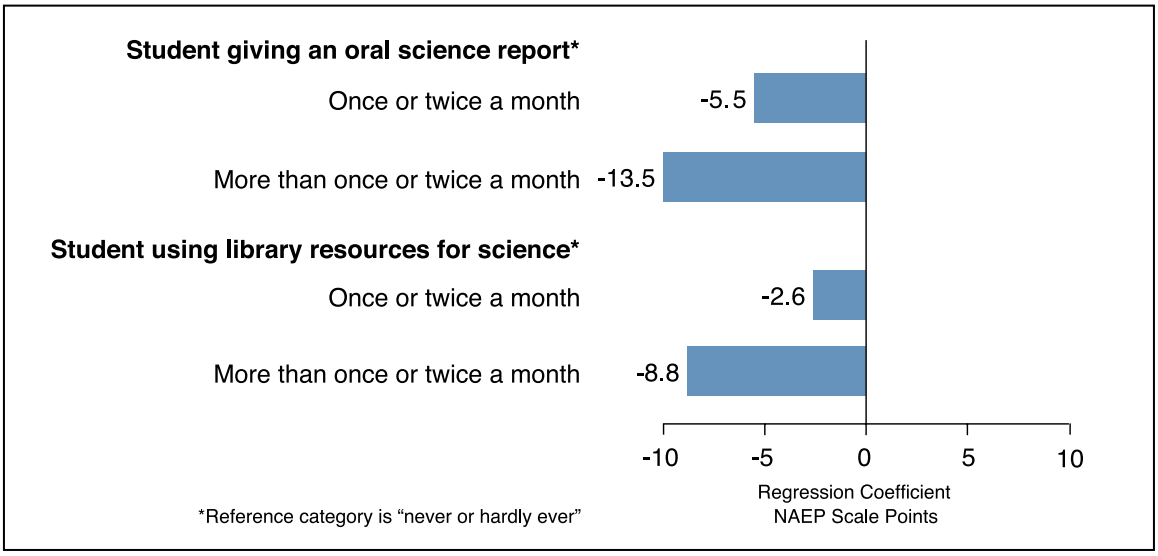
Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Figures 11, 12, and 13 reveal that all the retained pedagogical strategies are strongly related, in a statistical sense, to science performance. As one would expect, the magnitudes of the regression coefficients are larger, and sometimes considerably larger, than the corresponding coefficients in Phase 1 (Figures 5, 6, and 7). The signs of the coefficients are the same. Moreover, for those characteristics associated with multiple levels of response, the patterns in the regression coefficients are identical to those in Phase 1. In general, the sizes of the coefficients observed in Phase 2 are comparable to the sizes of the coefficients observed in Phase 1 for student home environment variables.

Thus, from a qualitative perspective, it appears that prior adjustment for student demographics and home environment does not materially affect the findings. What remains an issue is which set of regression coefficients (i.e., those reported in Phase 1 or those in Phase 2) is closer to what one would find in a controlled experiment. There is no way to answer the question with the present data; however, one can speculate that the two sets of results bracket the coefficients that would be obtained under ideal conditions.

Figure 12

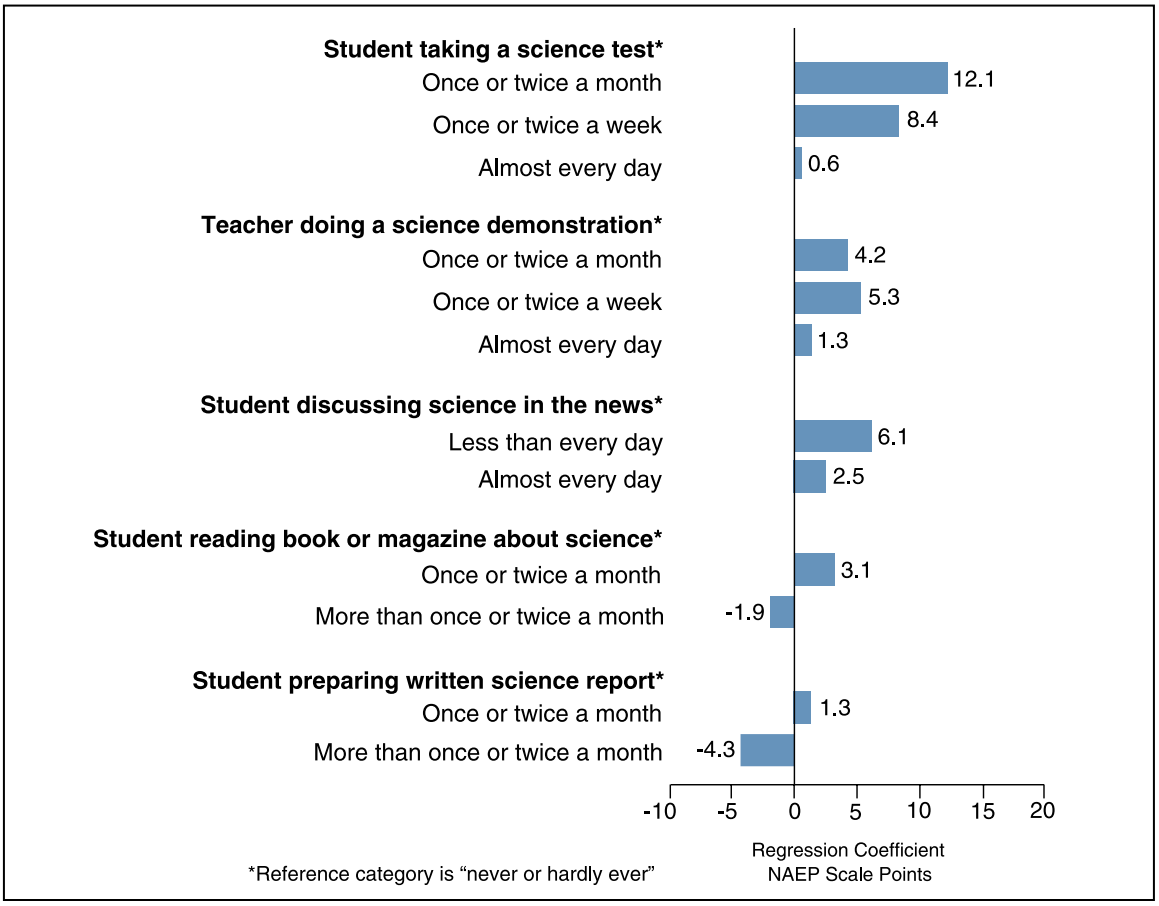
Estimated Unadjusted NAEP Science Score Gains and Losses Associated with Teacher Pedagogical Strategies, Grade 8, 2005



Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Figure 13

Estimated Unadjusted NAEP Science Score Gains and Losses Associated with Teacher Pedagogical Strategies, Grade 8, 2005



Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Phase 3

The findings of Phases 1 and 2 are encouraging in that they support the hypothesis that teachers' instructional practices do make a difference in their students' success. Unfortunately, the cross-sectional nature of the NAEP design precludes reaching more definitive conclusions. At the same time, there is value in determining whether the overall patterns in the regression coefficients in the final fitted models are also observed when the results are disaggregated by various student and school characteristics. Thus, for example, is the "Goldilocks" pattern in average scores for the strategy "students taking a science test" replicated when viewed for groups of students who are similar with respect to race/ethnicity and school characteristics? A reasonable consistency in the results would add to the credibility of the aggregate findings. In addition, examining the data at this level can show whether there is differential exposure to putatively supportive instructional strategies. In other words, to what extent are student characteristics associated with teachers' pedagogy? Do students have differential access to "effective" instructional strategies based on their characteristics?

