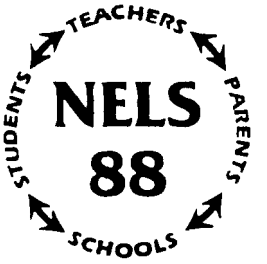
NATIONAL CENTER FOR EDUCATION STATISTICS

Statistical Analysis Report **May 1997**

National Education Longitudinal Study of 1988
Second Followup

Constructed Response Tests in the NELS:88 High School Effectiveness Study



National Opinion Research Center (NORC)
at the University of Chicago

Judith M. Pollock
Donald A. Rock
Educational Testing Service

Peggy Quinn
Project Officer
National Center for Education Statistics

U.S. Department of Education
Office of Educational Research and Improvement **NCES 97-804**

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Ramon C. Cortines
Acting Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policy makers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue NW
Washington, DC 20208–5574

May 1997

The NCES World Wide Web Home Page address is
http://www.ed.gov/NCES/

Suggested Citation

U.S. Department of Education. National Center for Education Statistics. *[National Education Longitudinal Study of 1988 Second Followup] Constructed Response Tests in the NELS:88 High School Effectiveness Study,* NCES 97-804, by J.M. Pollock and D.A. Rock. Project Officer: P. Quinn. Washington, DC: 1997.

contact

Peggy Quinn
(202) 219-1743

For free single copies of this publication, call the National Data Resource Center at **(703) 845–3151** or send a FAX request to **(703) 820–7465.**

Table of Contents

List of Tables

Page

List of Appendices

Acknowledgments

The development and scoring of constructed response tests is not a trivial undertaking. It requires the cooperation and dedication of a large number of talented individuals. We begin by thanking Jeffrey Owings and Peggy Quinn of the National Center for Education Statistics (NCES) and Anne Hafner (formerly of NCES) for their ideas, encouragement, and support. We are indebted to Larry Suter and the National Science Foundation (NSF) for sponsoring the High School Effectiveness Study (HSES) constructed response testing component of the National Education Longitudinal Study of 1988 (NELS:88). Steven J. Ingels, the NELS:88 Project Director, and Leslie A. Scott, both of the National Opinion Research Center (NORC) at the University of Chicago, provided guidance and support in the development and administration of the tests. All of these people provided the administrative environment that facilitated the completion of this complex undertaking. Thanks are also due to the many reviewers who provided helpful suggestions for revising this report: Bob Burton, Peggy Quinn, Marilyn McMillan, and Jeffrey Owings of NCES; Mary Naifeh of the Census Bureau; Steven Ingels and Leslie Scott of NORC; Richard Duran of University of California, Santa Barbara; Thomas Romberg of University of Wisconsin-Madison; Richard Snow of Stanford University; and David Burkham of University of Michigan.

Many people at Educational Testing Service shared their experience in constructed response testing with us and contributed to the development and scoring of the test items. We would especially like to thank Beth Brownstein, for her painstaking attention to every detail of the mathematics questions; Mary Gribben, who saw the science tests through to the end of the scoring and scaling procedures, long after it ceased to be her responsibility to do so; Dan Richman and Dick Devore for their contributions to development of the mathematics and science tests; Trudy Conlan, for generously sharing her tremendous experience and creativity; and Kalle Gerritz for her level-headed advice at all stages of the project.

Finally, we are grateful to the people whose cooperation in providing the data made this investigation possible: all of the students who voluntarily participated in the study, and the mathematics and science teachers who took time out from their summer vacations to score the tests.

Chapter 1: Introduction and Background

This report will describe an experiment in constructed response testing undertaken in conjunction with the National Education Longitudinal Study of 1988 (NELS:88). The term "constructed response" is used to describe test questions that require students to produce their responses themselves rather than to select the correct answer from several response options. Participants in this experiment took constructed response tests in mathematics or science, along with a battery of traditional multiple choice tests. Data on students' background and school experiences were also collected. The experiment was designed to explore the practical and psychometric issues involved in using constructed response test formats in the context of a large-scale, voluntary national survey.

We will begin with a brief description of the purpose and structure of the NELS:88 survey, and of its multiple choice test battery that measured gains in cognitive achievement during the high school years. The idea of incorporating a constructed response component in the NELS:88 test battery ultimately led to the experiment documented in this report. The High School Effectiveness Study, which is described below, provided the opportunity for collecting constructed response test data in conjunction with NELS:88 activities.

Later chapters will report on the objectives and issues involved in the development of the constructed response tests, and on the steps taken to address these issues. The scoring procedures and treatment of missing data will be described. Findings from analysis of the test data will be presented, including psychometric characteristics of the tests, response rates, performance differences for ethnic and gender groups, and comparisons with multiple choice test results.

The report concludes with a summary of the major issues and results, and with a description of the data file that will be made available to researchers wishing to conduct further investigations.

The National Education Longitudinal Study of 1988 (NELS:88)

The National Education Longitudinal Study of 1988 (NELS:88) is the third in a series of longitudinal studies sponsored by the National Center for Education Statistics (NCES). The first of these, the National Longitudinal Study of the High School Class of 1972 (NLS-72), began with high school seniors, while the second, the High School and Beyond (HS&B) study of 1980, started with both tenth and twelfth grade cohorts. The data collected from the students and from their teachers, schools, and parents provide policy-relevant information about student achievement, and about learning-related student experiences and attitudes.

NELS:88 is more comprehensive than the earlier longitudinal studies in the amount and type of data collected, as well as in the time period spanned by the data collection. NELS:88 began with a nationally representative core sample of eighth graders in 1,052 schools in the spring of 1988 and followed them through their high school years. The same students were followed and tested two and four years later. Students who remained on a normal sequence would have been in tenth and twelfth grades at the later testing times; however, dropouts, early graduates, and grade-retained students were also followed and tested. Adjustments were made to the sampling design in the followup years so that national estimates could be made for a cross-section of tenth and twelfth graders in the later years, as well as for a panel sample of eighth graders two and four years later.

Multiple choice tests in reading, mathematics, science and history/citizenship/geography were administered to NELS:88 participants in 1988, 1990, and 1992. The test scores were designed to provide researchers with longitudinal measures of gains in achievement over the four year time span that could be related to student background characteristics, curriculum exposure, out-of-school experiences, and other variables that were measured by survey questionnaires and school records.

During preparations for the final (1992) round of tests, the NELS:88 Technical Review Panel suggested the possibility of incorporating a constructed response component into the test battery. An objective of constructed response testing is to measure skills that cannot easily be assessed in multiple choice format. Constructed response questions, in which the student must solve a problem, write an explanation, draw a diagram, etc., require that the answer come entirely from the student's own knowledge and experience. There is no possibility of one of the options in a set of response choices providing a hint of the correct answer, or conversely, of a student being cued that his or her response is not correct by not finding it as one of the choices. Multiple choice format cannot easily give detailed information about the types of errors or misconceptions that led to an incorrect final answer; nor does it allow for the possibility of a test taker coming up with a different correct answer not envisioned by the test writer. Both of these are possible in constructed response format.

Replacing one or more of the NELS:88 multiple choice subject area tests with a constructed response test was not feasible. Tests with radically different formats and no overlap of test items could not be put on the same scale; thus longitudinal measurement of gains in achievement over time would be impossible. It was decided instead to preserve the structure of the core NELS:88 test battery, and to supplement it with a methodological experiment in constructed response testing. The information gathered in such an experiment could be used to inform future choices of test format with respect to issues such as content, difficulty, bias, omit rates, reliability, and costs.

The High School Effectiveness Study (HSES)

After the eighth grade base year, NELS:88 participants dispersed to a large number of high schools. This made analysis of school effects problematic for two reasons. First, the number of NELS:88 students within each school tended to be small, averaging 14 students per school in the 1990 first follow-up compared to approximately 24 students per school in the base year of NELS:88 and 30 students per school in the base year of HS&B. Second, the cluster of NELS:88 students in a high school could be expected to be unrepresentative of the population of the school, since the NELS:88 group typically had come from only one of many feeder schools represented in the high school population. The low numbers of students per school and the unrepresentative nature of the clusters did not permit school effects analyses or the use of hierarchical linear modeling techniques, which would normally be used to assess the effects of school policies and practices on students.

To compensate for this limitation, a probability subsample of 247 urban and suburban NELS:88 first followup schools in the thirty largest Metropolitan Statistical Areas (MSAs) were designated as High School Effectiveness Study (HSES) schools. In these schools, the NELS:88 national or "core" student sample was augmented to obtain a within-school representative student sample large enough to support school effects research. In HSES schools, the NELS:88 student sample was increased by 15 students on average to obtain within-school student cluster sizes of approximately 30 students. These schools and students were followed up again in 1992 as part of both the NELS:88 national survey and HSES survey, when the majority of the students were in twelfth grade.

The High School Effectiveness Study provided a convenient framework for a constructed response testing experiment in 1992. The full complement of NELS:88 core survey components were already being collected in the HSES schools: student questionnaires, multiple choice cognitive tests, parent, teacher and school questionnaires, and transcript records. Half of the HSES schools that agreed to commit the extra time required for students to take a four-question constructed response test were assigned to mathematics; in the other half of the schools, constructed response science tests were given.

Chapter 2: Constructed Response Field Test

This chapter will describe the field test activities undertaken in 1991 to determine the feasibility and costs of including a constructed response test component in the 1992 High School Effectiveness Study. Test formats, scoring procedures, and findings from the field test are reported.

One year prior to each of the three NELS:88 survey years (1988, 1990, and 1992), field tests were conducted that included multiple choice test items in reading comprehension, mathematics, science, and history/citizenship/geography. The objective was to develop and evaluate pools of items from which the final forms of the tests could be selected for the main survey years, so more items needed to be field tested than would eventually be chosen. In the 1991 field test, constructed response items in all subjects except history/citizenship/geography were tried out as well. Topics for the items were suggested by the NELS:88 Technical Review Panel and/or adapted from other sources. The results of this field test guided the selection and development of items for the second followup HSES constructed response tests.

With limited testing tune available, it was not possible to field test all subject areas for all students. Five different test booklets were assembled, each containing four constructed response questions in one subject area, along with multiple choice questions in the same subject. Mathematics and science questions each appeared in two booklets, and reading comprehension questions in one. Each of the constructed response questions was followed by several student reaction questions, asking for students' perceptions of the difficulty, timing, and clarity of that question, as well as whether they had given the best answer they could. Each of the five booklets was administered to about 400 students. Constructed response items were scored by a team of readers, most of whom were high school teachers, who were trained to apply a uniform set of criteria in evaluating the answers. The readers not only scored the items, but also provided feedback on the importance and curriculum-relevance of the topics, the presentation of the questions, and the appropriateness of the scoring procedures.

While the field test sample was not designed to be nationally representative, it did contain a wide range of ability levels, as well as a substantial number of black and Hispanic students. Results from the field test guided the design of the full-scale test administration the following year (see Dowd et al., 1991). Here is a summary of the relevant findings which aided the development of the main study tests:

- Test takers had more difficulty understanding what was expected of them in constructed response format than on the multiple choice tests, where the presence of answer choices clearly defined the objective of the question. For example, one problem asked, "What is the relationship between x and y?" and many students answered, "x and y are inversely proportional." This answer, while true, was not as complete as had been intended. In the rev ised test, the wording was made more precise: "Find an equation which shows the relationship between x and y." A challenge in writing the constructed response tests was to write questions that were explicit enough for students to understand just what was expected of them, but that did not hint at answers students would not otherwise have been able to provide.

- Field test participants were more likely to omit constructed response items than multiple choice items. Although a disproportionate number of the multiple choice items being field tested for the second follow-up were quite difficult, test takers tended to take a guess if they didn't know the answer. The percentage of omitted multiple choice items (aggregated across all test takers and all multiple choice questions) was 6 percent for the two mathematics forms, 5 percent for the two science forms, and 2 percent for the single reading form. Most of the constructed response items had higher omit rates, markedly so for questions that involved technical

mathematics and science material. Each of the constructed response mathematics questions was omitted by 2 to 34 percent of the total group, while 11 to 59 percent of test takers left each constructed response science item blank. Even the constructed response reading comprehension items, which did not contain any unfamiliar technical material, had omit rates of 5 to 12 percent, several times the multiple choice rate.

- Omit rates were examined for gender and ethnic subgroups on each test form, as well as for the total group. While the field test sample was not nationally representative, it included 53 to 77 black students and 68 to 84 Hispanic students taking each of the five test forms. In the multiple choice sections, omit rates for all population subgroups were very similar, in most cases differing by no more than one percentage point. The greater tendency to omit constructed response questions (relative to multiple choice) was similar for males and females, but considerably greater for black and Hispanic students than for white students. The ethnic group discrepancies were greatest for the most difficult mathematics and science items, but were present for the reading items as well.

- Students who had not taken advanced coursework in science and mathematics tended to be more likely to omit constructed response items than students who had taken these courses. The gaps in omit rates were greatest for questions with technical content, such as a mathematics question involving differences in relative area and perimeter of equilateral versus isosceles triangles, and a science question that required the test taker to compute the speed of railroad cars after a collision. Other questions were based on topics whose content would be familiar to most test takers, for example, reading a train schedule or describing an eclipse. For these non-technical questions, differences in omit rates between groups of students with different amounts of coursework were small. The items that had the greatest success in eliciting storable attempts from most test takers were the reading items.

- The most successful mathematics items, in terms of response rates, were those that had been designed as a series of increasingly complex steps, so that even a student with little mathematics background could attempt to answer some part of the problem, and by doing so demonstrate his or her level of competence. The least successful items were those that required specific mathematics or science knowledge to even begin to formulate a response.

- In a low-risk setting such as the NELS:88 survey, test takers know that they (and their schools and teachers) will not receive any feedback on their performance. They will not be rewarded or penalized for the quality of their answers, or even for answering the questions at all. In such a setting, it is incorrect from a measurement perspective to score "zero" for a completely blank problem because there is no way of knowing whether lack of ability or lack of motivation was responsible for the decision not to answer. One of the objectives in selecting and redesigning constructed response items from the field test was maximizing the number of students who could and would make an attempt to answer at least some part of each problem.

- For those who did answer the test questions, scores were analyzed to evaluate item difficulty and format-by-subgroup interactions. It has been suggested that standardized tests are biased against members of racial/ethnic minority groups, and that new modes of assessment may give students in these groups a better opportunity to demonstrate what they know (see Hartle and Battaglia, 1993). Subgroup performance on multiple choice versus constructed response sections of the field test was examined to determine whether the multiple choice format was

relatively disadvantageous to minority groups. Correlations of constructed response scores with variables for subgroup membership were calculated, with multiple choice scores martialled out. Black and Hispanic students tended to score lower on the constructed response sections than did white and Asian students, *even when multiple choice score was controlled for. In* other words, average score deficits for the black and Hispanic students, relative to white and Asian test takers, tended to be *greater in* constructed response format than on the multiple choice section of the test. As pointed out above, the field test sample was not systematic or nationally representative; however, the relative disadvantage of the constructed response format for minority students was consistent for all eight mathematics items. The performance differences in science and reading were less conclusive, but clearly showed no indication of any *advantage* for minority students in constructed response format. Score differences between male and female test takers were also analyzed. No substantial differences were found for any of the eight mathematics questions, while the science forms contained a mix of items that favored one gender or the other, as well as items with no substantial differences.

- The 12 minutes of testing time allowed for each extended constructed response item in the field test was reported by many students to be a little more than they needed to answer the question. The time was shortened to 10 minutes per item in the 1992 survey.

Chapter 3: Design of the 1992 Constructed Response Test: Objectives, Issues, Solutions

A decision was made to include a constructed response test component in the High School Effectiveness Study in 1992. This chapter will describe the factors considered in designing the test questions, and the steps taken to address these issues.

With the field test results and advice of the NELS:88 Technical Review Panel to guide them, test developers prepared constructed response test booklets for the 1992 High School Effectiveness Study (HSES) sample. Objectives in the design were selecting content that would be representative of what students might have learned by their senior year of high school; choosing appropriate difficulty level for the items; writing items that students could and would at least attempt to answer; and testing concepts and skills that were important for students to know, both as useful information in itself, and as a foundation for further study. Constructed response items were administered only in mathematics and science in 1992; reading comprehension was not included after the 1991 field test because of budget constraints. The topics below describe some considerations in construction of the multiple choice tests, and the parallel concerns for this constructed response experiment.

Domain Coverage

A test that claims to measure student achievement in a subject area must appropriately sample from the domain of knowledge the test claims to represent. The NELS:88 multiple choice mathematics tests taken by each participant contained 40 questions, which were administered in 30 minutes and covered a wide range of difficulty levels in arithmetic, algebra, geometry and advanced topics. The science test, with 25 questions, took 20 minutes and included questions in physical science, chemistry, and life science. The much longer time required for each constructed response item, 10 minutes per question, meant that only four problems could be administered in the limited time available.

An attempt was made to vary the content, context and format of the constructed response questions to cover as much of the domain as possible with this very limited number of test questions. Some of the material, such as a train schedule, a discussion of nuclear versus fossil fuels, or a lunar eclipse, would be familiar to students from their everyday life experiences or from exposure to issues in the news media. Other questions drew on content more closely related to school coursework, such as transfer of heat and computation of areas. Test takers were asked to interpret tables and graphs, draw diagrams, set up equations, and write explanations. However, even with a variety of format and content in the constructed response questions, it is obvious that four problems cannot pretend to even minimally represent all of the questions that could have been asked. Therefore, scores on the HSES constructed response tests should not be interpreted as representing students' overall level of math or science achievement.

Difficulty

Accurate measurement of individual achievement requires that each student answer test items of appropriate difficulty. Items that are much too hard for a given student provide very little information about the student's skill level; nor are items that are much too easy for the student very useful. Those test items that are slightly above and slightly below a particular student's ability level are the most valuable in pinpointing the precise standing of an individual relative to the skill being measured. Traditional multiple choice tests (that is, those that are not tailored or adaptive tests) attempt to match the range of item difficulty to the range of ability

levels found in the test-taking population. While most of the test items are likely to be either too easy or too hard for any given student, a few items will be at the right difficulty level to be valuable in determining the student's level of achievement.

An objective of the NELS:88 constructed response tests was that they be curriculum related. However, high school seniors have not all been exposed to the same curriculum. Some take no math courses after general math or algebra in ninth grade, while others continue a math sequence through calculus in grade twelve. A majority of students, though not all, take a biology course in high school, while fewer than half continue through chemistry and physics. Choosing items of appropriate difficulty for the NELS:88 constructed response tests meant trying to measure the wide range of mathematics or science knowledge to be expected in a sample of high school seniors, using the same four-item test for everyone. The difficulty of the tests needed to keep pace with student achievement in advanced courses in mathematics and science, while also accurately measuring achievement for students who had not taken these courses. Clearly, a four-question test cannot provide precise measurement for this wide range of knowledge. The NELS:88 test developers approached this challenge by designing each constructed response test item to provide information at different levels of achievement.

The constructed response mathematics questions consisted of multi-step problems, beginning with a near-trivial step, such as determining whether a student was able to read information from a table or graph. Subsequent steps required various manipulations of the data, or elaborations on the original simple procedure. In the hardest step of the problem, the student might be asked to write a general formula that described the process. Almost all test takers could be expected to be able to cope with the easiest steps of the problems, while only a small percentage would be able to complete all parts correctly. Thus, each problem would measure ability to perform across a fairly wide range of task difficulty rather than at a single point. The strategy of using multi-step problems was adapted from a study by Thomas Romberg (1982), as were some of the test items themselves.

Similarly, each constructed response science question was designed to be answered by students with a broad range of levels of science understanding. A question on the advantages and disadvantages of nuclear versus fossil fuels might be answered in a very simplistic or a much more comprehensive way, while an ecology question asked test takers not only to show relative numbers of predator and prey species on a graph, but also to explain the fluctuations of the animal populations over time. Most students would find the content of the questions familiar enough that they could attempt to respond, but only those with the most sophisticated understanding of the scientific concepts would be able to give the thorough and complete answers that would receive full credit.

Motivation

From the students' point of view, the NELS:88 tests were low-risk. That is, students knew that neither they nor their schools, teachers, or parents would ever receive copies of their scores. The results would not affect their grades, course credit, or college admission. They would receive neither reward nor punishment for performing well or poorly, or even for answering the questions at all. Students' only motivation to give their best answers on the tests was their willingness to cooperate with the objectives of NELS:88. Users of survey test scores have little choice but to assume that students have tried their best, and that their scores are good estimates of their achievement levels.

NELS:88 multiple choice test results have been consistent with this assumption. Several indirect indicators of motivation have been (a) high internal consistency reliabilities, (b) few unanswered items, (c) relatively small numbers of students with patterned responses (e.g., 1212121212), and (d) a very small percentage of scores around the chance level or below. All of these findings suggest that lack of motivation has not been a serious problem in the NELS:88 multiple choice tests (see Rock & Pollack, 1995).

Constructed response questions in the 1991 NELS:88 field test (Dowd et al., 1991) and in the 1990 NAEP survey (Swinton, 1993) had higher omit rates than did multiple choice questions, even if they were no more difficult. In a low-risk setting, students who are willing to cooperate with the relatively low-effort task of choosing between multiple response options may simply not be willing to exert the extra effort that constructed response questions require. Motivation may also interact with item difficulty. If test takers do not know the answer to a multiple choice question, it is easy to simply guess at random. Since most of the randomly guessed answers are likely to be wrong, such a response pattern provides a good indication of what the student did and did not know. However, in constructed response format, coming up with an answer from scratch when one has not mastered the material is much more difficult than simply guessing; students may simply leave the item blank. While inability to answer may account for many omitted constructed response items, field test results showed that it is clearly not responsible for all of them. Many students who performed well on the multiple choice sections of the test left at least one constructed response question blank. Others omitted constructed response items and then indicated in the followup questions that the material was not too difficult for them.

Since it can be difficult or impossible to draw valid implications from unanswered test items, it is important to try to motivate students to answer all questions to the best of their ability. Efforts were made to select constructed response questions for the HSES survey that students would find interesting and relevant to their lives and experiences rather than based strictly on abstract academic concepts. For example, there were questions related to train schedules, car stopping distances, nuclear fuels, and ecology. The multi-part structure of most of the problems was also designed to help to minimize nonresponse. Students might lack the skills necessary to complete *all* of a difficult math problem, or to give a thorough explanation of a scientific phenomenon. However, they still should have been able to *begin* each question and provide *some* storable response with very little effort.

Reaction Questions

As described above, in a low-risk testing situation it cannot be assumed that an unanswered item is equivalent to an incorrect response. In an attempt to identify their reasons for omitting responses, students were asked a series of questions about their reactions to each of the four constructed response test items. They were asked to evaluate the difficulty, clarity, and timing of each test item, as well as the quality of their response and the adequacy of their coursework background. These student reaction questions were designed to aid in distinguishing between items that were omitted because the student was unable to answer, which might legitimately be treated as incorrect responses, and those that *were* left blank for some other reason such as lack of motivation, that must be considered missing data. Test takers' self-report of finding questions too difficult or of not knowing how to answer could be used as a basis for deciding whether imputing scores for unanswered questions might be justified. The imputation procedure will be described in more detail in the section on scoring below. The text of the reaction questions can be found in Appendix A.

Explicit Instructions

The 1991 field test demonstrated that test takers did not always target their responses in the way test developers had anticipated the questions would be answered. This outcome probably resulted from several aspects of the interaction of the constructed response format with the differences between classroom tests and the NELS:88 survey setting.

First, in a classroom setting, the test takers and the test administrators/evaluators (teachers) know each other. From previous experience taking a teacher's tests and from the curriculum unit covered by a test, students know what is expected of them. They know how extensive their answer must be to receive full credit, and whether or not the teacher will take into account things like neatness, correct spelling and grammar, or showing intermediate steps in a problem solution. Similarly, teachers know the students: given previously-demonstrated capabilities, they may be able to guess whether a sketchy or incomplete response might or might not be indicative of lack of mastery of the material. This familiarity, which enables the test takers to correctly interpret the intentions of the test writers, and the evaluators to interpret the responses of the test takers, does not exist in a large-scale survey.

A second aspect of the format by setting interaction, once again, is the minimal motivation that must be expected in the low-risk survey setting. In a classroom or admissions test, it is in students' interest to give the *best answer they can.* But low-risk survey participants, even those who have chosen to respond to all of the test questions, may still give the minimal response that seems to answer the question without bothering to elaborate.

Differences in achievement scores should result only from differences in ability to answer the question, not from differences in test takers' interpretation of what was expected in the way of an answer. In revising the field test constructed response items in preparation for the 1992 HSES administration, test developers attempted to clarify the item stems to let students know how extensive and how precise their answers were expected to be. This may have sacrificed some of the "open-endedness" of the items, by restricting the range of possible responses to the ones that the test writers had in mind rather than allowing students to write everything they knew about a particular subject. It may also, in some cases, have given hints that enabled test takers to answer items they otherwise might not have understood, for example, in a math problem in which formulas and several examples were given. But if all responses were to be scored according to the same set of objective standards, it was essential to be certain that test takers understood the intent of each question.

Appendix A contains copies of the four mathematics and four science test items.

Chapter 4: High School Effectiveness Study Sample

This chapter will report on the characteristics of the students who took the constructed response tests. The test taking sample will be compared with estimates for the national population of twelfth graders, with respect to demographic proportions and average achievement.

Two hundred forty six NELS:88 second followup schools and over seven thousand students participated in the High School Effectiveness Study in 1992. About one-third of the participating students were members of the NELS:88 core sample (the national survey representative of the population of eighth graders four years later). The other two-thirds were additional students sampled in the HSES schools to achieve a representative within-school sample of a large enough size to support analysis of school effects and hierarchical linear modeling techniques, as described in Chapter 1. The 1992 HSES sample was intended for methodological purposes rather than for generating national estimates. Student questionnaires and multiple choice tests were administered in the HSES schools, and transcripts were collected In addition, students in half of the schools were targeted to receive constructed response tests in mathematics; in the other half of the schools, constructed response science tests were to be administered.

Table 4.1:
Counts of Schools, Participants, and Test Takers

	Mathematics	Science
HSES Schools	123	123
HSES Participants	3,553	3,535
Participants with Multiple Choice Tests	2,832	2,588
HSES Schools with Constructed Response Tests	110	108
Participants with Constructed Response Tests	2,415	2,239

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

However, not all of the participating schools agreed to allocate enough time for all of the survey instruments to be administered. In those that did, not all students participated in all aspects of the survey. Whether because of time constraints, scheduling conflicts, or student and/or school refusals, only about 68 percent of the HSES participants in mathematics test schools, and 63 percent of the participants in science test schools, took the constructed response tests. (Response rates for individual test questions are presented in the section on Missing Data in chapter 6.)

Only unweighted statistics are reported here, and no claims are made that the results are representative of a larger population. Findings of statistical significance in the results that follow are for the HSES sample alone, with no assumption of generalizability. However, to aid interpretation of the results of the HSES

constructed response analysis, it is helpful to see by just how much the unweighted HSES sample deviates, in its demographic characteristics and ability level, from the NELS:88 national sample population estimates. Table 4.2 shows the gender and racial/ethnic group proportions of the HSES constructed response sample compared with national estimates. Almost all of the HSES test takers were in twelfth grade (98 percent for the math test, 97 percent for science) so the relevant comparison group is the NELS:88 core twelfth grade sample rather than the full NELS:88 second followup sample, which also includes early graduates, dropouts, and students who had not progressed to grade twelve. The NELS:88 sample design intentionally oversampled Asian and Hispanic students, with sample weights for the NELS:88 core sample compensating for the oversampling. The proportions of Asian and Hispanic students in the HSES constructed response sample are each about 6 percentage points higher than in the grade twelve population, and the proportion of white students about 12 percentage points lower. These differences may be partly due to higher concentrations of Asian and Hispanic students in the urban setting of the 30 largest MSAs from which the HSES sample was drawn, and to differential rates of participation in the voluntary testing activities, as well as to the sample design. Since sample weights that generalize to a larger population will not be computed for the constructed response test takers, these comparisons are presented only to point out the most obvious similarities and differences.

Similarly, it is possible to compare the mathematics and science achievement levels of the HSES constructed response test takers with those of the NELS:88 nationally representative sample of twelfth graders. The same multiple choice tests in mathematics and science were taken by both groups.

Table 4.2:
Sample Sizes and Subgroup Proportions
National Estimates Compared with HSES Constructed Response Test Takers

	Estimated Grade 12 Population (weighted NELS:88 core sample)	HSES Test Takers Constructed Response Math (unweighted)	HSES Test Takers Constructed Response Science (unweighted)
Total N	2,537,024	2,415	2,239
Male	51%	52%	50%
Female	49%	48%	50%
Asian	4%	11%	10%
Hispanic	10%	16%	16%
Black	13%	14%	14%
white	71%	59%	59%
American Indian	1%	1%	1%

NOTE: HSES percentages arc unweighted because weights were not created for the HSES constructed response test methodological sample.

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

Examination of average scores on these tests for the *weighted* core sample (national estimates of the twelfth grade population) compared to the *unweighted* HSES group can give an idea of the size and direction of biases in the sample of students who took the constructed response tests.

Evidence from the multiple choice mathematics test scores shows that the HSES participants had slightly higher average levels of mathematics achievement than the national population by about 13 percent of a standard deviation. Potential differences due to oversampling of Asian students were approximately canceled out by comparable oversampling of Hispanics. The 7 percent oversampling of Asian students, who scored, on average, about half a standard deviation *higher* than *the* total HSES group, was approximately counterbalanced by the 6 percent overrepresentation of Hispanic students, with average scores half a standard deviation *lower* than the total. The gender and racial/ethnic subgroups in the HSES sample had consistently higher average mathematics achievement than the comparable groups in the weighted core sample, although the differences for Hispanic, black, and American Indian students were small and not statistically significant.

The group of students who took the HSES constructed response science test had about the same average achievement in science as the core sample, as measured by the multiple choice test taken by both groups. Differences in mean scores for gender and racial\ethnic subgroups were generally within about 10 percent of a standard deviation and were neither consistent in direction nor statistically significant.

Table 4.3:
Average Multiple Choice Test Scores by Subgroup
National Estimates and HSES Samples

	Mean Math Score		Sample N HSES Math	Mean Science Score		Sample N HSES Science
	National	HSES		National	HSES	
Total (s.d.)	48.8 (14.2)	50.7 (15.4)	2386	23.5 (6.2)	23.5 (6.7)	2200
Male	49.4	51.9	1235	24.4	24.1	1103
Female	48.3	49.3	1151	22.6	22.9	1097
Asian	53.1	58.3	253	24.0	24.9	230
Hispanic	42.1	42.4	378	20.6	20.1	347
Black	39.2	40.3	318	18.6	19.1	300
White	51.0	54.1	1404	24.7	25.2	1302
American Indian	39.8	41.6	31	19.5	18.1	13

NOTE: Only HSES students who had multiple choice as well as constructed response test scores are counted in this table. The multiple choice test scores reported here are estimates of performance on a selected item pool, scaled according to a complex Item Response Theory (IRT) based procedure, rather than simple counts of number of correct answers. This accounts for score means that may be higher than the actual number of items administered on a particular test form. The NELS:88 multiple choice mathematics test consisted of three different forms, varying in average item difficulty. Students who had taken the mathematics test in 1990 were assigned to the low, middle, or high difficulty form in 1992, based on their performance in the earlier year. Scores were equated to the same scale.

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

Despite the lack of sampling weights that would permit population estimates of performance on the constructed response tests, some generalizations are supported by the comparisons with the core group:

- HSES mathematics test takers were slightly higher achievers than the national population.
- HSES science test takers had achievement levels very similar to the national population estimates.
- Black and Hispanic students in the HSES sample differed by less than a tenth of a standard deviation from black and Hispanic students in the national population in both mathematics and science achievement.

Chapter 5: Scoring Procedures

Scoring constructed response test questions is a complex process, both conceptually and operationally. Unlike multiple choice questions, which have a single correct answer and can be scored by a computer, constructed response scoring generally requires subjective decisions in establishing the scoring criteria, and human judgment to determine how well test takers' responses meet these criteria. This chapter will present the criteria and procedures used in evaluating student responses to the HSES constructed response tests. The treatment of missing data will also be described.

The constructed response tests were scored by teams of readers, most of whom were high school math and science teachers. Readers were trained to apply a set of scoring protocols to ensure that a common set of standards was being applied to all papers and that the scoring was as objective as possible. One multi-part question was scored at a time, that is, all readers worked on scoring math question 1 until all of the tests had been read before moving onto training and scoring for another question. About twelve to fourteen readers and two coordinators took one week to score the tests in each of the two subject areas.

Analytic and Scale Scores

There are two types of scoring approaches typically used to evaluate constructed response questions: holistic and analytic. Holistic scoring assigns a single score that takes into account the overall impression or quality of the response according to an established set of criteria. Analytic scoring rates each of a number of features separately, for example, using the correct equation, doing computations accurately, using the correct metric, and labeling variables. The analytic method was chosen to score the HSES constructed response tests because it offers the opportunity to preserve the maximum amount of information for study by researchers: not only how well students answered the test questions overall, but also what parts of questions caused problems, and what types of errors were encountered.

The analytic scoring procedure used for the HSES constructed response tests broke down each feature of each problem into a separate score with several objective categories. The number of analytic scores varied for each of the eight test questions, depending on how many individual steps or features could be identified within each problem. Scoring guides were prepared listing each feature or step of each test question, and for every feature, all of the types of responses that were envisioned by the test developers or found in a review of the booklets prior to the scoring sessions. (Other categories were added to the lists during the scoring sessions when unanticipated responses were encountered.) Readers were asked to identify which of the descriptions in the scoring guide best fit each feature of the responses in the students' test booklets. Codes for the responses did *not* correspond to a point-count or relative value; they were strictly categorical.

For example, one analytic score was assigned for a step of the balance beam problem that required the students to determine the correct placement for a weight that would balance the system. There were several ways that test takers could get this step wrong or partially correct: by omitting it entirely, by misunderstanding the correct method in various ways, by making computational errors, etc. The categorical scores of O through 9 listed in the scoring guide do *not* correspond to increasing levels of correctness, but merely to different ways that test takers might have responded.

After all papers for each test question had been read, the readers and test developers discussed building an overall score scale, using different combinations of the analytic categories, that would correspond to identifiably different levels of performance. Final definitions of score scales utilized the judgments expressed

by the readers and information from analysis of the test data. Comparisons of constructed response scores with students' performance on the corresponding multiple choice test component served to validate the scale score definitions. (While these comparisons were useful in verifying that the translations of sets of analytic categories constituted meaningful scales, it is important to note that the use of the multiple choice test scores for validation may tend to produce bias toward a higher correlation between the multiple choice and constructed response sections.) The score scales were designed with a score of O indicating complete inability to understand or respond correctly to any part of the problem, and a score of 5 signifying a complete and correct response including the most difficult step. Scores of 1 to 4 were identified with combinations of analytic scores demonstrating increasing levels of competence. This 0-5 score scale was used for each of the four mathematics and four science questions, regardless of the number of steps or analytic scores.

The transformations of analytic scores to scale scores are based *on the subjective* judgments of the test developers, readers, and analysts about which categories of student responses demonstrated mastery of various concepts or skills, and also about the relative importance of the different skills in defining competence. Constructed response questions do not always have a single correct answer; score scales represent the choices, values, and emphasis of the people who developed them. For example, the score scales for these test questions could have rewarded good grammar, spelling or rhetoric in test items that required explanations, or neatness and artistic ability in diagrams. Instead the score scales were consciously defined to be limited as narrowly as possible to the mathematics or science concepts or skills that the items were designed to test. It is important to remember that these judgments could have been made differently, and that other definitions of scales might have resulted in findings very different from those reported here.

Complete descriptions of the analytic scores, the features of each response as categorized by the readers, may be found in Appendix A. Again, note that the codes for these categories do not imply a hierarchy of correctness. Descriptions of how the analytic scores were combined to develop a 0-5 score scale for each item are also included.

Imputation of Missing Scores

Some questions could not be scored because the test takers had not attempted to answer them. Rather than treat all unanswered questions as missing data, the student reaction questions following each test item were used to determine whether score imputation might be justified. If an omitted item was followed by an indication that the student had been unable to answer, a zero scale score was imputed (the student checked "hard" or "too hard" for the question on item difficulty; or "I really didn't know how to answer the question" or "No, I have not taken the courses needed to answer the question"). If the reason for the nonresponse could not be determined (no indication of inability in the reaction questions, or no response to the reaction questions), then low ability could not be assumed, and the scale score was left blank.