To begin with, students are cross-classified by two factors: (1) race/ethnicity (five levels), and (2) percentage of students in their school eligible for the school lunch program (three levels). This generates a 5x3 table with 15 cells (See Table 4). Thus the students in each cell are homogeneous with respect to two key characteristics, one individual (their race/

ethnicity) and one collective (school disadvantage). Note that for this analysis, we draw on the full set of 143,412 students rather than the smaller number employed in the regression analyses (see Appendix A).[19]

Table 4

Distribution of NAEP Science Scores and Eligibility for School Lunch Program by Racial/Ethnic Group, Grade 8

Percentage of Students in School Eligible for School Lunch			
Racial/Ethnic Group	0-25% Score (%)	26-75% Score (%)	76-100% Score (%)
White	166 (42)	155 (54)	143 (3)
Black	139 (12)	125 (56)	114 (33)
Hispanic	144 (12)	129 (52)	120 (36)
Asian/Pacific Islander	168 (42)	149 (45)	138 (13)
American Indian/ Alaska Native	150 (9)	137 (50)	113 (41)

Note: Entries represent reported average scaled score. Values in parentheses are percentages of students in the corresponding race/ethnicity classification enrolled in schools with the indicated level of the proportion of the student population eligible for the school lunch program.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

To carry out this analysis, we selected five strategies from the set we examined in Phase 1: two from group 1 (positive effects), one from group 2 (negative effects), and two from group 3 (quadratic, or "Goldilocks," effects). For each strategy, we disaggregated the students in each cell of Table 4 by their reported level of the strategy. For each such subclassification, we display the mean science score and the corresponding percentage. The results are found in Tables 5 to 9.

The presentation begins with Table 4, which displays the basic statistics for the cross-classification. Consider the left-most column, corresponding to schools in which 0 percent to 25 percent of students are eligible for the school lunch program. The table entry for Black students in that column indicates that the average NAEP science score for Black students attending such schools is 139 and that 12 percent of Black students attend such schools. Examination of Table 4 reveals that:

- White and Asian/Pacific Islander students are least likely to be enrolled in schools with the most disadvantaged populations (3 percent and 13 percent, respectively), while American Indian/ Alaska Native, Hispanic and Black students are the most likely to be enrolled in the most disadvantaged schools (41 percent, 36 percent, and 33 percent, respectively).
- For each racial/ethnic group, mean scores decrease with increasing levels of disadvantage of the school population. For White students, the mean scores are 166, 155 and 143. The differences in mean scores between the least and most disadvantaged categories of schools range from 23 points (166 – 143) for Whites to 37 points (150 – 113) for American Indian/Alaska Natives.
- For a given level of school disadvantage, mean scores are generally highest for White and Asian/ Pacific Islander students and lowest for Black and Hispanic students. The patterns displayed in Table 4 should be kept in mind for the remainder of the discussion.

Table 5 contains results for the strategy "reading a science textbook." The Phase 1 analysis showed that averaged over all racial/ethnic and school disadvantage groups, scores increase with the frequency of reading a science textbook. In each main panel of the table, there are four sets (rows) of entries, corresponding to the four different frequency categories. Consider the top main panel, which presents data for White students. The first column in that panel corresponds to White students attending schools with no more than 25 percent of the students eligible for the school lunch program. As indicated in Table 4, the students in this group represent 42 percent of all White students and have a mean score of 166. In Table 5, these students are further classified by the frequency of the strategy. There are four response categories. The first category is "never or hardly ever." Approximately 19 percent of the group falls in this category with a mean score of 158. The next category is "1-2 times a month." Approximately 20 percent of the group falls in this category with a mean score of 166. The third category is "1-2 times a week." Approximately 36 percent of the group falls in this

[19] Since the analyses reported here do not require a student-teacher match, it was decided to employ the full NAEP reporting sample in exploring the patterns of interest.

category with a mean score of 169. The last category is "almost every day." Approximately 26 percent of the group falls in this category with a mean score of 169.

There are a number of interesting patterns evident in this table:

- First, the disaggregated data support the findings in Phase 1; there do not appear to be any anomalies.
- Generally, the relationship between the frequency of reading a science textbook and average scores is the same for each combination of race/ethnicity and school disadvantage. The lowest mean corresponds to the base category "never or hardly ever" and rises with increasing frequency. The largest gap between categories is that between the base category and the successive, higher-frequency categories.
- Overall, this trend corresponds to the trend observed in Figure 5. However, the differences in mean scores between the base category and the others are rather larger than the sizes of the corresponding regression coefficients in Figure 5. Since the latter have been adjusted for the other strategies included in the model, this comparison suggests that most students tend to be exposed to either multiple supportive or multiple non-supportive strategies.
- With the exception of American Indian/Alaska Native students, for each racial/ethnic group, both the proportion of students in the base category and the gap in mean scores between the base category and the others is similar across levels of disadvantage.
- The overall ordering of racial/ethnic groups by mean scores is reflected in the ordering observed when students are cross-classified by both school disadvantage and frequency of the strategy. For example, consider those students in schools where 26 percent to 75 percent of students are eligible for the school lunch program and who fall in the base category for this strategy. Mean scores are highest for White and Asian/Pacific Islander students (146 and 139, respectively) and lowest for Black, Hispanic, and American Indian/Alaska Native students (120, 122, and 124, respectively).

The principal finding of this examination is that the trends in effects observed for this strategy in the regression analyses (Phase 1) are reflected in the results obtained when the data are disaggregated by a particular pair of student and school characteristics. This adds to the credibility of the conclusion regarding the apparent efficacy of this strategy and argues for further investigation. Moreover, the results indicate that the achievement gap in science scores cannot be explained by differences in students reading a science textbook. Thus, it is reasonable to suggest that science teachers should make some use of science textbooks in their teaching.