As a check on the reasonableness of the imputing procedure, average scores on the corresponding multiple choice test section were computed for students scoring at each step of the score scale, as well as for the omitted items that were and were not given an imputed zero score. For each of the four mathematics constructed response questions, the mean multiple choice score for the group of students with an imputed zero scale score closely resembled the mean multiple choice score of students who had actually answered the question and received a zero score. This supports the assumption that students who indicated that they were not able to answer the question would indeed have scored poorly if they had tried. Conversely, those who omitted a test item and did *not* provide a basis for imputation (that is, did not answer the reaction questions, or answered in a way that did not indicate inability to respond) had average scores closer to those of all students in the sample than to those with

actual (not imputed) zero scores. Factors other than sheer inability to answer clearly contributed to decisions to omit items for at least some of this group of the nonrespondents, so their scores were left blank.

Table 5.1:
Average Multiple Choice Test Scores for Each Scale Score Level
Mathematics Question 2

Scale Score	Number of Cases (Number with a Multiple Choice Score)	Multiple Choice Mathematics Test Score	
		Mean	S.D.
Total Sample	2415 (2386)	50.7	15.4
0	124 (118)	33.4	10.9
1	384 (378)	39.5	11.8
2	557 (554)	46.9	11.7
3	85 (84)	49.5	13.2
4	390 (386)	53.0	12.3
5	706 (701)	64.8	8.7
No Response (0 Imputed)	115 (112)	33.0	11.0
No Response (Missing Data)	54 (53)	44.4	15.5

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

For example, Table 5.1 illustrates multiple choice test statistics for students grouped according to their scores on constructed response mathematics question 2. Of the 115 test takers who omitted this item *and* indicated that they were unable to answer, 112 had taken the multiple choice math test. Their average score on this test was 33.0, very close to the 33.4 mean for the students who *did* respond to question 2 and produced a completely incorrect answer. The standard deviations for these groups (11.0 versus 10.9) are also close to identical. Therefore the decision to impute a zero score for question 2 for this subset of the nonrespondents is supported by comparison with another measure of mathematics achievement. On the other hand, the 54 nonrespondents who did *not* indicate that they were unable to answer question 2 appear to be very different from the lowest ability group, with a mean multiple choice score of 44.4 and about as much variance as the total sample. In other words, this group consists of both low and high achieving mathematics students. Imputing zero scores would not be a reasonable estimate of their ability to respond. Their nonresponse to question 2 cannot be assumed to be due entirely to inability rather than motivation or other factors; their scores have not been imputed but are treated as missing data.

Score comparisons for the imputed zero versus missing data groups for the four constructed response science items produced fairly similar results. The students for whom zero scores were imputed had average multiple choice science scores that were consistently lower than the average for those with unsuccessful attempts to respond. Average multiple choice science scores for nonrespondents who did not indicate inability (and were not imputed) fell somewhere between the averages for the total sample and for the actual (not imputed) zero-score group. As was the case for the math item described above, standard deviations for the nonrespondents whose scores were not imputed were generally at least as high as those of the total sample, indicating a mix of low and high achieving students in the missing data group.

The imputation procedure used applied only to test questions that were completely blank. It did not attempt to compensate for missing data on *parts* of multi-step problems. Capable students may have received low scores if they answered the first part of a problem and omitted the rest. In-depth study of these partial-omits is beyond the scope of the analysis reported here. However, the existence of the analytic scores, along with data on students' coursework background, grades, and performance on other measures, could be used in developing a more elaborate imputation scheme.

Appendix B contains scale score distributions for each of the four mathematics and four science questions, with the nonrespondents broken out into imputed-zero and missing data groups. Score means and standard deviations for the corresponding multiple choice test are included in the tables. (Note that the multiple choice scores are not simple counts of number of correct answers. Their scale is not the same as the number of items administered on each test form.) All of these statistics are also broken out according to students' perception of the tests and their performance on them as reported in the student reaction questions.

Chapter 6: Statistical Analysis of Test Results

This chapter will present findings from analysis of the constructed response mathematics and science tests in the High School Effectiveness Study. Reliabilities of the analytic and scale scores will be presented, as well as statistics on student performance and omit rates. Comparisons of the constructed response tests with the ax-responding multiple choice tests taken by the same students, and comparisons of test results for gender and racial/ethnic subgroups will be shown. The results of a factor analysis of the combined multiple choice and constructed response test sections will be presented. Finally, a summary of the test takers' responses to the student reaction questions will be reported.

Reliability

A test is said to be a reliable measure of a construct if it measures the construct consistently, that is, if the same measurement of the test taker's competence would be obtained under a variety of circumstances. The variation in circumstances might be the same test taken at another time, or a score on a parallel form of the test, that is, another test with items that have the same content and difficulty. Assuming that the characteristic being measured (the test taker's ability) has not changed, a reliable test should produce the same measurement of the characteristic under different circumstances.

Reader Reliability

Constructed response tests have an additional potential source of unreliability that is not present in multiple choice format: the possibility that different human scorers will evaluate a test taker's response differently. Any ambiguity in the definitions of the scoring criteria, or differences in the way the criteria are applied, may lead to different measurements of test takers' performance. In order to maximize objectivity, the HSES constructed response scoring procedures used analytic scoring (categorizing identifiable features of each answer) rather than holistic scoring (asking the readers to make a judgment on the overall quality of the response). While the readers' judgments played a part in *defining the* scales that were built from the analytic scores, the readers did not themselves assign the scaled scores. They assigned only the analytic scores; scale scores were later computed according to the specifications described in Appendix A.

About ten percent of the HSES constructed response test questions were selected at random to be scored by a second reader, who did not have access to the first reader's scores. These second readings provide a basis for evaluating the reliability, or consistency, of the scoring procedures. Table 6.1 summarizes the reader reliability statistics for the four mathematics and four science questions. For each test question, the table shows the lowest and highest proportion of first reader/second reader agreement of the 3 to 10 analytic scores used in construction of the scale score. The proportion of agreement of the scale score computed from each reader's analytic scores is also shown; both the proportion of scores that agree exactly, and the proportion that are either identical or discrepant by no more than one point on the 0-5 scale. (In most constructed response tests administered by Educational Testing Service, a one point difference between readers is not treated as a discrepancy needing resolution. Factors such as the length of the score scale, the location on the scale at which a discrepancy occurs, and the consequences of the score to the test taker may need to be taken into consideration in deciding whether small discrepancies are important.)

Table 6.1:
Reader Reliability
Percent of Reader l-Reader 2 Agreement

Test Question	# Reader Pairs	Agreement of Analytic Scores	Scale Score Exact Agreement	Scale Scores Within 1 Point
Mathematics				
Question 1	291	76 - 99%	89%	98%
Question 2	241	77 - 93%	84%	94%
Question 3	271	82 - 92%	83%	95%
Question 4	248	82 - 98%	94%	98%
S				
Question 1	244	50 - 98%	57%	89%
Question 2	323	62 - 89%	73%	90%
Question 3	293	66 - 98%	62%	83%
Question 4	395	63 - 73%	68%	89%

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

In general, there were higher levels of reader agreement for the mathematics analytic scores than for the science. Most of the mathematics features could be evaluated relatively unambiguously: a computation either resulted in the correct answer, or one of several incorrect answers; if incorrect, it was usually clear which of several mistakes had been made. Judging whether a diagram had the required lines, boxes or numbers, and in the right positions, was relatively straightforward. The science items, however, relied more heavily on descriptions or explanations. For example, the first question asked test takers to compare the use of nuclear fuels to the use of fossil fuels, describing at least one advantage and one disadvantage of each type. A response that stated "nuclear fuels are more expensive to produce than fossil fuels" might be interpreted by one reader as an advantage of fossil fuels, by a second reader as a disadvantage of nuclear fuels, and by still another reader as fulfilling two requirements of the question. The scale score definitions compensate for some of the individual-feature discrepancies: in the example above, one advantage of fossil fuels receives the same amount of credit as one disadvantage of nuclear fuels. So not all differences in categorical analytic scores result in scale score discrepancies.

It was not possible to compare reader l/reader 2 agreement for the total scale score summed across the four test questions. The 10 percent reliability sample was chosen independently for each test question; so very few papers had a second reader score for more than one question. Budget constraints precluded second readings for the whole sample of student responses, which would have made comparisons of total score reliability possible as well as in-depth study of the sources of variation that account for the score differences.

Complete counts of first reader/second reader judgments are presented in Appendix C, for each of the categorical analytic scores as well as for the scale score for each test question. In general, the highest reader reliability statistics are obtained for features that can be explicitly categorized as correct, or as incorrect in well-defined ways. The lowest levels of agreement correspond to aspects that depend more on the subjective judgment of the reader, such as whether a test taker's explanation shows understanding of the concept. This is an essential dilemma of constructed response testing. The very "open-endedness" of test questions that allow students to demonstrate what they know also makes them difficult to score reliably. Conversely, the reliable measurement possible with explicit questions that elicit specific answers may be obtained more economically with other item formats that are less time consuming to administer and less expensive to score.

Alpha Coefficient and Split Half Reliability

Adequacy of domain coverage affects reliability of both multiple choice and constructed response tests. For a fixed amount of testing time, this is a more serious issue for constructed response questions, since they take longer to answer, resulting in fewer questions possible in the time allotted. The test forms used in the High School Effectiveness Study included 40 multiple choice mathematics items, which were administered in 30 minutes, while 40 minutes were required for the 4 constructed response items. The science tests, with the same constructed response timing as mathematics, allowed 20 minutes for 25 multiple choice items. While the constructed response questions provide more information (a 0-5 score scale rather than a simple right/wrong), the range of topics they covered was necessarily quite limited.

For multiple choice tests, a commonly used measure of reliability is the alpha coefficient, which measures the internal consistency of the items, or the proportion of variance among people which is due to true or common variance (differences in test takers' levels of achievement) rather than error or unique variance (variation in scores caused by errors of measurement including test items that measure somewhat different constructs). Another standard measure is the split half reliability, which is a transformation of the correlation of scores on half of the test items with scores on the other half. This is a simulation of the idea that scores on parallel forms of a test should be closely related. Two halves of the test (odd/even items, randomly chosen items, or some other method) are treated as if they were parallel forms; an adjustment to the correlation of the two halves is necessary to compensate for the fact that each of the "forms" is half the length of the actual test.

Table 6.2 presents alpha coefficients and split half reliabilities for the multiple choice test alone, the constructed response section alone, and the two formats combined and treated as a single test. Only test takers who answered all four constructed response questions are included in the statistics in Table 6.2. Because computation of the reliability coefficients depends on the set of items being the same for all observations, the reliability statistics for the mathematics group are further restricted to students who took the "middle difficulty" form of the multiple choice math test (about 58 percent of the sample); test items on the low and high forms are not comparable.

Table 6.2:
Alpha and Split Half Reliability Coefficients,
By Test Format and Content Area

	Multiple Choice	Constructed Response	Combined Formats
Mathematics			
Alpha	.86	.74	.87
Split-Half	.87	.76	.90
Science			
Alpha	.84	.70	.85
Split-Half	.85	.71	.88

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

The split half reliability was based on the total number of correct odd-numbered versus even-numbered items for the multiple choice test. For the constructed response test, split half reliabilities were computed for each possible pairing of the four test questions (1+2 versus 3+4, 1+3 versus 2+4, and 1+4 versus 2+3) and averaged. Three pairings were also computed and averaged for the "combined formats" statistics, with the first element of each pairing (e.g., constructed response question 1+2) added to the odd-numbered multiple choice items and the second element (e.g., question 3+4) to the even-numbered item sum. Differences among the three item pairings were extremely small (.04 or less) for both mathematics and science.

The multiple choice mathematics and science tests appear to be nearly identical with respect to reliability. However, two unrelated factors, with opposite effects, influence these numbers. The first is the number of test items. In general, the longer a test is, the higher reliability it will have, assuming that the items maintain the same level of internal consistency. The mathematics test, with 40 items, should have had a substantially higher reliability than the 25-item science test. This potential advantage in reliability for the multiple choice mathematics test was counteracted by a second factor. The necessity of calculating reliabilities using only students who took the same test form (the middle difficulty mathematics form) meant that some of the lowest and highest ability students were not in the sample on which the reliability was computed. This restriction in range meant that the variance of total scores, and therefore the reliability (proportion of "true" variance to total variance), was lower than it would have been if the students who took the low and high difficulty forms of the multiple choice mathematics test had been included in the computation. This was not the case for the science test, where *all* students took the same test form. But the objective here is to compare the levels of reliability for the item formats, not for the mathematics versus science tests. No attempt was made to apply corrections for test length or for restriction in range that would have made the mathematics and science statistics really, instead of merely apparently, comparable with each other.

Results were remarkably consistent for both types of reliability coefficients, and for both the mathematics and science tests. The multiple choice tests alone had an acceptably high degree of reliability and the constructed response sections a substantially lower level. Combining the item formats produced reliability coefficients that were greater than the multiple choice tests alone, but only by a very small amount. If the purpose of adding constructed response items to a test were to increase its reliability, there is a faster and less expensive way to do so-by simply adding a few more multiple choice items. However, if constructed response questions are added for other reasons, for example, to increase the face validity of the test, there is no evidence here that doing so would necessarily have a negative impact on test reliability. Indeed, if inclusion of constructed response items improves the credibility of test results, their use may be justified for this reason alone.

Missing Data

Students showed a greater propensity to omit items in the constructed response tests than in the corresponding multiple choice section. This tendency was more pronounced for the science test than for the mathematics test, and also varied for gender and racial/ethnic subgroups. The results shown in Table 6.3 are consistent with findings from the National Assessment of Educational Progress (NAEP), in which omit rates for constructed response items (especially "extended open-ended" items, which are comparable in format to the HSES questions) are substantially higher than for multiple choice questions (Swinton, 1993). On the HSES multiple choice tests, most students answered most or all of the questions. Overall, only 3.5 percent of the 40 mathematics questions were omitted, and 2.5 percent of the 25 science questions.

Table 6.3:
Percentage of Omitted **Test Items**

	Mathematics		Science	
	Multiple Choice	Constructed Response	Multiple Choice	Constructed Response
Total	3.5	6.5	2.5	11.3
Male	3.1	7.6	2.2	11.3
Female	3.9	5.4	2.9	11.3
Asian	3.3	6.0	1.7	7.4
Hispanic	4.1	9.8	2.0	14.2
Black	4.0	12.3	5.9	23.9
White	3.3	4.4	2.0	8.0

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

The subgroup differences in omit rates for the multiple choice mathematics test were very small. Females were slightly more likely than males to leave multiple choice mathematics questions unanswered, while black and Hispanic students omitted slightly more items than did white students.

For the constructed response mathematics test the male/female nonresponse pattern was reversed. Although males in this sample omitted fewer multiple choice mathematics questions (and scored higher than females, by about a fifth of a standard deviation in both formats), they were *more* likely *than* females to leave constructed response test questions blank. This reversal strongly suggests that factors other than inability to answer enter into students' decisions to respond to constructed response test questions in a low-risk test.

There was no such reversal for the racial/ethnic groups in the HSES sample. Black and Hispanic students, who had only slightly higher omit rates than whites in the multiple choice mathematics section, were much more likely to leave constructed response mathematics questions blank (9.8 percent of questions for Hispanic and 12.3 percent for black test takers, compared to 4.4 percent for white test takers).

Science test nonresponse rates were similar to mathematics with respect to gender differences. Males, who on average scored higher than females, were less likely than females to omit multiple choice questions, but equally likely to omit constructed response items. As was the case for mathematics, their higher average achievement (about a fifth of a standard deviation on the multiple choice science test) did not translate to a greater propensity to answer constructed response questions.

Nonresponse rates for black and Hispanic students on the constructed response science items were dramatically higher than for whites, with 23.9 percent of the questions omitted by black students and 14.2 percent by Hispanic students, compared to 8.0 percent for whites.

The higher nonresponse rates for science than for mathematics items were probably related to the design of the test questions. While the science questions *could* be answered in a non-technical manner by students with limited knowledge of the material, they did not start out with an explicit low-level, non-technical first step that was designed to elicit a storable response from everyone. Students who had scored poorly on the multiple choice test in the corresponding subject area were more likely to attempt the first, trivial, step of the mathematics problems before giving up than they were to make an effort to respond to the science questions that had no such stepwise design (see the tables of score means in Appendix B).

For both mathematics and science, the raw nonresponse rates for black and Hispanic students in the constructed response tests would be unacceptably high for a test intended to support population estimates (although this experiment was not). The resolution procedure described in the earlier section on imputation of missing scores addressed this problem with considerable success. By imputing zero scores based on students' self report of their inability to answer the omitted questions, the nonresponse rates were drastically reduced, as shown in Table 6.4. The procedure was more successful for the mathematics test than for the science test in separating nonresponse due to inability from nonresponse due to motivation or other factors. This is evidenced by the average multiple choice test scores for the imputed versus the unresolved blank scores shown in Appendix B. For three of the four mathematics questions, the mean and standard deviation of the multiple choice achievement measure for the unresolved group was very close to that of the whole sample, indicating that the nonrespondents' ability to answer the question, had they been motivated to do so, was about the same as anyone else's. (The remaining question had too few nonrespondents to draw any conclusions.)

Table 6.4:
Percentage of Omitted Constructed Response Test Items
Before and After Imputation Procedures

	Mathematics		Science	
	Before Imputation	After Imputation	Before Imputation	After Imputation
Total	6.5	2.9	11.3	3.3
Male	7.6	3.5	11.3	3.8
Female	5.4	2.2	11.3	2.8
Asian	6.0	2.7	7.4	3.1
Hispanic	9.8	3.5	14.2	4.3
Black	12.3	6.0	23.9	7.5
white	4.4	2.0	8.0	1.9

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

Imputation of zero scores for the science question was also successful in drastically reducing the amount of missing data, although it was less successful than mathematics in separating inability from motivation. The unresolved nonrespondents for the science questions continued to have somewhat lower average multiple choice scores than the total sample (by about one quarter to one half of a standard deviation), indicating that a disproportionate number of low achieving science students failed to answer the student reaction questions that were necessary for imputing scores. As mentioned earlier, the simple imputation procedures used here are merely a first step in exploring ways to deal with missing data. A more elaborate scheme involving corollary information such as transcripts of coursework and grades could be investigated to determine appropriate imputations for unresolved omits.