Table 5

Average NAEP Science Scores and Frequency of Students Who Report Reading a Science Textbook, by Eligibility for School Lunch Program and Racial/Ethnic Group, Grade 8

Reading Science Textbook				
		Percentage Eligible for School Lunch		
	Frequency	0-25% Score (%)	26-75% Score (%)	76-100% Score (%)
White	Never or hardly ever	158 (19)	146 (20)	131 (19)
	1-2 times a month	166 (20)	156 (16)	142 (13)
	1-2 times a week	169 (36)	159 (33)	146 (29)
	Almost every day	169 (26)	158 (31)	148 (38)
Black	Never or hardly ever	132 (22)	120 (20)	107 (18)
	1-2 times a month	140 (18)	124 (14)	112 (14)
	1-2 times a week	140 (35)	127 (32)	118 (33)
	Almost every day	142 (25)	127 (34)	118 (35)
Hispanic	Never or hardly ever	135 (21)	122 (22)	113 (21)
	1-2 times a month	147 (19)	128 (19)	118 (18)
	1-2 times a week	146 (38)	133 (34)	123 (36)
	Almost every day	149 (23)	134 (24)	125 (25)
Asian/Pacific Islander	Never or hardly ever	159 (13)	139 (16)	128 (12)
	1-2 times a month	168 (21)	148 (24)	134 (22)
	1-2 times a week	170 (42)	151 (39)	141 (40)
	Almost every day	172 (25)	152 (22)	145 (25)
American Indian/Alaska Native	Never or hardly ever	‡ (24)	124 (19)	115 (8)
	1-2 times a month	‡ (20)	136 (17)	106 (13)
	1-2 times a week	151 (29)	137 (31)	120 (24)
	Almost every day	150 (27)	145 (34)	114 (54)

Note: Table values represent reported average scaled score and the percentage of students in each school lunch eligibility category (in parentheses).
‡ means not reported.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Table 6

Average NAEP Science Scores and Frequency of Students Who Report Working with Others on a Science Activity or Project, by Eligibility for School Lunch Program and Racial/Ethnic Group, Grade 8

Working with Others on Science Activity or Project				
		Percentage Eligible for School Lunch		
	Frequency	0-25% Score (%)	26-75% Score (%)	76-100% Score (%)
White	Never or hardly ever	159 (14)	149 (17)	140 (20)
	Once a month to daily	168 (86)	157 (83)	145 (80)
Black	Never or hardly ever	139 (16)	121 (17)	112 (20)
	Once a month to daily	139 (84)	126 (83)	116 (80)
Hispanic	Never or hardly ever	139 (16)	126 (19)	117 (18)
	Once a month to daily	146 (84)	131 (81)	121 (82)
Asian/Pacific Islander	Never or hardly ever	166 (15)	147 (15)	134 (17)
	Once a month to daily	169 (85)	149 (85)	140 (83)
American Indian/Alaska Native	Never or hardly ever	‡(16)	130 (21)	112 (15)
	Once a month to daily	152 (84)	139 (79)	115 (85)

Note: Values represent reported average scaled score (percentages).
‡ means not reported

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Table 6 contains the results for the strategy "working with others on a science/activity or project." It has the same format as Table 5, except that here there are only two categories: "never or hardly ever" (base category) and "once a month to daily."

Table 6 shows the following patterns.

- For the majority of combinations of race/ethnicity and school disadvantage, mean scores are generally lower and sometimes substantially lower in the base category ("never or hardly ever").
- These differences are larger but in the same direction as the effect for this strategy in Figure 5.
- The proportion of students in the base category is similar for each combination of race/ethnicity and school disadvantage.

Again, the detailed analysis adds to the credibility of the apparent effectiveness of this strategy, yet cannot account for the overall differences in mean scores among racial/ethnic groups.

Table 7 displays the results for the strategy "students giving an oral science report." There are three categories: "never or hardly ever" (base category), "1-2 times a month," and "more often than 1-2 times a month."

Table 7

Average NAEP Science Scores and Frequency of Students Who Report Giving an Oral Science Report, by Eligibility for School Lunch Program and Racial/Ethnic Group, Grade 8

Giving an Oral Science Report				
		Percentage Eligible for School Lunch		
	Frequency	0-25% Score (%)	26-75% Score (%)	76-100% Score (%)
White	Never or hardly ever	167 (63)	157 (64)	146 (61)
	1-2 times a month	166 (33)	155 (31)	144 (30)
	More often than 1-2 times a month	154 (4)	142 (5)	125 (8)
Black	Never or hardly ever	143 (56)	128 (52)	118 (45)
	1-2 times a month	137 (36)	126 (37)	117 (38)
	More often than 1-2 times a month	123 (8)	112 (11)	106 (17)
Hispanic	Never or hardly ever	148 (57)	133 (57)	124 (51)
	1-2 times a month	143 (36)	129 (34)	121 (34)
	More often than 1-2 times a month	126 (7)	113 (9)	105 (14)
Asian/Pacific Islander	Never or hardly ever	171 (60)	154 (53)	145 (50)
	1-2 times a month	167 (36)	147 (38)	141 (37)
	More often than 1-2 times a month	154 (4)	129 (9)	114 (13)
American Indian/Alaska Native	Never or hardly ever	149 (61)	139 (65)	120 (39)
	1-2 times a month	152 (33)	136 (28)	115 (44)
	More often than 1-2 times a month	‡(6)	120 (7)	110 (17)

Note: Values represent reported average scaled score (percentages).
‡ means not reported

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Table 7 shows that:

- For each combination of race/ethnicity and school disadvantage, mean scores are highest in the base category ("never or hardly ever") and decrease with increasing frequency of students giving an oral report. In fact, the score deficit is the most severe for students reporting giving an oral report more often once or twice a month.
- For each level of school disadvantage, the percentage of Black students in the base (highest scoring) category is lower than for other racial/ethnic groups.
- For each racial/ethnic group, the proportions of students in the base (highest scoring) category are generally high, but trend lower in schools with greater levels of disadvantage. In addition, students in more disadvantaged schools are less likely to be exposed to the base strategy. Since Black and Hispanic students are more likely to attend schools with high proportions of disadvantaged students, this finding may account for some of the achievement gap.

- Since excessive use of this instructional strategy is more likely to be seen in more disadvantaged schools, and is associated with lower average scores, schools employing this strategy might be advised to limit or curtail it.

This detailed analysis reveals a score pattern that mimics the one seen in Figure 12, although the magnitudes of the differences are larger here. Disadvantaged minority students are somewhat more likely to report experiencing higher frequencies of this strategy, corresponding to lower NAEP science scores.

Table 8 contains the results for the strategy "teacher doing a science demonstration." It has four categories: "never or hardly ever" (base category), "1-2 times a month," "1-2 times a week," and "almost every day." Scores are highest in the category "one or two times a week" and lowest in the category "never or hardly ever." Thus students who experience this instructional pedagogy at intermediate frequencies are more likely to score higher.

There are two notable patterns in Table 8:

- For each combination of race/ethnicity and school disadvantage, mean scores are lowest in the base category ("never or hardly ever") and highest in the category "1-2 times a week."
- For each racial/ethnic group, the proportions of students in the base category (lowest scoring) are generally low but increase in schools with greater levels of disadvantage.

The mean score patterns revealed here mimic those observed in Figure 13, though again the magnitudes of the gaps are somewhat greater here. Again, too, since Black and Hispanic students are more likely to attend schools with high proportions of disadvantaged students, the overall differences in mean scores among racial/ethnic groups can be accounted for, in part, by the patterns evident in this table. Thus, schools may make progress in addressing the achievement gap by encouraging teachers to include demonstration once or twice a week.

Table 9 contains the results for the strategy "students discussing science in the news." It has three categories: "never or hardly ever" (base category), "less often than every day," and "almost every day." Higher scores are associated with the moderate use of this strategy.

Table 8

Average NAEP Science Scores and Frequency of Students Who Report Their Teacher Doing a Science Demonstration, by Eligibility for School Lunch Program and Racial/Ethnic Group, Grade 8

Teacher Doing a Science Demonstration				
		Percentage Eligible for School Lunch		
	Frequency	0-25% Score (%)	26-75% Score (%)	76-100% Score (%)
White	Never or hardly ever	157 (15)	146 (20)	134 (24)
	1-2 times a month	167 (27)	157 (28)	144 (26)
	1-2 times a week	171 (37)	161 (32)	150 (28)
	Almost every day	164 (21)	154 (21)	146 (23)
Black	Never or hardly ever	128 (16)	118 (22)	109 (24)
	1-2 times a month	141 (26)	127 (25)	116 (23)
	1-2 times a week	143 (28)	129 (26)	118 (26)
	Almost every day	140 (30)	127 (28)	118 (27)
Hispanic	Never or hardly ever	139 (15)	121 (20)	114 (21)
	1-2 times a month	146 (23)	131 (25)	121 (25)
	1-2 times a week	149 (35)	134 (30)	124 (28)
	Almost every day	142 (28)	131 (25)	122 (26)
Asian/Pacific Islander	Never or hardly ever	159 (13)	139 (16)	133 (18)
	1-2 times a month	170 (29)	150 (26)	140 (25)
	1-2 times a week	172 (35)	154 (32)	141 (31)
	Almost every day	167 (23)	148 (26)	140 (26)
American Indian/Alaska Native	Never or hardly ever	‡(16)	128 (25)	123 (12)
	1-2 times a month	‡(24)	139 (24)	114 (24)
	1-2 times a week	160 (31)	141 (30)	117 (29)
	Almost every day	153 (29)	139 (20)	113 (35)

Note: Values represent reported average scaled score (percentages).
‡ means not reported

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

Table 9

Average NAEP Science Scores and Frequency of Students Who Report Discussing Science in the News, by Eligibility for School Lunch Program and Racial/Ethnic Group, Grade 8

Student Discussing Science in the News				
		Percentage Eligible for School Lunch		
	Frequency	0-25% Score (%)	26-75% Score (%)	76-100% Score (%)
White	Never or hardly ever	162 (47)	151 (50)	141 (53)
	Less often than every day	171 (48)	161 (43)	149 (39)
	Almost every day	165 (5)	154 (6)	136 (7)
Black	Never or hardly ever	136 (56)	123 (59)	114 (56)
	Less often than every day	145 (37)	130 (33)	119 (35)
	Almost every day	131 (8)	123 (8)	110 (9)
Hispanic	Never or hardly ever	142 (57)	128 (58)	119 (56)
	Less often than every day	149 (38)	133 (37)	123 (38)
	Almost every day	143 (5)	130 (5)	115 (6)
Asian/Pacific Islander	Never or hardly ever	164 (49)	147 (53)	138 (50)
	Less often than every day	174 (47)	152 (42)	142 (45)
	Almost every day	169 (5)	149 (5)	‡(5)
American Indian/Alaska Native	Never or hardly ever	147 (55)	133 (55)	118 (46)
	Less often than every day	152 (40)	143 (39)	116 (47)
	Almost every day	‡(5)	141 (6)	‡(6)

Note: Values represent reported average scaled score (percentages).

‡ means not reported

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

In summarizing the data patterns in Table 9:

- For nearly all combinations of race/ethnicity and school disadvantage, the average scores for students reporting the middle category "less often than every day" are higher than those for students reporting either of the other two categories.
- The gaps in average scores typically fall in the range of 5 – 10 points.
- In most cases, the base category is the modal category.
- Black and Hispanic students were less likely to be exposed to the optimal level of this strategy.

The mean score patterns revealed here mimic those observed in Figure 13, though, again, the magnitudes of the gaps are somewhat greater here. The proportions of White students in the middle (optimal) category are somewhat larger than the proportions for Black students and Hispanic students. Consequently, the overall differences in mean scores among racial/ethnic groups may be due, in part, to differences in the frequency of exposure to this strategy.

We undertook this analysis to identify and examine the characteristics — school, student, teacher, and instructional — of the eighth-grade science classroom that are statistically associated with performance on the NAEP science assessment. Indeed, a number of strong, interesting patterns were discovered. But what is one to make of these findings? On the one hand, NAEP is an unrivaled source of data; it presents results from a large, nationally representative sample of students. On the other hand, information on instructional practices is derived from student responses, which can be affected by misunderstanding of the questions, recall bias, and so on. In addition, limitations of the administrative context preclude obtaining more fine-grained information relating to the nature and quality of the instructional practices. Finally, NAEP is an observational study, so that statistical results cannot blithely be interpreted causally. Notwithstanding these caveats, the consistency across the different analyses conducted in this study suggests that these findings should be taken seriously and could serve as the basis for further research studies.

Phase 1 examined the relationships among student demographic, background, and home environment characteristics and teacher instructional practices on the one hand, and NAEP science scores on the other. This phase of the analysis identified instructional practices that were associated with differences in NAEP science scores after differences among students with respect to demographic characteristics and home environment, as well as teacher characteristics, have been taken into account, or controlled for.