The nonresponse patterns for question formats and population subgroups described above illustrate several points: the importance of designing constructed response questions in a low-risk test in ways that minimize nonresponse, especially for members of racial/ethnic minority groups; the need to interpret nonresponse appropriately rather than scoring all blank questions as incorrect; and the utility of making it easy for test takers to indicate that they cannot answer a question. Test questions of a technical nature, such as in mathematics and science, will probably have lower nonresponse rates if they begin with a step so trivial that almost anyone *could* attempt to answer. Eliciting a storable response-even a completely incorrect one-makes it possible to avoid the problematic necessity of interpreting missing data. In tests where it is not practical to collect the extensive student reactions used for imputation here (the page of 5 questions following each constructed response test item), perhaps a place for test takers to check "I don't know how to answer this question" would serve a similar purpose.

For readers who are interested in nonresponse patterns for different test questions, Appendix D contains more detail on omit rates for each test question, before and after imputation, in addition to the four questions combined. Nonresponse percentages are presented for gender and racial/ethnic subgroups as well as for the total sample. Although the groups are not systematic samples of a larger population, standard errors based on the sample sizes for each test section are included to give the reader an indication of the stability of the mean estimates.

Average Scale Scores

Constructed response test questions were scored on a 0-5 scale, with a total score of 0-20 computed only for those test takers who had storable (or imputable) responses to all four mathematics or science questions. Total scores were available for 90 percent to 95 percent of each gender and racial/ethnic subgroup, with the exception of black students. For this group, 87 percent of those with math tests, and 83 percent of those with science tests, answered all four questions or had imputed scores. Students with complete/imputed data (total scores) scored higher on the corresponding multiple choice test section by about one third (science) to one half (math) standard deviation than those who had one or more unresolved omits. The complete-data **mathematics** group had achievement levels (as measured by the multiple choice test) about 15 percent of a standard deviation higher than estimates for the national population, while the science complete-data students exceeded national estimates by only about 3 percent of a standard deviation. As pointed out earlier, the HSES constructed response test taking sample was not designed to be representative of all twelfth graders in the nation. However, in interpreting performance on the constructed response tests, it is useful to keep in mind the evidence that the HSES group appears to be slightly more able than the national population.

Average constructed response total scale scores in both mathematics and science were lower for females than for males, and for Hispanic and black students than for white test takers. Estimates of the proportion of these gaps that may be due to differences in course-taking patterns or other factors have not been attempted for this report.

The score statistics in Tables 6.5 and 6.6 report comparisons of test formats and of demographic subgroups for the test takers in the HSES sample who took the multiple choice tests and also had scores (original or imputed) for all four constructed response questions. Table 6.5 shows mean mathematics scores by gender and racial/ethnic subgroup for the two types of formats. Differences between each group and a reference group are expressed in total group standard deviation units (effect sizes). The standardized metric is used for comparisons because the two formats and two subject areas have different score scales. Thus, direct comparisons of differences in terms of raw score points are meaningless.

For example, females with complete data, on average, scored 3.1 points lower than males on the multiple choice mathematics test, which is equivalent to 20 percent of a standard deviation. The gap in male/female performance is almost identical for constructed response format, 21 percent of a standard deviation. The difference in format does not appear to be relatively advantageous for either gender group. It should be remembered, however, that males had higher nonresponse rates for the constructed response test section, and that less able students tended to omit more of these items. If all students had scores on all four constructed response test items, this bias would have had the effect of shrinking the male/female difference somewhat, although probably not significantly so, since the amount of missing data was small.

Means for each of the racial/ethnic minority groups were compared with those for *white test* takers. *The* Asian students maintained their score advantage in both formats, while the black test takers had about the same disadvantage in each. The Hispanic/white gap in performance was smaller for the constructed response format than for the multiple choice test.

Table 6.5:
Mean Mathematics Scores, By Format and Subgroup
And Difference from Reference Group
in Standard Deviation Units (Effect Sizes)

	Multiple Choice		Constructed Response	
	Mean	S.D. Units	Mean	S.D. Units
Total (S.D.)	51.2 (15.3)		11.3 (5.3)	
Male	52.7		11.9	
Female	49.6	-20%	10.7	-21%
Asian	59.2	32%	13.8	28%
Hispanic	42.2	-80%	9.0	-63%
Black	41.0	-87%	7.6	-88%
White	54.4		12.3	

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

Once again, the greater tendency for lower achieving students to omit constructed response questions-and thus to be absent from these score means—must be considered in interpreting these comparisons. Assuming that the students who omitted each constructed response question would have scored lower, on average, than those who answered would indicate that score means would have been somewhat lower if all test takers had received scores. It is reasonable to assume that the higher the omit rate for a subgroup, the more its average score would be lowered if there were no missing data. Thus observing whether the omit rate for a subgroup is higher or lower than for another group gives an indication of whether the gap in constructed response score means would be larger or smaller if all data were present.

The situation for the Asian/white contrast is comparable to the male/female picture in that the *higher* scoring groups (males and Asians) have higher omit rates on the constructed response items. If all subgroup members had scores available, the Asian students would have somewhat lower average constructed response scores than are shown in the table, corresponding to a smaller advantage for Asian students, that is, a relative disadvantage of constructed response format for this group. For the other racial/ethnic minority groups the situation is reversed: the *lower* scoring group (Hispanic and black test takers) had higher omit rates. If they had no missing data, their average scores would be lower still. The effect would be to slightly increase the small

relative disadvantage of constructed response format for black students, and to decrease but not eliminate the relative advantage for Hispanics.

Table 6.6 shows the comparable statistics for the science test. Average scores on the constructed response science items were substantially lower than in mathematics, with a mean score of only 6.5 out of a possible 20 points. However, since there was no attempt to make the difficulty of the test items or the scoring algorithms comparable across the two subject areas, it would be incorrect to assume that student achievement in science, on some absolute scale, is lower than in mathematics. In other words, a score of 3 out of 5 on a test question does not necessarily correspond to a judgment of a particular level of competence in the subject area. It merely measures the quality of the student's response *on that item*, relative to a complete and correct answer. The skewed distribution of science scores must be considered in drawing conclusions from the score results.

Constructed response format appears to be relatively disadvantageous to females in the HSES science sample. Examination of results for individual items shows a large relative disadvantage for the first two test items, dealing with nuclear versus fossil fuels and eclipses, but not for the last two, an ecology item and one concerning a temperature graph.

Table 6.6:
Mean Science Scores, By Format and Subgroup
And Difference from Reference Group
in Standard Deviation Units (Effect Sizes)

	Multiple Choice		Constructed Response	
	Mean	S.D. Units	Mean	S.D. Units
Total (S.D.)	23.7 (6.6)		6.5 (3.9)	
Male	24.4		7.2	
Female	23.0	-21%	5.7	-38%
Asian	25.0	-4%	6.8	-11%
Hispanic	20.4	-74%	5.0	-59%
Black	19.2	-93%	4.2	-78%
White	25.3		7.3	

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

Asian students scored lower than whites on the constructed response science items, more so than could be attributed to their very slightly lower level of achievement on the multiple choice test. While the relative differences in performance between Asian and white test takers were small and not statistically significant, they

were replicated for each of the four test questions as well as for the total score. Black and Hispanic students, however, scored higher on the constructed response section, relative to whites, than their multiple choice test scores would have predicted. Part of this result may be due to a slight floor effect in the items, since average scores were low for all groups. Still, the relative format advantage appeared for each science item as well as for the total score.

Factor Structure

Given the high cost of constructed response testing in terms of administration time and scoring complexity, it is important to examine the benefits of this format relative to multiple choice tests. Preliminary factor analyses were conducted to determine whether the construct measured by the constructed response test questions was identifiably different from that of the multiple choice test.

The factor analyses in each subject area were performed on eight scores: the four constructed response scale scores, plus four scores based on subsets of the multiple choice items. The multiple choice test questions were grouped by content for this analysis, with number-right scores on the arithmetic, algebra, geometry, and data/probability/advanced topics items for the mathematics test, and life science, earth science, chemistry and physics scores on the science test. The mathematics factor analysis was restricted to the group of students who had taken the middle-difficulty mathematics form (over half of the sample) since the groupings of test items by content required that all students in the factor analysis received the same set of questions. All students took the same form of the science test.

Two distinct (although highly correlated) factors were identified in each of the two subject areas, and were associated with the two different test formats, that is, all of the constructed response questions had high factor loadings on one factor, and all of the multiple choice item subsets loaded on the other. Correlations of the multiple choice and constructed response factors with demographic variables showed similarities with the patterns found in the analysis of effect sizes (differences in standard deviation units) reported above. It must be remembered that there were slightly more unresolved missing scores on the constructed response questions for males than for females, and for black and Hispanic students than for whites. Thus, the constructed response format would appear to be slightly more advantageous to the group with the greater amount of missing data than is actually the case.

The tables below present the results of a confirmatory analysis of the factor structure of the two modes of measurement. The maximum likelihood (mle) confirmatory solution was used here in order to:

1) statistically reproduce the results of the exploratory solutions,

2) estimate the internal consistency reliabilities of the individual constructed response items and multiple choice item subsets,

3) arrive at a "true" score estimate of the correlations between the constructed response and multiple choice factors, and

4) extend selected demographic variables on the two-factor solution to see if the two question formats have differing relationships with background variables.

Table 6.7:
Confirmatory Factor Analysis
Mathematics

	Structure Coefficients		Reliability
	First Factor (CR)	Second Factor (MC)	
Constructed **Response:**			
Question 1	.60	--	.36
Question 2	.71	--	.50
Question 3	.73	--	.52
Question 4	.61	--	.37
Multiple **Choice Subsets:**			
Arithmetic	--	.80	.62
Algebra	--	.86	.71
Geometry	--	.72	.52
Data/Adv.	--	.62	.39
Factor Extension **Variables:**			
Female	-.09	-.08	
Hispanic	-.07	-.35	
Black	-.19	-.29	
Socioeconomic Status	.18	.43	

-- The confirmatory solution contrains these entries to be zero, that is, potential relationships other than those specified in the model are not calculated.

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

The results of the mathematics confirmatory solution shown in Table 6.7 suggest that the maximum likelihood estimates of the reliabilities of the single constructed response items are somewhat lower than the multiple choice parcels. The correlation between the two factors is .86. While this is relatively high, it still is low enough to suggest that while they share much in common, the two formats still have some unique variance. The extension coefficients in the table can be interpreted as the correlation between the factor "true" scores and either a continuous variable (socioeconomic status) or dummy coded variables (gender, Hispanic-white, and black-white comparisons). Inspection of the extension of the demographic characteristics on the two factor solution gives additional evidence for some unique measurement properties associated with each of the two factors. That is, while there is no difference between the gender extensions on the two factors, there are relatively large differences for the socioeconomic status and Hispanic-white comparisons. There is also a significant but smaller difference for black-white extensions. The negative sign of the extended Hispanic-white and black-white correlations indicates that in both cases the minority group is doing worse than the majority group. The greater size of the negative coefficient for the multiple choice factor shows that the minority groups are doing *differentially* worse in this format. The higher positive correlation for socioeconomic status on the multiple choice factor than the comparable loading on the constructed response factor indicates that students from high socioeconomic background do proportionately better on the multiple choice items than do students from low socioeconomic backgrounds.

Table 6.8:
Confirmatory Factor Analysis
Science

	Structure Coefficients		Reliability
	First Factor (CR)	Second Factor (MC)	
Constructed **Response:**			
Question 1	67	--	.45
Question 2	58	--	.34
Question 3	.63	--	.39
Question 4	.58	--	.33
Multiple Choice **Subsets:**			
Life Science	--	.73	.52
Earth Science	--	.77	.59
Chemistry	--	.77	.58
Physics	--	.71	.49
Factor Extension **Variables:**			
Female	-.21	-.12	
Hispanic	-.22	-.39	
Black	-.33	-.40	
Socioeconomic Status	.41	.56	

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National Center for Education Statistics.

The confirmatory solution for the science tests presented in Table 6.8 shows an even higher correlation between constructed response and multiple choice factors (.90). Patterns of format effects for population subgroups are similar to those found for mathematics: a relative advantage for Hispanic and, to a lesser extent, black students in constructed response format, while high socioeconomic status students tended to do better on the multiple choice tests. Unlike mathematics, where neither format appeared to be relatively advantageous for gender groups, females who took the science tests had a smaller score deficit on the multiple choice than on the constructed response section of the test. The reliabilities of the constructed response items were consistently lower than those of the multiple choice item subsets. It should be kept in mind that these reliabilities are internal consistency estimates based on a single factor underlying the constructed response items and a different but highly correlated factor underlying the multiple choice items.

These results suggest that whatever the common component of the four constructed response items is, it does have some unique reliable variance unrelated to the component underlying the multiple choice item subsets. The question that needs to be answered is whether or not the unique variance in the constructed response items is useful valid variance. This can be tested by studying, for example, whether the constructed response scores predict school achievement as well as the multiple choice items do.

Several generalizations about the interactions of format differences in the HSES tests with the gender and ethnicity of test takers are evident from examination of effect sizes for individual constructed response items as well as total scores, and for correlations of demographic dummy variables with factors:

- Females found some of the constructed response science items more difficult than did males. The score differences were greater than could be accounted for by differences in achievement as measured by the multiple choice science tests. No format differences in relative difficulty for the gender groups were found in the mathematics tests.

- Format differences did not have a substantial effect on the performance of the Asian students in the HSES sample. While some of the constructed response questions appear to be differentially more difficult for the Asian students, this effect is small and not completely consistent for all test items and analytic methods. The apparent differences may be due more to item context than to format.

- Hispanic constructed response test takers had less of a score deficit, relative to the white students in the sample, than would have been predicted by their scores on the multiple choice test. While this relative format advantage might be attenuated somewhat by a correction for missing data, it was found for each of the mathematics and science questions in this survey. This result should be interpreted with caution, however, since the field test of the same constructed response questions (prior to revisions) found a relative *disadvantage* of constructed response format for Hispanic students. The different findings may be due to differences in the samples, or to some other factor. A similar situation exists for the Advanced Placement tests taken by high school students, and administered by Educational Testing Service: analysis of performance differences on multiple choice sections compared to constructed response sections of the tests for gender and racial/ethnic subgroups has detected significant differences, but the patterns of differences are inconsistent.

- The effect of format differences for black versus white students on the mathematics test was inconclusive. Analysis of effect sizes indicated no format difference, while factor analysis results suggest a small constructed response format advantage for black students. Differences in results may be related to the necessity of restricting the factor analysis sample to students who took the middle difficulty form of the mathematics test. If a constructed response format advantage operates primarily for low-achieving students, fewer of them were present in the factor analysis sample. On the science test, an apparent reduction in the size of the black-white score gap for constructed response items may be partly due to differential omit rates, and partly to floor effects. Whether or not corrections for these factors would eliminate the apparent advantage entirely is inconclusive.

Exploration of the language background and use variables and transcript records in the data files described in Appendix E may be useful in explaining some of the ethnic group differences in performance.

Correlations

The relationships of constructed response test scores and omit rates with student background characteristics and achievement as measured by the multiple choice tests have been documented earlier in this report. Correlation analysis of these variables supports earlier conclusions concerning higher omit rates for low achieving students and members of racial/ethnic minority groups. Strong correlations between scores on the constructed response test and scores on *all four* NELS:88 multiple choice tests were also found. In each sample,

the total constructed response score correlated most strongly with the corresponding multiple choice section (.82 for math, .70 for science). In both samples, the correlation of multiple choice mathematics with science was .80. It was not possible to determine the relationship between constructed response scores in mathematics and science since each student received constructed response questions in only one of the subject areas. Tables of correlation coefficients are included in Appendix D.

It is important to remember that the size of correlation coefficients is constrained by the reliability of the measurements. Two aspects of the HSES constructed response tests limit their reliability and thus tend to attenuate the size of correlation coefficients. The short test length, 4 items in each subject, severely limits the coverage of items in the content domain. And the constructed response format is dependent on human scorers, with the possibility of unreliability of scores due to differences in reader judgment. Both of these considerations have been discussed at length in the earlier section on reliability. They are noted again here in order to point out that correlations of constructed response scores with other variables would be somewhat higher without these constraints. The one exception to this is the confirmatory factor analysis where the relationship between the constructed response items and the background variables is corrected for the unreliability of the constructed response items.

In the NELS:88 mathematics and science multiple choice tests, clusters of test questions *were* selected that marked distinct levels of proficiency in skills within the content area. Five such levels were identified in the mathematics test, and three in science. *The* levels were shown to follow a building-block pattern, that is, proficiency at a higher level implied mastery of the skills at all lower levels. The development and scaling of these scores is documented in the *NELS:88 Second FollowUp Student Component Data File User's Manual*, *as* well *as in the Psychometric Report for the NELS:88 Base Year Through Second Follow Up. The* correlation coefficients in Table 6.9 show the relationships between mastery of these hierarchical proficiency levels and the total score on the corresponding constructed response mathematics or science test.

Table 6.9:
Correlations of Proficiency Level with Constructed Response Total Score

Proficiency Level	Constructed Response Total Score	
	Mathematics	Science
Level 1	.44	.41
Level 2	.64	.65
Level 3	.72	.63
Level 4	.77	(none)
Level 5	.43	(none)

SOURCE: National Education Longitudinal Study of 1988 (NELS:88), Second Followup Survey, National *Center for* Education Statistics.

For the mathematics test, performance on the constructed response test is most closely identified with mastery of proficiency levels 2, 3, and 4 (operations with decimals, fractions, powers and roots; simple problem solving, requiring the understanding of low level mathematical concepts; and intermediate level concepts/multi-step solutions to word problems). Levels 1 and 5 (simple arithmetical operations on whole numbers; and complex problem solving linked to knowledge of mathematics material found in advanced mathematics courses) were less highly correlated with the constructed response tests, primarily because the content of the constructed response questions overlapped most closely with the difficulty of the middle levels. (The extreme splits observed for the lowest and highest proficiency levels would preclude high correlations in any case.) While competence in arithmetic was necessary to solve the constructed response problems, it was not in itself sufficient. At the other end of the scale, high achieving mathematics students did tend to score higher on the constructed response tests. However, the test items did not require advanced mathematics, and students at a somewhat lower level of proficiency could perform nearly as well.