To investigate the possibility that controlling for those background characteristics might underestimate the relationships found in Phase 1, Phase 2 was conducted using models that employed teacher characteristics and instructional pedagogies without controlling for other variables.[20] Results of Phase 2 indicated that the Phase 1 findings with regard to the pattern of associations between instructional strategies and NAEP science achievement were not materially affected by prior adjustment for student demographic and home environment variables. This provides further support for the hypothesis that teachers' instructional practices do make a difference in students' success on the NAEP science assessment.

We conducted Phase 3 to examine how the apparently effective and ineffective pedagogies identified in the multivariate analyses are distributed across students when the data are disaggregated by racial/ethnic group and school disadvantage categories. We designed this phase to answer two questions: (1) Are the relationships between instructional strategies and NAEP science scores found in the aggregate analyses replicated when viewed at the subgroup level? And (2) Do different groups of students have differential access to effective (or ineffective) instructional strategies?

The Phase 1 analyses revealed that student demographic characteristics had statistically significant associations with achievement. Black and Hispanic students scored considerably lower than White students, and males scored higher than females. English-language learners and students with disabilities scored much lower than other students. Analysis of students' home characteristics revealed that students with many books in the home scored considerably higher than students with fewer books; and students who were absent frequently scored much lower than other students. Finally, students whose teachers held a standard certificate scored slightly higher than other students and students whose teachers' years of total experience exceeded their years of science experience scored slightly lower than other students.

From the Phase 1 and Phase 2 analyses, we identified three subgroups of teacher instructional practices with distinct patterns of association between their reported frequencies and NAEP science achievement. One group comprises strategies for which increasing frequency is associated with higher average NAEP science scores:

- Using a science textbook
- Doing hands-on activities in science
- Writing long answers to science tests and assignments
- Talking about measurements and results from hands-on activities
- Working with others on a science activity or project

[20] As explained earlier, under- or overestimation can occur as a result of selection bias that is not captured by the variables incorporated into the model.

A second group comprises instructional activities for which increasing frequency is associated with lower average scores:

- Students giving an oral science report
- Students using library resources for science

A final group is characterized by the "Goldilocks" analogy. For these instructional strategies, higher scores are associated with moderate frequency:

- Students taking a science test
- Teachers doing a science demonstration
- Students discussing science in the news
- Students reading a book or magazine about science
- Students preparing a written science report

For the Phase 3 analyses, we chose five instructional strategies from the multilevel analyses. For each one, students are cross-classified by racial/ethnic group and the percentage of students in school who are eligible for the school lunch program (a measure of school disadvantage).

- **Reading a science textbook** – Across all racial/ethnic and school disadvantage groups, scores increase with the frequency of reading a science textbook. The percentage of students in the optimal categories is similar for each combination of race/ethnicity and school disadvantage. Thus, the analysis suggests that mean score differences among racial/ethnic groups cannot be accounted for by differences in exposure to this instructional strategy. These data suggest that it is reasonable to recommend that science teachers should make some use of science textbooks in their teaching.
- **Working with others on a science project** – Across all racial/ethnic and school disadvantaged groups, scores increase with the frequency of working with others on a science project. Thus, the analysis also indicates that mean score differences among racial/ethnic groups cannot be accounted for by differences in exposure to this instructional strategy. A reasonable inference from these data is that science teachers make use of group work on science projects.
- **Students giving an oral science report** – Across all racial/ethnic and school disadvantage groups, scores decrease with the frequency of students giving an oral science report. The score deficit is most severe for students reporting giving an oral report more often than once or twice a month. Black and Hispanic students and students attending more disadvantaged schools were more likely to experience higher frequencies of this strategy. Since excessive use of this strategy is more likely to be seen in more disadvantaged schools, curtailing this practice may help to close the achievement gap.
- **Teacher doing a science demonstration** – Across all racial/ethnic and school disadvantage groups, scores are lowest in the "never or hardly ever" category and highest in the category of "one or two times a week." For all racial/ethnic groups, students attending schools with high levels of disadvantage are less likely to be exposed to the optimal use of this strategy (one or two times a week). Thus, schools may make progress in closing the achievement gap by encouraging teachers to employ this strategy to a moderate degree.
- **Discussing science in the news** – Across all racial/ethnic and school disadvantage groups, scores are highest in the middle response category, "less often than every day." Black and Hispanic students are less likely than other students to fall into that category, suggesting that the achievement gap may be due, in part, to differences in the frequency of exposure to this instructional strategy.

In this report, we have provided a view of the eighth-grade science classroom in 2005 using HLMs, which are well-suited to this type of data. Our intent was to identify teacher characteristics and pedagogies that are strongly related to student science achievement. We conducted three sets of analyses that yielded similar findings with regard to the relationships between instructional practices and NAEP eighth-grade science scores. With due regard to the usual cautions attendant upon drawing policy conclusions from the analysis of cross-sectional data, the consistency of the findings speaks to the plausibility of the results. Many of the findings are in line with the predictions one would make based on the arguments found in the National Academy of Sciences report *Taking Science to School.* Other findings are somewhat puzzling and call for further investigation, not least because of the methodological issues to which we previously alluded. In any case, more detailed information on how these practices are implemented in classrooms and information about the contexts in which they are employed should be gathered to provide further evidence with regard to the efficacy of these strategies.

The 2005 NAEP grade 8 science assessment contains records for 148,595 students from more than 6,300 schools. A total of 5,183 students who could not be reported because of the nature or severity of their disability were first excluded. The analysis sample for HLM was obtained by excluding students sequentially from the full reporting sample (143,412) as indicated below:

a) Students from International Department of Defense Schools (n = 1,473).

b) Students missing the "teacher match code" variable (n = 27,611).

c) Students with missing gender (n = 73).

d) Students whose answers to the question "Which best describes the science course you are taking?" are "I am not taking a science course this year," multiple or omitted (n = 5,283).

e) Students whose teacher holds a degree below the bachelor's level (i.e., their response to "What is the highest academic degree the teacher holds?" was "High-school diploma" or "Associate's degree/voc certificate"); multiple or omitted (n = 1,039).

A total of 107,933 students were included in the reduced sample for regression analyses. Table A.1 shows the percentages of students by their demographic characteristics using the reduced sample, along with comparisons to the percentages reported in the NAEP Data Explorer (NDE) (http://nces.ed.gov/nationsreportcard/nde/).

Table A.1

Percentage of Students by Demographic Characteristics, NAEP Grade 8 Science, 2005

Student Demographic Characteristics		Percentage in Reduced Sample	Percentage in NDE
Gender			
	Male	50%	50%
	Female	50%	50%
Ethnicity			
	White	63%	61%
	Black	16%	17%
	Hispanic	14%	16%
	Asian/Pacific Islander American/Pacific Islander	4%	4%
	American Indian/Alaska Native	1%	1%
	Unclassified	1%	1%
English-language learner (ELL)			
	ELL	4%	5%
	Not ELL	96%	95%
School lunch program eligibility			
	Eligible	35%	37%
	Not eligible	58%	55%
	Info. not available	7%	8%
Parental education level			
	Did not finish H.S.	7%	7%
	Graduated H.S.	18%	18%
	Some ed. after H.S.	18%	17%
	Graduated college	49%	48%
	Unknown	10%	11%
Title 1 participation			
	Title 1	25%	N/A
	Not Title 1	75%	N/A
Student classified as having a disability			
	SD	9%	10%
	Not SD	91%	90%

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Science Assessment.

There are a number of approaches to conducting a multilevel analysis of a large data set with numerous potential explanatory characteristics. The approach adopted for this report follows the one implemented in a study of charter schools.[21] It is a variant of what is often referred to as stagewise regression.

The multilevel or hierarchical regression analysis was implemented using the program HLM6-PV that is designed to accommodate the special features of the NAEP database.[22] To make these ideas more concrete, consider the following model:

Level 1: $y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \ldots + \beta_{Pj} X_{Pij} + e_{ij}$ (1)

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01} W_{1j} + \ldots + \gamma_{Q1} W_{Qj} + u_{0j}$

$$\beta_{1j} = \gamma_{10}$$
$$\vdots$$
$$\beta_{Pj} = \gamma_{P0}$$

where i indexes students within schools, j indexes schools;

y_{ij} is the outcome for student i in school j;

$X_1, \ldots, X_P$ are P student- and teacher-related covariates, centered at their grand means, and indexed by i and j as above;

β_{0j} is the mean for school j, adjusted for the covariates $X_1, \ldots, X_P$;

$\beta_{1j}, \ldots, \beta_{Pj}$ are the regression coefficients for school j, associated with the covariates $X_1, \ldots, X_P$;

e_{ij} is the random error (i.e., residual term) in the level 1 equation, assumed to be independently and normally distributed with mean zero and a common variance σ^2 for all students;

$W_{1j}, \ldots, W_{Qj}$ are school-related covariates;

γ_{00} is the intercept for the regression of the adjusted school mean on school characteristics;

$\gamma_{01}, \ldots, \gamma_{Q1}$ are the regression coefficients associated with the school level covariates $W_{1j}, \ldots, W_{Qj}$;

u_{0j} is the random error in the level 2 equation, assumed to be independently and normally distributed across schools with mean zero and variance τ^2;

and $\gamma_{10}, \ldots, \gamma_{P0}$ are constants denoting the common values of the P regression coefficients across schools. For example, γ_{10} is the common regression coefficient associated with the first covariate in the level 1 model for each school.

In the level 1 equation, HLM estimates an adjusted mean for each school. In the level 2 equation, these adjusted means are, in turn, regressed on the school-level covariates. The level 1 regression intercept β_{0j} is assumed to be a random variable that is regressed on the level 2 indicators, while the level 1 regression coefficients $\beta_{1j}, \ldots, \beta_{Pj}$ are all assumed to be common values (non-stochastic). Therefore, the estimated values of $\beta_{1j}, \ldots, \beta_{Pj}$ will be fairly insensitive to the inclusion of any level 2 covariates.

It should be noted that the criterion or dependent variables in these analyses are the plausible values generated during NAEP operational analyses. The regression models that are used to create the plausible values do not include predictor variables that directly reflect the nested structure of the data, as do the HLMs. Consequently, there may be some bias in the estimates of the HLM parameters. However, the bias should be quite small, given the extensive set of predictors that are used in NAEP operations.

Determining appropriate weights to be employed in an HLM analysis is a complex matter. The general recommendation is to apply school weights to schools, and the student-within-school weights to students.[23] Since the student-within-school weights are fairly constant within each school, only school weights were incorporated at level 2 of the models in the Phase 1 and 2 analyses.

[21] Henry Braun, Frank Jenkins, and Wendy Grigg, *A Closer Look at Charter Schools Using Hierarchical Linear Modeling* (NCES 2006-460), U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, D.C.: U.S. Government Printing Office, 2006.

[22] Stephen Raudenbush et al., *HLM6: Hierarchical Linear and Nonlinear Modeling*, Lincolnwood, IL: Scientific Software International, 2004. See also, Stephen Raudenbush and Anthony S. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods (2nd Ed.)*, Thousand Oaks, CA: Sage Publications, 2002.

[23] Danny Pfefferman, Chris J. Skinner, David J. Holmes, Harvey Goldstein, and Jon R. Rasbash, "Weighting for Unequal Selection Probabilities in Multilevel Models," *Journal of the Royal Statistical Society, Series B,* 60 (1), pp. 23 – 40, 1998.

For this study, the characteristics are first organized into a small number of categories, each containing substantively related characteristics. In this study, there are five such categories, described in the main body of the report. Student attitude questions, such as attitudes toward science, were excluded from the analysis. Before conducting the multilevel analysis, a systematic sequence of exploratory unweighted regression models was first fitted using the General Linear Models (GLM) methodology. The goal was to select a set of candidate covariates for the more intensive HLM analysis. SAS GLMSELECT procedure was used, with variable selection based on stepwise regression and the Schwarz Bayesian Information (SBC) criterion.

Variables emerging from the GLM analysis are entered into the regression by category, according to a predetermined order, based on both statistical considerations and interpretive goals. At each stage, the set of variables corresponding to a particular characteristic is retained only if the regression coefficients associated with the variables in the set generally exceed a predefined statistical threshold (i.e., significant at the 0.05 level). At the succeeding stage, the variables remaining from the previous stage together with all the variables in the next category are entered into the regression. The process continues until the last category has been entered.

As indicated earlier, categories of characteristics were introduced in a particular order. For the Phase 1 analysis, a model with no explanatory characteristics was run at the first stage in order to obtain a preliminary decomposition of score variance into within-school and between-school components. The full set[24] of student characteristics was then entered at the second stage of analysis. All but one of the regression coefficients were statistically significant and so all characteristics were retained for the next stage. At the third stage, the full set of home environment characteristics was entered as student-level covariates. Based on the sizes of the regression coefficients and their associated p-values, seven characteristics were retained for the next stage. At the fourth stage, all teacher pedagogy characteristics were entered at the student level and 13 were retained for further analysis. At the fifth stage, all the teacher characteristics were entered into the student level and only two characteristics were retained for the further analysis. At the last stage, the full set of school characteristics was entered into the school level of the two-level model, and three were retained in the model.