The science tests showed a similar pattern. Performance on level 1 science tasks (understanding of everyday science concepts; "common knowledge" that can be acquired in everyday life) was significantly correlated with the constructed response total score. Relationships were even stronger with the two highest science proficiency levels (understanding of fundamental science concepts upon which more complex science knowledge can be built; and understanding of relatively complex scientific concepts, typically requiring an additional problem solving step). The constructed response science questions were not dependent on content of advanced level science courses such as physics and chemistry.

Student Reactions

Students' self report of their performance, in addition to providing a basis for score imputation, may be useful as a guide in designing constructed response questions for low-risk survey tests in the future. The HSES test takers were asked to provide feedback on the difficulty, clarity and timing of the questions, as well as on their perceptions of their performance. Response rates for the reaction questions were quite high, with about 95 percent of the sample responding to most of the questions. Omit rates tended to be higher for the questions at the end of the test forms, and were also somewhat higher for black students than for other subgroups. Appendix A contains the complete text of the student reaction questions. Tables of students' responses, broken down by gender and racial/ethnic group, may be found in Appendix D, and are summarized below.

"How hard was the question ?"

Test developers and advisors feared that the constructed response tests would be too easy for a sample of high school seniors. This did not prove to be the case. In addition to the evidence provided by the scaled scores (no clustering of students at the top of the total scale score distribution), the students' self report indicated that the questions were of appropriate difficulty. For a majority of the test questions in both mathematics and science, and for most of the gender and racial/ethnic subgroups examined, the most frequently chosen response to the difficulty question was "about right. " With the exception of one mathematics and one science question, more students indicated that each question was "hard" or "too hard" than "easy" or "too easy. " Asian students tended to report that the mathematics (but not the science) questions were too easy, while a larger proportion of Hispanic and black test takers than the other racial/ethnic groups found the questions hard or too hard. In general, the perceived difficulty of the science questions tended to be higher than the mathematics problems.

"How good** was your **answer?"

Students tended to choose the extremes in responding to this question ("really didn't know how to answer," or "gave a pretty good answer") in preference to the middle option ("partly right") more often than was justified by their actual performance. For most of the test questions there were fewer zero (and imputed-zero) scores than students who said they didn't know how to answer. At the other end of the scale, more students thought that they gave a "pretty good answer" than actually received a score of 4 or 5 on each question. Differences between the mathematics and science tests appear to be related to the higher mean and wider spread of scores in mathematics compared to science, which in turn is probably a consequence of the stepwise structure of the mathematics test items.

There were substantial gender differences in students' perceptions of their answers to the constructed response questions. For all four mathematics and all four science questions, a much higher proportion of females than males said they really didn't know how to answer the questions, and many more males than females thought they gave a pretty good answer. While the differences in performance (actual scores) *did, in* fact, favor males, the score differences were relatively small compared to the differences in self-evaluations.

The tendency to overestimate performance appears to be somewhat greater for black test takers as well as for males, particularly in the mathematics test. Systematic analysis of the self report versus actual performance data in conjunction with other variables may reveal whether or not the apparent gender and racial/ethnic group differences in perceptions are related to differences in the courses taken or schools attended by members of different subgroups.

"Have you** taken the courses you would **need to answer the question?"

A majority of test takers reported having had enough background in their school coursework to answer each of the mathematics questions. Hispanic and black students were more likely than whites to feel unprepared for the questions, with about one-third to one-half of students in these subgroups indicating that they had not taken the courses needed to solve the mathematics problems. Fewer students felt prepared to answer the science questions-about half of all test takers did not feel that they had the necessary background for three of the four questions. Subgroup differences in response to the question about course background were generally fairly small for *science* test takers. Transcript records are available for further study of comparisons of actual course taking patterns with students' self report of adequate preparation.

"Did you understand the question?"

Students who took the mathematics test did not seem to be making a distinction between difficulty and clarity in answering this question. There was a close correspondence between the number of test takers who found the question "a little confusing" or "very confusing" and those who had said it was "hard" or "too hard" (up to about half of the sample). For most questions, this was also about the same number of students who indicated, "No, I have not taken the courses needed to answer the question." This similarity of responses suggests that their lack of understanding was probably related more to insufficient mastery of the material than to flaws in the question design. The pattern of responses was similar for the last two science questions, which were relatively technical and had diagrams as part of the question stem (as did the mathematics questions). The first two science questions, on the other hand, had fairly short stems that consisted only of text. Only about a quarter of test takers thought these questions were unclear, although closer to half of the group found them difficult.

Comparison of students' perceptions with their scores on the last science question, however, suggests that this test question (heating curve) may *not* have made clear to the test takers what was expected of them, although they thought it did. Only about a quarter of the test takers reported finding the question unclear or difficult. But freer than 25 percent gave a reasonably complete answer to the question (scores of 3 or more on the 0-5 scale). In fact, of the students who thought they gave a "pretty good answer," about 40 percent actually demonstrated little or no understanding of the concept.

"Did you have enough time to answer the question?"

The constructed response items were "paced," that is, separately timed, at 10 minutes each. In constructed response format, there is the potential for students to get bogged down in writing a much more complex response than test designers anticipated, and thus to jeopardize their ability to finish the rest of the test. It then becomes impossible to tell whether unanswered items at the end of the test were too difficult, or whether the student simply ran out of time. To avoid this problem, students were told when the time was up for each question, and were instructed to move onto the next one. Tabulations of the student reaction questions showed that the 10 minutes allotted for each question was adequate. Nearly half of the test takers responded that the timing was "about right," with more students saying that too much time was allowed than not enough. Most students could probably have finished each item in a slightly shorter time, perhaps 8 minutes. However, nearly 20 percent of black and Hispanic students reported that 10 minutes was not long enough for several of the test questions. This assessment was intended to be a "power" test rather than a speed test, that is, it was designed to measure how much students could do rather than how quickly they could do it. It was important to ensure that time constraints did not adversely affect test scores for some subgroups and thus contaminate interpretation of subgroup differences in performance.

Chapter 7: Summary/Conclusions/Recommendations

The methodological experiment described in this report was designed to investigate issues in constructed response test design, administration, scoring and interpretation in the context of a large-scale, voluntary national survey. The study investigated practical issues such as communication, nonresponse, time, and cost, as well as psychometric issues including reliability, factor structure, and differential subgroup performance. The major findings from analysis of the mathematics and science constructed response test results in the High School Effectiveness Study are summarized below.

In deciding whether these results are applicable to other settings, it is important to consider how similarities or differences in the major features of the High School Effectiveness Study compared to other tests may impact results. HSES tests were low-risk: the test takers knew that their scores would not be reported to their schools, parents, or teachers, or even to themselves, which may have affected their motivation to try to give their best answers to the questions. The participants were twelfth grade students selected without regard to their course-taking history or future educational plans, so the tests had to be written to accommodate a wide range of achievement. The tests were given to students across the nation who were strangers to the test writers and scorers, so it was essential that the questions be explicit enough that answers could be evaluated without any extraneous information about what was required. The score scales were designed to measure only competence in mathematics or science, and not other factors such as writing ability or effort. To the extent that a classroom test or a college entrance exam may differ from this survey in incentives to answer, homogeneity of the test takers, acquaintance of test takers with test givers, or measurement objectives, it is necessary to consider how results might differ from those found in the High School Effectiveness Study.

Omit Rates. Constructed response test questions require more effort than multiple choice questions. In a low-risk setting, test takers may not be willing to give the extra effort required. In the High School Effectiveness Study, omit rates for constructed response questions were consistently higher than for the multiple choice tests in the same subject area. The Asian, black and Hispanic students in the HSES sample were more likely than white students to omit constructed response items, although subgroup differences in multiple choice response rates were small. Unanswered items present a particular problem on a low-risk test, since it is not appropriate to score "zero" or "no credit" when students have no incentive to attempt to answer. It is therefore desirable to minimize the amount of missing test data by:

- attempting to induce students to give their best answers by "selling" them on the value of their participation, and by making test questions interesting and relevant, especially for members of racial/ethnic minority groups.

- making each test question accessible to all test takers at some level, using a stepwise design and non-technical language as much as possible, while still managing to convey the information that a technical response is required for full credit if that is the case.

- making it convenient for students who really don't know the answer to demonstrate their lack of knowledge (perhaps by simply checking a box that says "I don't know how to answer this question") rather than simply leaving the question blank.

- planning in advance for an imputation scheme for missing items or parts of items that takes into account, if possible, corollary information such as coursework, grades, performance on other test questions, or self-evaluations of ability to respond. Evidence from multiple choice test

scores and self-reports in this sample demonstrates that scoring all omits as wrong is inappropriate.

Reliability. The NELS:88 HSES constructed response tests had somewhat lower levels of reliability than the multiple choice tests in the same subject area. Combining the formats resulted in a slight increase in reliability over that for the multiple choice items alone, but not as great an increase as could have been achieved by adding several more multiple choice items. Two factors that may contribute to lower reliability for constructed response test questions are:

- reader reliability-the possibility of different readers giving different scores to the same answer. Reader reliability imposes an upper limit on the overall reliability (consistency of measurement) that can be achieved by a constructed response test. Problems may be minimized by making questions and scoring criteria as explicit and unambiguous as possible. Second readings obtained for field test samples are useful in identifying and correcting aspects of the questions and scoring procedures with a high potential for difficulties.

- domain coverage—the longer time required for the HSES constructed response questions compared to multiple choice meant that many fewer items could be given in the same period of time. Limited coverage of possible question topics may result in measurements that are too greatly influenced by the content of particular questions rather than being a reliable measure of overall mathematics or science achievement.

Analysis of Scores. Average scores for males were higher than for females in both multiple choice and constructed response format, and in both mathematics and science. The white students in the HSES sample scored higher, on average, than the Hispanic and black students in both formats and both content areas.

Correlations of the constructed response tests with multiple choice test total scores in the same subject were high However, factor analysis of the tests did reveal separate (although highly correlated) factors for the two item formats. The constructed response format appears to have been relatively advantageous for HSES Hispanic students in both mathematics and science, and to a lesser extent for black test takers, although HSES field test results and analysis of group differences on Advanced Placement tests have found a great deal of inconsistency in relative format advantage for racial/ethnic groups. Students of high socioeconomic status tended to do relatively better on multiple choice items. Gender differences and contrasts between Asian and white students were inconsistent and may be due to interactions with item content. Evaluation of the size of the format effect is complicated by nonresponse rates that differ for students of different ability levels and racial/ethnic groups, and by a possible floor effect in the science test.

Just as constructed response format provides test takers the opportunity to respond in many different ways, it also allows the test user to judge the value of the responses according to any arbitrary set of criteria. Had the scoring scales been designed differently, for example, giving weight to features such as writing style, other factors and subgroup differences might have emerged.

Scoring Costs. The greatest single constraint on the use of constructed response questions in the HSES survey (in addition to administration time) was the cost of scoring. Unlike multiple choice questions, which can be scored by computer at negligible cost, constructed response questions must be read individually by human readers with some expertise in the test content. A rough estimate of the cost of scoring the HSES constructed response questions is approximately $2 per test item per student. This includes the cost of recruiting, training and supervising the readers, and of preparing data files of the analytic scores. It does not include the higher cost

(relative to multiple choice) of developing the items, or of developing analytic scoring procedures and building and evaluating score scales. Per-item costs might be reduced somewhat in a larger-scale survey; however, economies of scale might be offset by the necessity of recruiting readers from a wider area, which would add travel and maintenance costs in addition to reader stipends.

Constructed response tests are time consuming to administer and expensive to score. However, they may provide diagnostic information and measurements of skills that are difficult to evaluate with multiple choice questions. Choices of appropriate test format must be based on the constructs to be measured and the interpretations that will be made from the scores.

Bibliography

Bennett, R.E. and Ward, W. C., eds. 1993. *Construction Versus Choice in Cognitive Measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Collis, K. F., and Romberg, T.A. 1992. *Mathematical Problem Solving Profiles.* Melbourne, Victoria, Australia: Australian Council for Educational Research.

Collis, K. F., Romberg, T.A., and Jurdak, M.E. May 1986. "A Technique for Assessing Mathematical Problem-Solving Ability." *Journal for Research in Mathematics Education* 17(3): 206-221.

Dorans, N. J., and Schmitt, A.P. "Constructed Response and Differential Item Functioning: A Pragmatic Approach." In *Construction Versus Choice in Cognitive Measurement, 1993* Eds. R.E. Bernet and W.C. Ward. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dowd, K.L., et al. 1991. *NELS:88 Second FollowUp Field Test Report.* Chicago: NORC. ERIC ED 335-418.

Hartle, T. W., and Battaglia, P.A. 1993. "The Federal Role in Standardized Testing. In *Construction Versus Choice in Cognitive Measurement.* Eds. R.E. Bennet and W.C. Ward. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ingels, S. J., et al. 1994. *NELS:88 Second FollowUp: Student Component Data File User's Manual.* Washington, DC: NCES 93-374.

Morgan, R Personal communication concerning relative gender and racial/ethnic subgroup differences in performance on multiple choice versus constructed response sections of Advanced Placement tests, Educational Testing Service.

Robinson, S.P. 1993. "The Politics of Multiple-choice Versus Free-response Assessment. In *Construction Versus Choice in Cognitive Measurement.* Eds. R.E. Bennet and W.C. Ward. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rock, D. A., and Pollack, J. M. 1995. *NELS:88 Base Year through Second Follow Up Psychometric Report.* Washington, DC: NCES 95-382.

Romberg, T.A. 1982. *The Development and Validation of a Set of Mathematical Problem-Solving Superitems.* Wisconsin Center for Education Research, University of Wisconsin.

Swinton, S. 1993. *Differential Response Rates to Open-Ended and Multiple-Choice NAEP Items By Ethnic Groups.* Paper Presented at AERA Annual Meeting, Atlanta, GA.

Traub, R.E. 1994. *Reliability for the Social Sciences, Theory and Applications.* Thousand Oaks, CA: Sage Publications.

Appendix A

- Test items
- Student Reaction Questions
- Analytic Scores
- Scale Scores

NORC - 4521
Form Approved
OMB No. 1850-0652
App. Exp.: 7/92

NATIONAL EDUCATION LONGITUDINAL STUDY OF 1988

SECOND FOLLOW-UP

SCHOOL EFFECTS SUPPLEMENT FREE RESPONSE TEST

SCIENCE

Prepared for the U.S. Department of Education
National Center for Education Statistics

By the National Opinion Research Center (NORC),
A Social Science Research Center
at the University of Chicago

Science Free Response

4 Questions

10 Minutes Each

Each of the following questions has several parts. Write your answers in the space provided. Answer each part as completely as you can. After you have finished your work, answer the brief questionnaire following each question.

Question 1.

Fossil fuels (such as coal, oil, and natural gas) and nuclear fuels are both used to generate electricity. Compare the use of nuclear fuels to the use of fossil fuels, including in your discussion at least one advantage and one disadvantage of@ of these two types of fuel.

For each of the following, circle the phrase that best describes how you did on this question.

1. How hard was the question?

   (A) Too easy
   (B) Easy
   (C) About right
   (D) Hard
   (E) Too hard

2. How good was your answer?

   (A) I really didn't know how to answer the question.
   (B) My answer was partly right.
   (C) I think I gave a pretty good answer.

3. Have you taken the courses you would need to answer the question?

   (A) Yes, I have had enough background in my coursework.
   (B) No, I have not taken the courses needed to answer the question.

4. Did you understand the question?

   (A) It was very clear.
   (B) It was clear enough.
   (C) It was a little confusing.
   (D) It was very confusing.

5. Did you have enough time to answer the question?

   (A) Not enough time at all
   (B) Could have used a little more time
   (C) About the right amount of time
   (D) A little too much time
   (E) Way too much time

Question 2.

(A) Draw a diagram below of the relative positions of the Earth, Moon, and Sun during a solar eclipse. Label your diagram.

(B) Draw a diagram below of the relative positions of the Earth, Moon, and Sun during a lunar eclipse. Label your diagram.

(C) Explain why a lunar eclipse can be seen from a greater geographic area on the Earth than a solar eclipse can.

For each of the following, circle the phrase that best describes how you did on this question.

1. How hard was the question?

   (A) Too easy
   (B) Easy
   (C) About right
   (D) Hard
   (E) Too hard

2. How good was your answer?

   (A) I really didn't know how to answer the question.
   (B) My answer was partly right.
   (C) I think I gave a pretty good answer.

3. Have you taken the courses you would need to answer the question?

   (A) Yes, I have had enough background in my coursework.
   (B) No, I have not taken the courses needed to answer the question.

4. Did you understand the question?

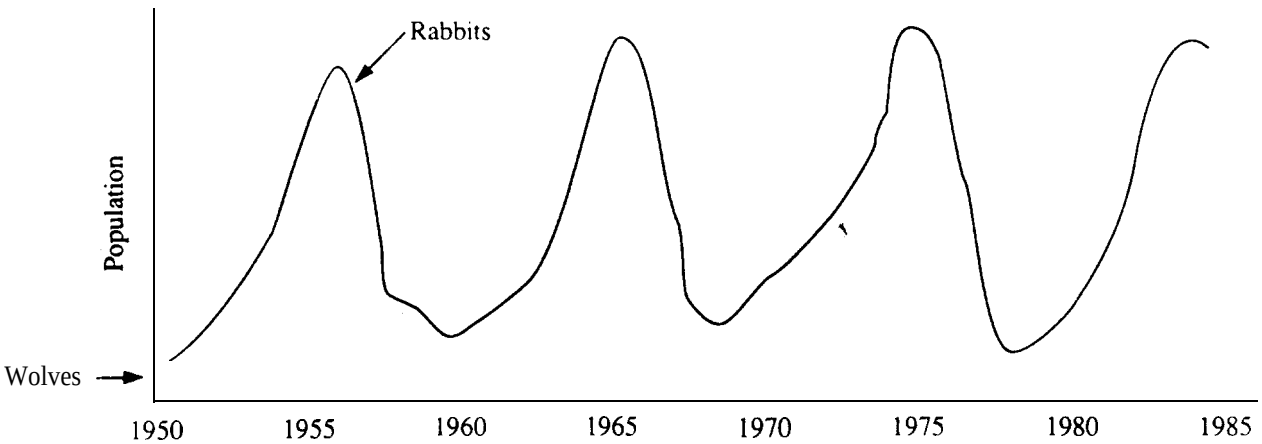
   (A) It was very clear.
   (B) It was clear enough.
   (C) It was a little confusing.
   (D) It was very confusing.