As a result of such model-fitting sequence in Phase 1, NAEP science scores are first adjusted for both student demographics and student home environment before the set of teacher pedagogy variables are introduced. Consequently, the estimated regression coefficients of the teacher pedagogy variables represent the strength of the relationship between NAEP science scores and teacher pedagogy after taking account of (some of the) differences among students that are not influenced by the teacher. Finally, teacher characteristics and school characteristics are entered into the model to provide further insight.

Ordinarily, the introduction of a new variable into a regression model results in changes in the regression coefficients of the variables already present. (The changes are due to the correlations between the new variable and the old ones.) That is certainly the case in this study. The observed changes, however, are rather modest. For example, when the home environment characteristics are introduced in stage 3, the coefficients of the student demographic variables are slightly reduced in magnitude but the signs and patterns are unaffected. This holds true for both demographic characteristics and home environment characteristics at stage 4, when teacher pedagogy characteristics are entered into the model. Accordingly, the discussion in the text focuses on the results obtained after the conclusion of stage 6, when all five categories of characteristics have been entered and those with significant regression coefficients retained.

Each HLM model estimated in Phase 1 also yields a decomposition of the total variance of NAEP scores into the fraction attributable to the differences among students within schools and the fraction attributable to differences among schools. These variance decompositions are a convenient summary indicator of the success of the model in accounting for the heterogeneity in student outcomes (NAEP science scores). Table B.1 presents the variance decompositions corresponding to the model sequence described previously. The fully unconditional model on the first row of the table yields the basic decomposition. The total variance is simply the sum of the two displayed components: 966=601+365. That is, about 62 percent of the total variance (601/966) is attributable

[24] Refer to Tables 2 and 3 for the full set of variables entered into the models at each stage.

to within-school heterogeneity, and about 38 percent (365/966) is attributable to between-school heterogeneity. For the succeeding models, the numbers in the fourth and sixth columns of the table represent the percentage reduction in the residual variance at the corresponding level of the model, treating the residual variance in the fully unconditional model as the baseline.

Note that including covariates at Level 1 yields a greater reduction in the between-school variance than in the within-school variance. This is a common finding in such analyses and is explained by the fact that there is greater student homogeneity within schools than between schools. Indeed, 80 percent of the variance between school means is accounted for by measured differences among students with respect to demographics and home environment. The inclusion of teacher pedagogy covariates results in a further reduction of 6 percent in within-school variance and of 4 percent in between-school variance. When viewed, instead, in terms of the reduction in the residual variance of the previous model, the corresponding percentages are 8 and 20. These are substantial and lend support to the hypothesis that teacher instructional practices do have an impact on student science achievement.

Table B.1

Variance Decomposition for NAEP Science Scale Scores in Phase 1 Analysis, NAEP Grade 8, 2005

Model		Between Students, Within Schools		Between Schools	
Level 1 Covariates	Level 2 Covariates	Variance	Percent of Variance in Baseline Model Accounted for	Variance	Percent of Variance in Baseline Model Accounted for
None	None	601	†	365	†
Student demographics	None	473	21	90	75
Student demographics †Home environment	None	438	27	72	80
Student demographics †Home environment †Teacher pedagogy	None	404	33	57	84
Student demographics †Home environment †Teacher pedagogy †Teacher background	None	404	33	55	85
Student demographics †Home environment †Teacher pedagogy †Teacher background	School characteristics	404	33	50	86

†Not applicable.

Data for Figure 3

Estimated Regression Coefficients for Student Demographic Characteristics, Level 1 Variables, NAEP Grade 8 Science, 2005

Student Demographic Characteristics	Regression Coefficient	p Value
Gender (male)	7.86 (0.40)	.00
Race/Ethnicity[1]		
Black	-18.17 (0.66)	.00
Hispanic	-7.12 (0.62)	.00
Asian/Pacific Islander	-0.29 (0.89)	.75
American Indian/Alaska Native	-9.12 (2.07)	.00
English-language learner	-18.73 (0.99)	.00
Eligible for free/reduced-price school lunch	-4.23 (0.55)	.00
Parents had bachelor's or higher degree	6.05 (0.59)	.00
Title 1 participation	-4.15 (0.82)	.00
Disability (IEP)	-22.69 (0.66)	.00

[1] The reference group is White.
Note: Standard errors of the estimates appear in parentheses.

Data for Figure 4

Estimated Regression Coefficients for Student Home Environment Characteristics, Level 1 Variables, NAEP Grade 8 Science, 2005

Student Home Environment Characteristics	Regression Coefficient	p Value
Books at home[1]		
26-100 books	6.26 (0.42)	.00
More than 100 books	13.29 (0.48)	.00
More than 10 pages read for homework	2.51 (0.40)	.00
Computer at home	2.22 (0.68)	.00
Encyclopedia at home	2.03 (0.54)	.00
Talking about things studied in school with family at least once a week	1.65 (0.44)	.00
Magazines at home	1.13 (0.43)	.01
Days of absence per month[2]		
1-2 days	-1.30 (0.43)	.01
3-4 days	-4.20 (0.60)	.00
5-10 days	-5.62 (0.75)	.00
More than 10 days	-11.44 (1.11)	.00

[1] The reference category is 0-25 books.
[2] The reference category is "none."
Note: Standard errors of the estimates appear in parentheses.

Data for Figure 5

Estimated Regression Coefficients for Teacher Pedagogical Strategies, Level 1 Variables, NAEP Grade 8 Science, 2005

Teacher Pedagogical Strategies	Regression Coefficient	p Value
Student reading a science textbook[1]		
Once or twice a month	3.08 (0.74)	.00
Once or twice a week	4.72 (0.64)	.00
Almost every day	6.23 (0.79)	.00
Student doing hands-on activities in science	2.72 (0.65)	.00
Student doing hands-on with magnifying glass/microscope	3.02 (0.48)	.00
Student doing hands-on with chemicals	2.86 (0.50)	.00
Student doing hands-on with barometer	1.86 (0.41)	.00
Student doing hands-on with simple machines	1.74 (0.40)	.00
Student writing long answers to science tests/assignments	3.17 (0.55)	.00
Student talking about the measurements and results from hands-on activities or investigations	1.77 (0.59)	.00
Student working with others on a science activity/project	1.61 (0.50)	.00
School resources for teachers[2]		
Most of the needed resources	0.56 (0.47)	.23
All the needed resources	1.62 (0.70)	.03

[1] The reference category is "never or hardly ever."
[2] The reference category is "teacher not getting any of the resources needed."
Note: Standard errors of the estimates appear in parentheses.

Data for Figure 6

Estimated Regression Coefficients for Teacher Pedagogical Strategies, Level 1 Variables, NAEP Grade 8 Science, 2005

Teacher Pedagogical Strategies	Regression Coefficient	p Value
Student giving an oral science report[1]		
Once or twice a month	-3.92 (0.45)	.00
More often than once or twice a month	-9.46 (0.85)	.00
Student using library resources for science[1]		
Once or twice a month	-2.45 (0.47)	.00
More often than once or twice a month	-6.72 (0.59)	.00

[1] The reference category is "never or hardly ever."
Note: Standard errors of the estimates appear in parentheses.

Data for Figure 7

Estimated Regression Coefficients for Teacher Pedagogical Strategies, Level 1 Variables, NAEP Grade 8 Science, 2005

Teacher Pedagogical Strategies	Regression Coefficient	p Value
Student taking a science test[1]		
Once or twice a month	8.33 (1.51)	.00
Once or twice a week	6.05 (1.54)	.00
Almost every day	0.21 (1.48)	.89
Teacher doing a science demonstration[1]		
Once or twice a month	3.54 (0.55)	.00
Once or twice a week	4.45 (0.55)	.00
Almost every day	1.87 (0.62)	.00
Student discussing science in the news[1]		
Less often than every day	3.33 (0.39)	.00
Almost every day	0.51 (0.88)	.56
Student reading a book or a magazine about science[1]		
Once or twice a month	1.65 (0.49)	.00
More often than once or twice a month	-1.28 (0.73)	.09
Student preparing a written science report[1]		
Once or twice a month	1.01 (0.35)	.01
More often than once or twice a month	-2.91 (0.60)	.00

[1] The reference category is "never or hardly ever."
Note: Standard errors of the estimates appear in parentheses.

Data for Figure 8

Estimated Regression Coefficients for Teacher Characteristics, Level 1 Variables, NAEP Grade 8 Science, 2005

Teacher Characteristics	Regression Coefficient	p Value
Teacher holding standard teaching certificate	1.49 (0.59)	.01
Years of teaching experience greater than years of science teaching experience	-1.35 (0.45)	.00

Note: Standard errors of the estimates appear in parentheses.

Data for Figure 9

Estimated Regression Coefficients for School Characteristics, Level 2 School Variables, NAEP Grade 8 Science, 2005

School Characteristics	Regression Coefficient	p Value
Percentage of minority students (Black and Hispanic)	-0.06 (0.01)	.00
Percentage eligible for school lunch program[1]		
26-75%	-1.98 (0.57)	.00
76-100%	-3.75 (1.44)	.01
Region of the country[2]		
Midwest	-0.46 (0.94)	.63
South	-0.17 (0.85)	.84
West	-3.22 (0.91)	.00

[1] The reference group is 0-25%.
[2] The reference region is Northeast.
Note: Standard errors of the estimates appear in parentheses.

Data for Figure 10

Estimated Regression Coefficients for Teacher Characteristics, Level 1 Variables, NAEP Grade 8 Science, 2005

Teacher Characteristics	Regression Coefficient	p Value
Teacher holding standard teaching certificate	1.79 (0.80)	.03
Years of teaching experience greater than years of science teaching experience	-3.84 (0.77)	.00

Note: Standard errors of the estimates appear in parentheses.

Data for Figure 11

Estimated Regression Coefficients for Teacher Pedagogical Strategies, Level 1 Variables, NAEP Grade 8 Science, 2005

Teacher Pedagogical Strategies	Regression Coefficient	p Value
Student reading a science textbook[1]		
Once or twice a month	4.65 (0.80)	.00
Once or twice a week	7.83 (0.71)	.00
Almost every day	10.38 (0.87)	.00
Student doing hands-on activities in science	4.33 (0.69)	.00
Student doing hands-on with magnifying glass/microscope	4.81 (0.56)	.00
Student doing hands-on with chemicals	3.98 (0.56)	.00
Student doing hands-on with simple machines	3.48 (0.46)	.00
Student doing hands-on with barometer	2.21 (0.44)	.00
Student writing long answers to science tests/assignments	5.86 (0.66)	.00
Student talking about the measurements and results from hands-on activities or investigations	2.83 (0.61)	.00
Student working with others on a science activity/project	1.69 (0.55)	.00
School resources for teachers[2]		
Most of the needed resources	2.56 (0.71)	.00
All the needed resources	4.52 (1.06)	.00

[1] The reference category is "never or hardly ever."
[2] The reference category is "teacher not getting any of the resources needed."
Note: Standard errors of the estimates appear in parentheses.

Data for Figure 12

Estimated Regression Coefficients for Teacher Pedagogical Strategies, Level 1 Covariates, NAEP Grade 8 Science, 2005

Teacher Pedagogical Strategies	Regression Coefficient	p Value
Student giving an oral science report[1]		
Once or twice a month	-5.50 (0.52)	.00
More often than once or twice a month	-13.47 (1.08)	.00
Student using library resources for science[1]		
Once or twice a month	-2.58 (0.48)	.00
More often than once or twice a month	-8.83 (0.62)	.00

[1] The reference category is "never or hardly ever."
Note: Standard errors of the estimates appear in parentheses.

Data for Figure 13

Estimated Regression Coefficients for Teacher Pedagogical Strategies, Level 1 Covariates, NAEP Grade 8 Science, 2005

Teacher Pedagogical Strategies	Regression Coefficient	p Value
Student taking a science test[1]		
Once or twice a month	12.10 (1.65)	.00
Once or twice a week	8.39 (1.64)	.00
Almost every day	0.60 (1.70)	.73
Teacher doing a science demonstration[1]		
Once or twice a month	4.17 (0.61)	.00
Once or twice a week	5.29 (0.70)	.00
Almost every day	1.30 (0.75)	.09
Student discussing science in the news[1]		
Less often than every day	6.10 (0.45)	.00
Almost every day	2.49 (0.94)	.01
Student reading a book or a magazine about science[1]		
Once or twice a month	3.13 (0.49)	.00
More often than once or twice a month	-1.86 (0.76)	.02
Student preparing a written science report[1]		
Once or twice a month	1.28 (0.39)	.00
More often than once or twice a month	-4.32 (0.75)	.00

[1] The reference category is "never or hardly ever."
Note: Standard errors of the estimates appear in parentheses.

About ETS

At nonprofit ETS, we advance quality and equity in education for people worldwide by creating assessments based on rigorous research. ETS serves individuals, educational institutions and government agencies by providing customized solutions for teacher certification, English-language learning, and elementary, secondary and post-secondary education, as well as conducting education research, analysis and policy studies. Founded in 1947, ETS develops, administers and scores more than 50 million tests annually — including the TOEFL® and TOEIC® tests, the GRE® test and *The Praxis Series*™ assessments — in more than 180 countries, at over 9,000 locations worldwide.

Listening. Learning. Leading.®

www.ets.org