5. Did you have enough time to answer the question?

   (A) Not enough time at all
   (B) Could have used a little more time
   (C) About the right amount of time
   (D) A little too much time
   (E) Way too much time

Question 3.

A particular species of rabbit is infected with a virus that only affects rabbits, and that is only active when the population of rabbits reaches a specific density. The virus kills most of the rabbits at fairly regular intervals, as shown on the graph below. The rabbits share their ecosystem on an isolated island with a species of wolf for which the rabbit is the predominant prey. On the same graph below, draw a curve that might reasonably represent the population of wolves over the same time period, starting with the population point given for the year 1950.



On the lines below, briefly explain the reasons for the height and the position of the curve you drew, compared to the rabbit curve.

For each of the following, circle the phrase that best describes how you did on this question.

1. How hard was the question?

   (A) Too easy
   (B) Easy
   (C) About right
   (D) Hard
   (E) Too hard

2. How good was your answer?

   (A) I really didn't know how to answer the question.
   (B) My answer was partly right.
   (C) I think I gave a pretty good answer.

3. Have you taken the courses you would need to answer the question?

   (A) Yes, I have had enough background in my coursework.
   (B) No, I have not taken the courses needed to answer the question.

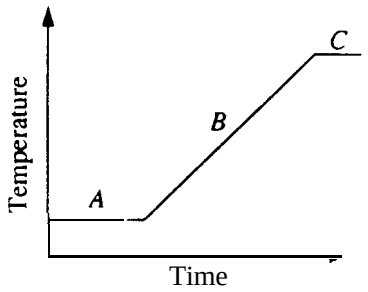
4. Did you understand the question?

   (A) It was very clear.
   (B) It was clear enough.
   (C) It was a little confusing.
   (D) It was very confusing.

5. Did you have enough time to answer the question?

   (A) Not enough time at all
   (B) Could have used a little more time
   (C) About the right amount of time
   (D) A little too much time
   (E) Way too much time

Question 4.



A beaker contains a mixture of water and ice. A thermometer is placed in this mixture, and the mixture is continuously stirred as it is heated to boiling over a flame. At regular intervals, the temperature of the mixture is recorded. These data are then used to produce the graph above. In the space provided below, briefly explain the appearance of each labelled section of the curve.

Why is the temperature constant in section A of the curve even though heat is being added?

Why does section B of the curve slope upward?

Why is the temperature constant in section C of the curve?

For each of the following, circle the phrase that best describes how you did on this question.

1. How hard was the question?

   (A) Too easy
   (B) Easy
   (C) About right
   (D) Hard
   (E) Too hard

2. How good was your answer?

   (A) I really didn't know how to answer the question.
   (B) My answer was partly right.
   (C) I think I gave a pretty good answer.

3. Have you taken the courses you would need to answer the question?

   (A) Yes, I have had enough background in my coursework.
   (B) No, I have not taken the courses needed to answer the question.

4. Did you understand the question?

   (A) It was very clear.
   (B) It was clear enough.
   (C) It was a little confusing.
   (D) It was very confusing.

5. Did you have enough time to answer the question?

   (A) Not enough time at all
   (B) Could have used a little more time
   (C) About the right amount of time
   (D) A little too much time
   (E) Way too much time

NORC - 4521
Form Approved
OMB No. 1850
App. Exp.: 7/9

NATIONAL EDUCATION LONGITUDINAL STUDY OF 1988

SECOND FOLLOW-UP

SCHOOL EFFECTS SUPPLEMENT FREE RESPONSE TEST

MATHEMATICS

Prepared for the U.S. Department of Education
National Center for Education Statistics

By the National Opinion Research Center (NORC),
A Social Science Research Center
at the University of Chicago

Mathematics Free Response

4 Questions

10 Minutes Each

Each of the following questions has several parts. Write your answers in the space provided. Answer each part as completely as you can. After you have finished your work, answer the brief questionnaire following each question.

Question 1.

SUMMER TRAIN SCHEDULE FOR **TRAINS** GOING
FROM CITY A TO CITY B

Train #	Leave City A	Arrive City B
#1	6:05 a.m.	6:50 a.m.
#2	6:55	7:40
#3	7:23	8:12
#4	7:42	8:17
#5	8:03	8:43
#6	9:20	10:05
#7	10:35	11:20
#8	11:35	12:20 p.m.
#9	2:08 p.m.	2:53

(A) In the **summer,** what is the latest train from City A you can get if you want to reach City B by 11:30 a.m.?

SHOW YOUR WORK **HERE**:

Answer: The latest train I can get if I want to reach City B by 11:30 a.m. is train # _____

(B) In the **summer,** what train from City A should you take if you want to spend the least amount of time traveling from City A to City **B?**

SHOW YOUR WORK **HERE**:

Answer: The train that spends the **least** amount of time traveling from City A to City B is train # _____

GOON TO THE NEXT PAGE

(C) A person whose home is 30 minutes from the City A train station has an appointment in City B at 1:30 p.m. The appointment is 20 minutes from the City B train station. If it is during the summer, what *is the* latest time that the person can choose to leave home for this appointment?

SHOW YOUR WORK HERE:

Answer: The latest time the person can choose to leave home for this appointment is _____

(D) During the winter months:

(i) Trains take 10 percent more time to go from City A to City B.
(ii) People prefer that trains leave City A 5 minutes later than in the summer.

These factors are to be taken into account in making up the winter train schedule.

Let t = the time a train leaves City A in the summer
y = the time, in minutes, it takes a train to travel from City A to City B in the summer.

Write an algebraic expression, using t and y, which can be used to calculate the time a train arrives in City B in the winter.

SHOW YOUR WORK HERE:

Answer: The time a train arrives in City B in the winter = _____

GOON TO THE NEXT PAGE

For each of the following, circle the phrase that best describes how you did on this question.

1. How hard was the question?

   (A) Too easy
   (B) Easy
   (C) About right
   (D) Hard
   (E) Too hard

2. How good was your answer?

   (A) I really didn't know how to answer the question.
   (B) My answer was partly right.
   (C) I think I gave a pretty good answer.

3. Have you taken the courses you would need to answer the question?

   (A) Yes, I have had enough background in my coursework.
   (B) No, I have not taken the courses needed to answer the question.

4. Did you understand the question?

   (A) It was very clear.
   (B) It was clear enough.
   (C) It was a little confusing.
   (D) It was very confusing.

5. Did you have enough time to answer the question?

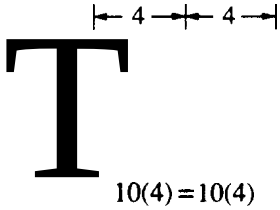
   (A) Not enough time at all
   (B) Could have used a little more time
   (C) About the right amount of time
   (D) A little too much time
   (E) Way too much time

Question 2.

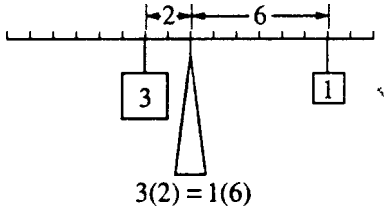
Below are some diagrams of a 16-foot long beam which is centered over a pivot.

- The dash marks are at one-foot distances along the beam.
- Each box is a weight attached to the beam, and the numbers indicate the weight, in pounds, of each box.

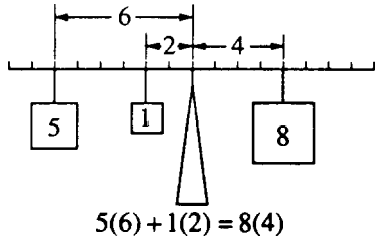
IN BALANCE



10(4) = 10(4)

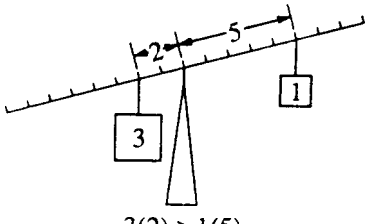


3(2) = 1(6)

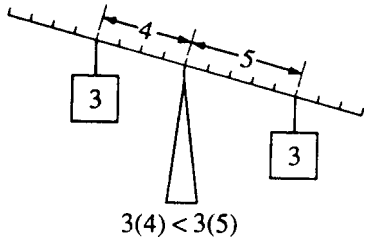


5(6) + 1(2) = 8(4)

OUT OF BALANCE

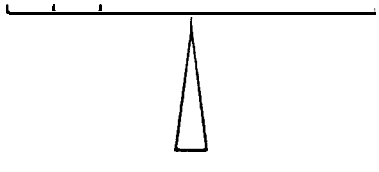


3(2) > 1(5)



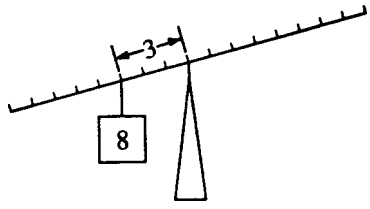
3(4) < 3(5)

(A) Draw two 9-pound weights attached to the beam so that the beam will be in balance. Label the 9-pound weights and their distance from the pivot.



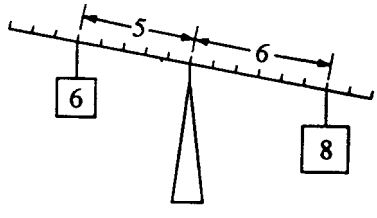
(B) Draw **one** 4-pound weight to balance the beam. Label the 4-pound weight and its distance from the pivot.

SHOW YOUR COMPUTATION HERE:



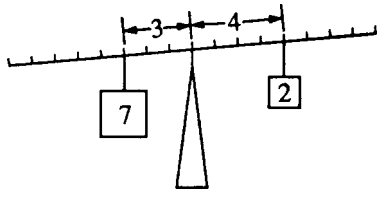
(C) Draw **one** additional 6-pound weight so that the beam will be in balance. Label the 6-pound weight and its distance from the pivot.

SHOW YOUR COMPUTATION HERE:



(D) The beam can be balanced by placing **one** additional weight of **x** pounds at a distance of **y** feet to the right side of the center of the beam. Find an equation which shows the relationship between **x** and **y**.

SHOW YOUR WORK HERE:



My equation is: _____

GOON TO THE NEXT PAGE

For each of the following, circle the phrase that best describes how you did on this question.

1. How hard was the question?

   (A) Too easy
   (B) Easy
   (C) About right
   (D) Hard
   (E) Too hard

2. How good was your answer?

   (A) I really didn't know how to answer the question.
   (B) My answer was partly right.
   (C) I think I gave a pretty good answer.

3. Have you taken the courses you would need to answer the question?

   (A) Yes, I have had enough background in my coursework.
   (B) No, I have not taken the courses needed to answer the question.

4. Did you understand the question?

   (A) It was very clear.
   (B) It was clear enough.
   (C) It was a little confusing.
   (D) It was very confusing.

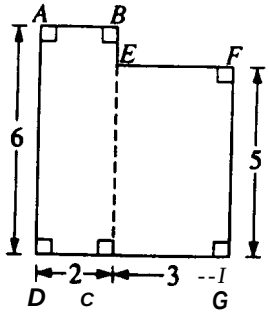
5. Did you have enough time to answer the question?

   (A) Not enough time at all
   (B) Could have used a little more time
   (C) About the right amount of time
   (D) A little too much time
   (E) Way too much time

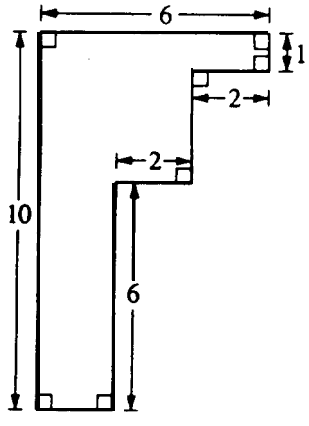
Question 3.

To find the area of a figure made up of two or more **rectangles,** we can find the area of each rectangle and add the areas **together.** For **example:**

Area of **rectangle** $ABCD = 6 \times 2 = 12$ square units
Area of rectangle $CEFG = 3 \times 5 = 15$ square units
Area of figure $= 12 + 15 = 27$ square **units**



(A) Draw lines in the figure **below** to show that it is made up of several **rectangles.**



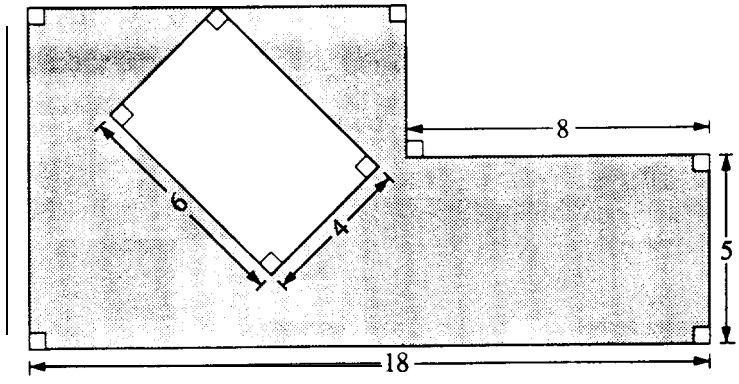
(B) What is the area of the figure in **(A)?**

SHOW YOUR WORK **HERE:**

Answer: The area of the figure in **(A)** is _____ square units.

GO ONTO THE NEXT PAGE

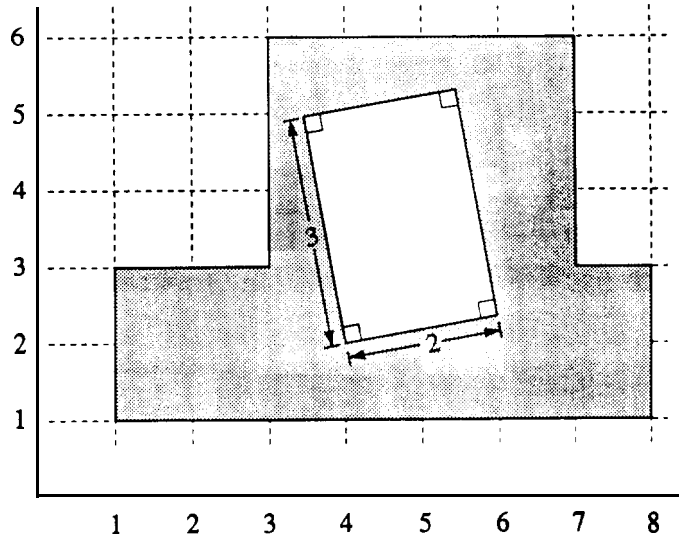
(C) What is the area of the shaded region in the figure **below?**



SHOW YOUR WORK **HERE**:

Answer: The area of the shaded region is square **units.**

(D) What is the area of the shaded region in the figure **below?**



SHOW YOUR WORK **HERE**:

Answer: The area of the shaded region is _____ square **units.**

GOON TO THE NEXT PAGE

For each of the following, circle the phrase that best describes how you did on this question.

1. How hard was the question?

   (A) Too easy
   (B) Easy
   (C) About right
   (D) Hard
   (E) Too hard

2. How good was your answer?

   (A) I really didn't know how to answer the question.
   (B) My answer was partly right.
   (C) I think I gave a pretty good answer.

3. Have you taken the courses you would need to answer the question?

   (A) Yes, I have had enough background in my coursework.
   (B) No, I have not taken the courses needed to answer the question.

4. Did you understand the question?

   (A) It was very clear.
   (B) It was clear enough.
   (C) It was a little confusing.
   (D) It was very confusing.

5. Did you have enough time to answer the question?

   (A) Not enough time at all
   (B) Could have used a little more time
   (C) About the right amount of time
   (D) A little too much time
   (E) Way too much time

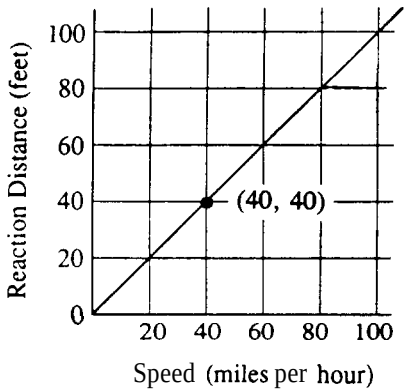
Question 4.

The distance a car travels after the driver has decided to stop (stopping distance) is related to how fast the car was moving. The graphs below show how the components that make up stopping distance increase for faster speeds.

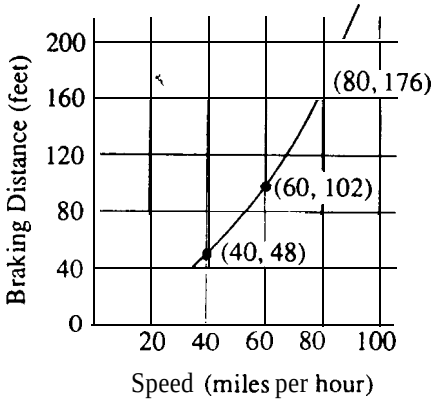
The distance a car travels from the time its driver first decides to apply the brakes until the driver actually applies the brakes is shown in the graph labelled Reaction Distance.

The distance the car travels from the time the brakes are applied until it comes to a complete stop is shown in the graph labelled Braking Distance.

REACTION DISTANCE



BRAKING DISTANCE



Example: According to the graphs above, if a car is traveling at 40 miles per hour the driver's reaction distance is 40 feet and the car braking distance is 48 feet.

(A) What is the driver's reaction distance for a car travelling at 80 miles per hour?

Answer: _____

What is the braking distance for this car?

Answer: _____

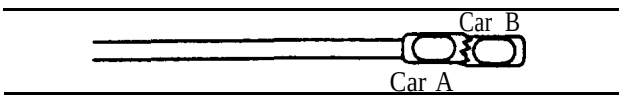
(B) A car is travelling at 60 miles per hour. How far will the car travel from the time its driver first decides to apply the brakes until it comes to a complete stop.

SHOW YOUR WORK HERE:

Answer: _____

GO ON TO THE NEXT PAGE

(C)



In the diagram of a collision between Car A and Car B shown above, the skid marks of Car A's tires are about 100 feet long. (Note: The skid marks made by Car A did not begin until after its driver had actually applied the brakes).

What is the closest that Car A could have been to the collision point when its driver first decided to apply the brakes?

SHOW YOUR WORK HERE:

Answer: _____

(D) Explain why Car A might have been farther away than the answer you gave above.

GOON TO THE NEXT PAGE

For each of the following, circle the phrase that best describes how you did on this question.

1. How hard was the question?

   (A) Too easy
   (B) Easy
   (C) About right
   (D) Hard
   (E) Too hard

2. How good was your answer?

   (A) I really didn't know how to answer the question.
   (B) My answer was partly right.
   (C) I think I gave a pretty good answer.

3. Have you taken the courses you would need to answer the question?

   (A) Yes, I have had enough background in my coursework.
   (B) No, I have not taken the courses needed to answer the question.

4. Did you understand the question?

   (A) It was very clear.
   (B) It was clear enough.
   (C) It was a little confusing.
   (D) It was very confusing.

5. Did you have enough time to answer the question?

   (A) Not enough time at all
   (B) Could have used a little more time
   (C) About the right amount of time
   (D) A little too much time
   (E) Way too much time

Student Reaction Questions

The following five questions were answered after each math or science problem. Codes in the database are alphabetic.

1. How hard was the question?
   A. Too easy
   B. Easy
   C. About right
   D. Hard
   E. Too hard

2. How good was your answer?
   A. I really didn't know how to answer the question.
   B. My answer was partly right.
   C. I think I gave a pretty good answer.

3. Have you taken the courses you would need to answer the question?
   A. Yes, I have had enough background in my coursework.
   B. No, I have not taken the courses needed to answer the question.

4. Did you understand the question?
   A. It was very clear.
   B. It was clear enough.
   C. It was a little confusing.
   D. It was very confusing.

5. Did you have enough time to answer the question?
   A. Not enough time at all
   B. Could have used a little more time
   C. About the right amount of time
   D. A little too much time
   E. Way too much time

Analytic Scores: Math Question 1 (Train Schedule)

Students were given a train schedule and asked to select the trains that met various time criteria, to figure out how much time to allow for a trip counting travel time before and after the train trip, and to write an equation for a transformation of the train schedule to allow for 5 minute later departure times and 10% increase in travel time.

— A. What is the latest train that will get to City B by **11:30 a.m.?**

O = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Any incorrect train number or time
4 = the train that arrives at 11:20; or the train that leaves at 10:35; or train #7 (all correct)

— B. What train takes the least amount of **time?**

O = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not **attempting** to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Any incorrect train number or time; no work shown
4 = Any incorrect train number or time; work shown but not appropriate procedure
5 = Any incorrect train number or time: correct procedure but contains arithmetic error(s), including not knowing how to subtract hours and minutes
6 = the train that arrives at 8:17; or the train that leaves at 7:42; or train #4 (all correct)

— C. What is the latest time the person can leave home for this **appointment?**

0 = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not **attempting** to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Any incorrect time; no work shown
4 = Any incorrect time; work shown but procedure was incorrect
5 = Any incorrect time; correct procedure but contains arithmetic error(s), including not knowing how to subtract hours and minutes
6 = 11:35 (correct train, but omitted travel time to station in City A)
7 = 11:00-11:05 (range since might choose to leave extra time for trip) at 7:42

— D. Algebraic expression for winter arrival time in City B

0 = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Makes an attempt, but no evidence of understanding; does not recognize what information is appropriate
4 = Understands problem statement and can do arithmetic involved in getting the winter schedule but cannot express the relationship algebraically (e.g., gives a numerical example)
5 = Uses appropriate information in an inappropriate way (e.g., mentions 1.1 or +5 but cannot set up the formula)
6 = Basically the formula is correct, except t is left out
7 = " " " " " except uses. 1 instead of 1.1, or takes 10% of something else
8 = " " " " " except sign error
9 = " " " " " except leaves out 5 (writes t + 1. 1y)
A = Correct formula: $t + 5 + (1.1)y$

Analytic Scores: Math Question 2 (Balance Beam)

Drawings of balance beams that were in and out of balance demonstrated the relationship between weight and distance from the pivot. Students were asked to demonstrate their understanding of the mathematics by drawing weights on partially-complete diagrams after determining distances of increasing complexity. The last step required the equation for the relationship between the weight and distance of the missing weight.

— **A.** Draw two **9-pound** weights so that the beam will be in balance

O = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Any incorrect attempt: attempted to draw weight or weights, but distances are not the same, or there are not 2 weights
8 = Correct placement but other than 9-pound weights, or weights are not labelled, or distances are not labelled
9 = Correct (9-pound weights are placed the same distance from the pivot, and on opposite sides of the pivot, distances labelled)

— **B.** Draw one **4-pound** weight to balance the beam

0 = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Any incorrect placement: attempted to draw weight or weights, but distance is incorrect and no work is shown
6 = Incorrect attempt: work is shown; method of solving is incorrect
7 = Incorrect attempt: work is shown; correct method; arithmetic error (including using weight other than a 4 pound weight)
8 = Uses correct method to solve problem but misreads diagram (e.g. counts fulcrum as 1 instead of O and writes 8(4) = 4(8) and places 4 pound weight 8 units to the right of the pivot: must show arithmetic)
9 = A 4 pound weight is placed six units to the right of the pivot

— **C.** Draw one additional **6-pound** weight so that the beam will be in balance

O = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Incorrect placement with no work shown (doesn't fit category 6)
4 = Puts weight on the wrong (right-hand) side of the pivot; incorrect method
5 = Puts weight on the correct (left) side of the pivot; incorrect method
6 = Puts new 6 pound weight 3 units to the left of the other 6 pound weight

7 = Uses correct method to solve problem, but makes minor computational error
8 = Uses correct method to solve problem, but misreads diagram--must show arithmetic
9 = A 6 pound weight is placed 3 units to the left of the pivot

— **D.** Equation which shows the relationship between **x** and **y**

0 = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = No evidence of understanding; does not recognize what information is appropriate (including giving an incorrect example)
4 = Method of solving problem is incorrect but relates to the problem (i.e., uses some appropriate information but in an inappropriate way)
5 = Uses correct method to solve problem, but makes minor computational error
6 = Uses correct method to solve problem, but misreads diagram--must show arithmetic
7 = General relationship between x and y indicated (e.g., as x increases, y decreases; x and y are inversely proportional)
8 = Correct example but not general formula
9 = xy = 13

Analytic Scores: Math Question 3 (Area of Figure Made of Rectangles)

Test takers were shown how to compute area by decomposing figures into rectangles, and then asked to decompose and find areas of increasingly complex figures, including ones with another area embedded, and one in which they had to determine the dimensions from a graph.

— A. Draw lines in the figure to show that it is made of rectangles

O = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Incorrect attempt (decomposition into figures other than rectangles, misreads problem, etc.)
8 = Begins but does not complete decomposition into rectangles
9 = Any correct decomposition into rectangles (may include extraneous lines filling out full rectangle)

— B. What is the area of the figure in **(A)?**

0 = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Incorrect attempt; work is not shown
4 = Incorrect attempt; calculates perimeter (32) instead of area
5 = Incorrect attempt; any other incorrect method or misreads problem
6 = Correct method; errors in both addition and determining areas
7 = Correct method; areas of individual rectangle(s) are determined incorrectly
8 = Correct method; arithmetic errors in addition of areas of rectangles
9 = Correct area: 30 (whether or not work is shown)

— C. What is the area of the shaded region in the figure **below?**

O = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Incorrect answer; work is not shown
4 = Made some attempt related to the problem but did not get far (includes incorrect method; misreading of problem)
5 = Attempted to break figure into rectangles but did not do anything else
9 = Calculated at least some area(s):

Score C1, C2, and C3 **only** if C=9; otherwise leave these scores blank:

— **C1.** Area of Small Rectangle
O = ignored small rectangle
1 = attempt to calculate area: incorrect method
8 = correct method; arithmetic error
9 = correct area (24)

— **C2.** Area of Large Figure
O = ignored large figure
1 = attempt to calculate area: incorrect method (includes not decomposing into rectangles; misreading problem)
8 = correct method; minor computational error
9 = correct area (130), or each part handled separately (90 - 24 + 40)

— **C3.** Subtract Small from Large Figure
O = ignored need to subtract (e.g., didn't differentiate shaded vs. unshaded area)
1 = attempt to calculate area: incorrect method (includes adding instead of subtracting; misreading problem)
8 = correct method; arithmetic error
9 = correct area (106) (C1 and C2 should be 9 unless errors are present)

— **D.** What is the area of the shaded region in the figure below?
0-9 (Same definitions as part C)

Score D1, D2, and D3 only if D=9; otherwise leave these scores blank:

— **D1.** Area of Small Rectangle
0, 1, 8, 9 (Same definitions as part C 1; correct area = 6)

— **D2.** Area of Large Figure
0, 1, 8, 9 (Same definitions as part C2; correct area = 26; code 8 includes misreading sidelength due to not properly using axis markings)

— **D3.** Subtract Small from Large Figure
0, 1, 8, 9 (Same definitions as part C3; correct area = 20)

Analytic Scores: Math Question 4 (Car Stopping Distance)

Given graphs relating speed to reaction distance and braking distance, students were asked to determine reaction, braking, and total stopping distances for cars traveling at different speeds, as well as to infer the minimum distance between cars from the length of skid marks before a collision. The last part, a request for an exp lariat ion of why the distance could have been greater was generally unsuccessful, both because it was apparently too difficult, and because the judgments of second readers showed unsatisfactory levels of reliability.

— A1. What is the driver's reaction distance?

O = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Incorrect: reads wrong graph (176)
6 = Incorrect (any other)
9 = Correct (80)

— A2. What is the braking distance for this car?

O = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Incorrect: reads wrong graph (80)
6 = Incorrect (any other)
9 = Correct (176)

— B. How far will the car travel until it comes to a complete stop?

0 = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Incorrect: 60
4 = Incorrect: 102
5 = Incorrect: uses 60 and 102 but does not know what to do with them
6 = Incorrect: any other
8 = Uses correct method but reads wrong numbers from graph, or makes arithmetic error
9 = Correct (60 + 102 = 162)

— **C.** What is the closest Car A could have been to the collision **point?**

0 = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
3 = Incorrect: answer is less than 100 feet
4 = Incorrect: uses 100 as total distance, Reaction and Braking. Proceeds correctly based on this assumption (braking distance is about 55)
6 = Incorrect: any other
8 = Uses correct method but reads wrong numbers from graph, or makes arithmetic error
9 = Correct (100 + 60 = 160; accept 155-160 as correct)

— **D.** Explain why Car A might have been farther **away.**

Some acceptable reasons:

- Skid marks don't begin immediately
- Car B moving and absorbs impact
- Car A in motion at time of impact and stopped by crash
- Reaction Distance graph may not apply to all cases: reaction time may be greater for a particular driver due to alcohol, fatigue, etc.
- Braking Distance graph may not apply in all cases; variations due to road/weather conditions; up/down hill; type/condition of tires, etc.

0 = No answer: section completely blank
1 = Off topic (such as doodles, irrelevant remarks, etc. indicating that the student was not attempting to answer the question)
2 = I don't know; I haven't learned this; I can't do this
6 = Incorrect: unacceptable reason or illogical explanation
9 = Correct: any acceptable reason

Analytic Scores: Science Question 1 (Nuclear vs. Fossil Fuels)

Students were asked to compare the use of nuclear fuels to the use of fossil fuels, including at least one advantage and one disadvantage of each.

— Any answer?

O = No answer: page completely blank
1 = Incomprehensible or irrelevant remarks; doodles
2 = I don't know; I haven't learned this
3 = Attempt to respond

(If score is code 0-2, leave the rest of the scores blank.)

Nuclear fuels: Advantages

— Number of advantages listed
— Which ones?

A. Clean air
B. Ample supply of fuel available
C. Small amounts of fuel needed/nuclear is more fuel-efficient

Nuclear fuels: Disadvantages

— Number of disadvantages listed
— Which ones?

F. Danger of release of radiation
G. Expensive to build
H. Thermal pollution
I. Expensive fuel (more processing)
J. Disposal of radioactive waste
K. Meltdown

Fossil fuels: Advantages

— Number of advantages listed
— Which ones?

M. Abundant fuel (coal)
N. Cheap fuel (easy to mine and use)
O. Economical to build power plants
P. Petroleum byproducts

Fossil fuels: Disadvantages

— Number of disadvantages listed
— Which ones?

R. Atmospheric pollution: acid rain, CO_2, nitrogen oxides, sulfur oxides
S. Non-renewable resource
T. Collateral environmental damage associated with mining, oil spills
U. Dependence on foreign supply
V. Relatively small amount of power for amount of fuel consumed

Incorrect, Invalid, or Emotional Statements

— Number of statements
— Which

X. Misunderstandings about nuclear energy (e.g., non-polluting, inexhaustible)
Y. Misunderstandings about fossil fuels (e.g., more expensive; imminent shortage of coal)
Z. Emotional statements (e.g., fossil fuels are "natural")

Awareness of Social Issues

— Any mention of unemployment; responsibility for cleanup; nuclear accidents; govt regulation, etc.

0 = no
1 = yes

Other Alternatives

— Any mention of solar energy, wind, geothermal, etc.)

0 = no
1 = yes

Analytic Scores: Science Question 2 (Eclipses)

Students were asked to draw diagrams of the relative positions of the earth, moon and sun during a solar eclipse and during a lunar eclipse, and to explain why a lunar eclipse can be seen from a greater geographic area on earth.

— A. Solar Eclipse Diagram

0 = No diagram: section completely blank
1 = Incomprehensible or irrelevant response; doodles
2 = I don't know; I haven't learned this
3 = Incomplete response: fewer than 3 bodies; or unclear which is which
4 = Earth, sun, moon arrangement
5 = Sun, earth, moon arrangement
6 = Other incorrect arrangements (including triangle) or misconceptions
7 = Sun, moon, earth arrangement (correct)

— B. Lunar Eclipse Diagram

0 = No diagram: section completely blank
1 = Incomprehensible or irrelevant response; doodles
2 = I don't know; I haven't learned this
3 = Incomplete response: fewer than 3 bodies; or unclear which is which
4 = Earth, sun, moon arrangement
5 = Sun, moon, earth arrangement
6 = Other incorrect arrangements (including triangle) or misconceptions
7 = Sun, earth, moon arrangement (correct)

— C. Explanation of Visibility of Eclipse

0 = No explanation
1 = Incomprehensible or irrelevant explanation; doodles
2 = I don't know; I haven't learned this
3 = Incomplete understanding of concept of eclipse: partial explanation, e.g., earth is larger than the moon
4 = Explanation based on relative frequency of lunar vs. solar eclipses rather than geographic area
5 = Explanation based on the size of the sun without comparison to the other bodies
6 = Explanation based on solar eclipse only: the sun is much larger than the moon (or moon is smaller than sun) and the moon therefore can't cover it
7 = Shadow cast by the moon onto the earth is relatively small, and the eclipse is visible only in the area of the shadow. The shadow cast by the earth onto the moon blocks all the sunlight to the moon and the eclipse is visible to all areas of the earth from which the moon can be seen (correct)
8 = Explanation confuses or reverses solar and lunar eclipses

Analytic **Scores:** Science Question **3 (Rabbit and Wolf Populations)**

Students were given a partially completed graph of population fluctuations of rabbits and wolves in an isolated ecosystem whose numbers are affected by a rabbit-specific virus. They were asked to complete the graph (draw acme for the wolf population) and to explain the height and position of the curve they drew compared to the rabbit curve.

— Any **Drawing?**

O = No drawing: graph is completely blank
1 = Incomprehensible or irrelevant marks on graph; doodles
2 = I don't know; I haven't learned this
3 = Attempt to draw graph, even if incorrect

Two features of drawing--score these onlv if there is an attempt to draw the graph (code 3). Otherwise leave the next three scores blank (but score the explanation separately).

— **A.** Phase of Wolf Curve

O = Phase is inconsistent or is a straight line
1 = The wolf curve leads the rabbit curve
2 = The wolf curve changes direction at the same time as the rabbit curve
3 = The wolf curve lags the rabbit curve
4 = Wolf curve is opposite to rabbit curve

— **B.** Relative heights of curves

O = Relative height inconsistent
1 = Wolf curve is higher than the rabbit curve
2 = Same height
3 = Wolf curve is lower than the rabbit curve

— Any **Explanation:** O (score even if there is no drawing)

= No explanation; completely blank
1 = h-relevant or incomprehensible explanation; doodles
2 = I don't know; I haven't learned this
3 = Comprehensible explanation, even if incorrect

Four features of explanation--score these onlv if there is a comprehensible explanation (code 3). (Otherwise leave the next four scores blank.)

— **A.** The **lower** amplitude of the wolf curve

O = not mentioned
1 = mentioned but incorrectly
2 = explained correctly

— **B.** The wolf curve lags behind the rabbit **curve.**

O = not mentioned
1 = mentioned but incorrectly
2 = explained correctly

— **C.** Rabbit population affects wolf **population:** More rabbits

makes possible more wolves **and/or** fewer rabbits
results in fewer **wolves.**

0 = not mentioned
1 = mentioned but incorrectly (Example: Wolves eat rabbits and then die from the virus)
2 = explained correctly

— **D.** Wolf population affects rabbit **population:** Fewer wolves makes possible more rabbits **and/or** more wolves result in fewer **rabbits.**

O = not mentioned
1 = mentioned but incorrectly
2 = explained correctly

Analytic Scores: Science Question 4 (Heating Curve)

A graph of time vs. temperature was presented for the mixture of water and ice being heating over a flame. Students were asked to explain why the 3 sections of the graph had horizontal or sloping lines.

— A. **Why** is the temperature constant in Section A

0 = No response: section is completely blank
1 = Incomprehensible or irrelevant response; doodles
2 = I don't know; I haven't learned this
3 = Doesn't understand graph; explanation does not mention temperature or heat
4 = Understands graph relates to temperature/heat over time (but no explanation of why: no mention of melting (change of phase) or heat absorption (heat of fusion)) e.g. takes time to heat up
5 = Explanation focuses on melting only (change of phase)
6 = Explanation focuses on melting (change of phase) and/or absorption/addition of heat (heat of fusion), but includes incorrect statements
7 = Explanation focuses on both melting (change of phase) and absorption of heat (heat of fusion)
8 = Explanation focuses on potential energy change (correct)
9 = Explanation focuses on potential energy change but was incorrect

— B. Why does Section B of the Curve Slope Upward

0 = No response: section is completely blank
1 = Incomprehensible or irrelevant response; doodles
2 = I don't know; I haven't learned this
3 = Doesn't understand graph; explanation does not mention temperature or heat
4 = Explanation focuses on increasing temperature only
5 = Explanation focuses on absorption/addition of heat only
6 = Explanation focuses on increasing temperature and/or absorption of heat but includes incorrect statement(s)
7 = Explanation focuses on both increasing temperature and absorption of heat
8 = Explanation focuses on kinetic energy change (correct)
9 = Explanation focuses on kinetic energy change but was incorrect

— C. Why is the temperature constant in Section C

0 = No response: section is completely blank
1 = Incomprehensible or irrelevant response; doodles
2 = I don't know; I haven't learned this
3 = Doesn't understand graph; explanation does not mention temperature or heat
4 = Understands graph relates to temperature over time (but no explanation of why: no mention of boiling/evaporation (change of phase) or heat absorption (heat of vaporization)
5 = Explanation focuses on boiling/evaporation only (change of phase)
6 = Explanation focuses on boiling/evaporation (change of phase) and/or absorption of heat (heat of vaporization) but includes incorrect statement(s)
7 = Explanation focuses on both boiling/evaporation (change of phase) and absorption of heat (heat of vaporization)
8 = Explanation focuses on potential energy change (correct)
9 = Explanation focuses on potential energy change but was incorrect

Scale Score: Math Question 1 (Train Schedule)

0 = any part code 2 or more, but nothing correct

1 = part A = 4 (correct train) or part B=5 (correct procedure with errors)
or part D=5 (appropriate information used in inappropriate way)

2 = part B=6 (correct train) or part C = 5 or 6 (correct procedure, with errors)
or part D=4 (correct arithmetic but not general formula)
or D=6,7,8,9 (partially correct formula)

3 = part C=7 (correct time)

4 = part C=7 (correct time) AND part D=6,7,8,9 (partially correct formula)
or D=A (completely correct formula) but C not correct

5 = part C=7 (correct time) AND D=A (correct formula)

Scale Score: Math Question 2 (Balance Beam)

0 = any part code 2 or more, but nothing correct

1 = part A=9 (correct) or B=7 (correct method with errors)
or C=5 (incorrect method)

2 = part B=8 or 9, or C=6 (correct method, may have problems with diagram)

3 = part C=7 or 8; or D=5 or 6 (correct method; computational error or misreads diagram)
or D=7 (general relationship indicated but not formula)

4 = C=9 (correct)
or D=8 (correct example but not general formula)
or C=7,8 AND D=5,6,7 (correct method, minor error in both parts)

5 = D=9 (correct)

Scale Score: Math Question 3 (Area of Figure Made of Rectangles)

O = any part code 2 or more, but nothing correct

1 = A=8 or 9, or C=5 or D=5 (attempts decomposition, nothing else)

2 = B=6 or 7 (decomposition ok, correct method for area, but errors)
C1 or D1=8 or 9 (correct method for area of small figure)
C3 or D3=8 or 9 (subtraction of small from large figure)

3 = B=8 or 9 (correct decomposition and area; may have addition error)
C2 or D2=8 or 9 (correct method for large figure; may have error)

4 = C1, C2 and C3=8 or 9 (correct method for decomposition, computing area, and subtraction for complex figure)

5 = D1, D2 and D3=8 or 9 (correct method for decomposition, computing area, and subtraction for more complex figure)

Scale Score: Math Question 4 (Car Stopping Distance)

O = any part code 2 or more, but nothing correct

1 = A1=9 or A2=9 or B=3,4,6 (graph reading only)

2 = B=3 or 5; or C=4 (parts of method correct, but not good progress)

3 = B=9 (correct stopping distance, sum of parts)

4 = no score 4 (large step in difficulty to next part)

5 = C=8 or 9 (correct method, may have minor error)

Scale Score: Science Question 1 (Nuclear vs. Fossil Fuels)

0 = attempted problem ("any answer" = 2 or 3), or indicated inability to answer (question was "hard" or "too hard"; "didn't know how to answer"; or "have not taken the courses")

1-5 = count of how many distinct and valid (codes A-V) advantages and disadvantages of nuclear and/or fossil fuels are described. Categories (nuclear advantages; nuclear disadvantages; fossil fuels advantages; fossil disadvantages) are not itemized separately, even though the question asks for it, because it is not always possible to make a distinction. For example, "Power plants using fossil fuels are cheaper to build than nuclear reactors" could be interpreted to be an advantage of one or a disadvantage of the other.

Add 1 point if any mention made of awareness of social issues and/or alternative energy sources (only one point even if both are mentioned). The extra point is added only if:

the count of valid advantages/disadvantages is at least 2 (that is, the student has basically answered the question,

and

the additional point does not exceed the maximum score of 5

Subtract 1 point if one or more incorrect, invalid or emotional statements (codes X,Y,Z). Only one point is subtracted, even if more than one incorrect statement is present. A score that includes at least one valid response will not be lowered to less than 1. The point is subtracted after the cap of 5 has been applied. For example, a response containing 7 valid statements and a discussion of alternative energy sources would only receive a score of 4 if there are also incorrect statements present.

Scale Score: Science Question 2 (Eclipses)

0 = attempted problem (any part = 2 or above, but no correct or partially correct answer), or indicated inability to answer (question was "hard" or "too hard"; "didn't know how to answer"; or "have not taken the courses")

1 = explanation is code 3 or higher, but no diagram is correct

2 = one correct diagram (A=7 or B=7), nothing added for explanation

3 = both diagrams correct (A=7 and B=7), nothing added for explanation
or
one diagram correct (A=7 or B=7) and partial explanation (C=4,5,6 or 8)

4 = both diagrams correct (A=7 and B=7) and partial explanation (C=4,5,6 or 8)
or
one diagram correct (A=7 or B=7) and complete explanation (C=7)

5 = both diagrams correct; correct explanation (C=7)

Scale Score: Science Question 3 (Rabbit and Wolf Populations)

0 = attempted problem (score of 2 or more on any part), but no correct or partially correct answer), or indicated inability to answer (question was "hard" or "too hard"; "didn't know how to answer"; or "have not taken the courses")

1 = wolf height correct (graph score B=3)
or
correct "rabbit affects wolf' explanation (expl. C=2)

2 = wolf height correct (graph score B=3) AND explanation C=2
or
wolf lag correct (graph score A=3)
or
explanation A, B or D correct (=2) but no graph correct

3 = both graph features correct (A=3 and B=3), but no explanation
or
confirms understanding of amplitude (graph B=3 and expl. A=2)
or
confirms understanding of lag (graph A=3 and expl. B=2)

4 = at least one indicator for each feature (amplitude, lag); must include at least one valid explanation:
height indicator (graph B=3 OR expl. A=2)
AND
lag indicator (graph A=3 OR expl. B=2)
AND
at least one correct explanation A, B, C or D = 2

5 = both graph features correct (graph A=3 and B=3) plus at least one explanation in A, B, or D (correct=2)

Scale Score: Science Question 4 (Heating Curve)

0 = attempted problem (score of 2 or more on any part), but no correct or partially correct answer), or indicated inability to answer (question was "hard" or "too hard";"didn't know how to answer"; or "have not taken the courses")

1 = any part = 4: understands that graph relates temperature to time, but does not deal with the addition of heat; or B=6 which may mention absorption of heat but has an incorrect statement

2 = A=6 or C=6: melting/heat/boiling/evaporation, but incorrect or any part =9: mention of potential or kinetic energy, but incorrect

3 = At least 2 parts equal to 5, 7 or 8

4 = All 3 parts equal 5, 7 or 8; with at least one 7 or 8

5 = All 3 parts equal to 7 or 8

Appendix B

Test Score Statistics and Breakdowns by Responses to Student Reaction Questions

- Counts for All Constructed Response Test Takers
- Counts of Subset Who Also Had Multiple Choice Tests
- Multiple Choice Test Means for Each Scale Point
- Multiple Choice Test Standard Deviations for Each Scale Point

Math Question 1: Train Schedule
Counts of **All** Constructed Response Test Takers
By Constructed Response Scale Score Level and Responses to Reaction Questions

Sample Counts	Total	**0**	**1**	**2**	**3**	**4**	**5**	Imputed **0**	**Blank**
All Test Takers	2415	59	347	554	988	222	226	14	5
Difficulty Too Easy	147	6	12	22	41	16	49	1	0
Difficulty Easy	427	2	22	63	156	87	97	0	0
Difficulty Right	780	16	77	177	380	72	56	0	2
Difficulty Hard	765	21	155	207	314	43	21	4	0
Difficulty Too Hard	219	11	64	64	68	2	1	9	0
Omitted Question	77	3	17	21	29	2	2	0	3
Didn't Know Answer	576	23	158	178	192	14	2	9	0
Answer Partly Right	873	21	107	212	429	70	30	2	2
Pretty Good Answer	879	13	61	141	335	136	191	2	0
Omitted Question	87	2	21	23	32	2	3	1	3
Had Courses	1768	26	169	355	782	210	222	2	2
Didn't Have Courses	539	29	154	173	163	10	0	10	0
Omitted Question	108	4	24	26	43	2	4	2	3
Question Very Clear	447	9	38	71	165	66	98	0	0
Clear Enough	734	11	67	147	330	88	87	3	1
A **Little** Confusing	931	23	166	249	387	62	38	5	1
Very Confusing	209	14	54	63	69	4	0	5	0
Omitted Question	94	2	22	24	37	2	3	1	3
Not Enough Time	103	10	26	35	25	1	2	2	2
A Little More Time	208	10	38	52	89	12	5	2	0
Right Amount of Time	1108	24	174	280	476	89	63	2	0
A Little Too Much	454	5	41	86	198	58	64	2	0
Way Too Much Time	436	8	44	73	160	59	88	4	0
Omitted Question	106	2	24	28	40	3	4	2	3

Math Question 1: Train Schedule
Counts of Subset with Multiple Choice Tests
By Constructed Response Scale Score Level and Responses to Reaction Questions

Number with M.C. Test	Total	**0**	**1**	**2**	**3**	**4**	**5**	**Imputed 0**	Blank
All Test **Takers**	2386	56	341	546	977	222	225	14	5
Difficulty Too Easy	146	6	11	22	41	16	49	1	0
Difficulty Easy	423	1	22	62	154	87	97	0	0
Difficulty Right	773	16	76	176	376	72	55	0	2
Difficulty Hard	753	21	151	203	310	43	21	4	0
Difficulty Too Hard	215	9	64	62	68	2	1	9	0
Omitted Question	76	3	17	21	28	2	2	0	3
Didn't Know Answer	565	21	155	175	189	14	2	9	0
Answer Partly Right	864	20	105	210	426	70	29	2	2
Pretty Good Answer	872	13	60	138	332	136	191	2	0
Omitted Question	85	2	21	23	30	2	3	1	3
Had Courses	1750	24	166	351	774	210	221	2	2
Didn't Have Courses	530	28	152	169	161	10	0	10	0
Omitted Question	106	4	23	26	42	2	4	2	3
Question Very Clear	438	9	34	69	162	66	98	0	0
Clear Enough	728	10	67	145	327	88	87	3	1
A Little Confusing	922	22	165	246	384	62	37	5	1
Very Confusing	205	13	53	62	68	4	0	5	0
Omitted Question	93	2	22	24	36	2	3	1	3
Not Enough Time	98	9	25	34	23	1	2	2	2
A Little More Time	203	9	37	51	87	12	5	2	0
Right Amount of Time	1093	23	171	275	470	89	63	2	0
A Little Too Much	453	5	41	85	198	58	64	2	0
Way Too Much Time	434	8	43	73	160	59	87	4	0
Omitted Question	105	2	24	28	39	3	4	2	3

Math Question 1: Train Schedule
Multiple Choice Math Test Means
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test Mean	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	50.7	28.9	37.5	44.7	52.5	65.5	69.7	29.9	33.7
Difficulty Too Easy	60.5	24.8	42.1	50.0	61.6	67.6	71.4	21.2	*
Difficulty Easy	61.3	20.1	40.1	51.6	58.6	67.7	71.0	*	*
Difficulty Right	51.0	28.1	38.1	46.3	51.9	63.9	68.3	*	34.5
Difficulty Hard	46.0	29.2	36.7	42.3	50.8	63.1	64.1	34.4	*
Difficulty Too Hard	40.3	34.3	37.2	38.3	46.1	64.1	77.2	28.9	*
Omitted Question	44.9	25.6	36.0	47.8	48.9	64.5	59.1	*	33.2
Didn't Know Answer	41.1	27.7	35.8	39.3	47.2	61.6	63.9	30.8	*
Answer Partly Right	49.6	30.5	39.4	45.8	51.6	63.5	66.9	28.9	34.5
Pretty Good Answer	58.4	28.9	38.6	49.3	56.8	67.0	70.3	27.6	*
Omitted Question	45.6	25.3	36.9	47.8	50.2	64.5	62.6	29.0	33.2
Had Courses	54.8	29.3	41.7	47.7	54.4	65.7	69.8	40.2	34.5
Didn't Have Courses	38.2	29.1	32.7	38.1	44.4	61.3	*	28.5	*
Omitted Question	45.5	24.9	38.4	47.2	49.4	64.5	64.8	27.0	33.2
Question Very Clear	59.0	32.0	42.8	49.8	57.5	68.1	69.8	*	*
Clear Enough	54.8	29.6	40.6	49.9	54.0	65.1	71.0	27.4	24.6
A Little Confusing	46.5	26.3	36.4	41.7	50.4	63.6	67.1	29.2	44.4
Very Confusing	39.6	31.1	33.1	38.2	46.7	61.2	*	32.4	*
Omitted Question	45.4	25.3	37.8	46.5	49.3	64.5	62.6	29.0	33.2
Not Enough Time	39.5	30.0	34.4	39.8	45.9	65.4	63.1	34.1	34.5
A Little More Time	47.2	28.1	36.4	43.8	53.1	63.0	63.2	34.1	*
Right Amount of Time	48.5	30.1	37.3	43.8	50.7	64.4	67.7	38.5	*
A Little Too Much	54.8	27.7	36.8	47.4	54.9	65.3	69.6	25.5	*
Way Too Much Time	57.1	26.6	41.8	46.4	56.1	68.1	72.1	26.7	*
Omitted Question	45.8	25.3	36.9	48.3	49.4	62.6	64.1	24.0	33.2

* No data for this cell.

Math Question 1: Train Schedule
Multiple Choice Math Test Standard Deviations
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test S.D.	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	15.4	8.3	11.6	13.8	12.0	7.4	6.0	9.8	9.3
Difficulty Too Easy	16.5	6.2	16.1	15.2	11.1	12.7	6.3	0.0	*
Difficulty Easy	13.2	0.0	13.8	15.5	10.9	5.6	5.4	*	*
Difficulty Right	14.1	7.4	12.0	12.8	12.0	6.4	5.3	*	9.9
Difficulty Hard	13.5	8.1	10.6	12.4	10.6	7.9	4.3	13.0	*
Difficulty Too Hard	13.2	9.7	11.0	13.7	11.8	12.1	0.0	7.5	*
Omitted Question	15.5	4.7	13.4	14.7	14.1	1.5	0.9	*	8.8
Didn't Know Answer	13.0	8.6	10.4	12.2	11.6	7.1	2.6	11.5	*
Answer Partly Right	13.2	8.1	11.5	12.6	11.0	6.4	6.3	7.7	9.9
Pretty Good Answer	14.7	7.9	13.6	15.1	11.8	7.6	5.7	2.6	*
Omitted Question	15.2	5.8	12.9	14.8	13.4	1.5	5.0	0.0	8.8
Had Courses	14.1	9.0	12.0	13.5	11.3	7.3	5.9	3.6	9.9
Didn't Have Courses	12.3	8.0	8.9	12.0	11.4	8.5	*	10.3	*
Omitted Question	14.6	4.3	12.6	14.2	12.5	1.5	5.7	2.0	8.8
Question Very Clear	13.9	8.6	13.8	14.5	10.8	6.0	6.1	*	*
Clear Enough	14.1	10.7	13.0	12.9	11.4	7.5	4.7	6.6	0.0
A Little Confusing	14.1	4.3	10.1	12.1	11.6	7.7	7.1	8.7	0.0
Very Confusing	14.3	9.9	9.9	14.8	13.4	9.2	*	12.5	*
Omitted Question	14.7	5.8	12.7	15.0	12.8	1.5	5.0	0.0	8.8
Not Enough Time	14.7	10.8	10.4	16.3	12.3	0.0	3.4	9.7	9.9
A Little More Time	14.5	9.0	11.8	12.9	11.3	8.3	12.0	2.5	*
Right Amount of Time	14.4	8.4	11.5	13.2	11.5	7.0	6.1	17.3	*
A Little Too Much	14.5	4.1	10.6	14.2	10.9	8.2	5.4	5.4	*
Way Too Much Time	15.8	5.0	12.2	13.6	13.2	6.5	4.5	4.4	*
Omitted Question	15.0	5.8	12.6	14.8	12.5	3.0	5.0	5.0	8.8

* No data for this cell.

Math Question 2: Balance Beam
Counts of All Constructed Response Test Takers
By Constructed Response Scale Score Level and Responses to Reaction Questions

Sample Counts	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	2415	124	384	557	85	390	706	115	54
Difficulty Too Easy	225	5	11	15	1	32	157	3	1
Difficulty Easy	456	1	18	48	9	82	294	3	1
Difficulty Right	701	20	104	171	38	163	200	3	2
Difficulty Hard	594	48	149	217	26	81	46	27	0
Difficulty Too Hard	340	46	89	89	9	24	5	78	0
Omitted Question	99	4	13	17	2	8	4	1	50
Didn't Know Answer	662	72	187	202	26	61	16	98	0
Answer Partly Right	732	23	139	240	35	169	115	8	3
Pretty Good Answer	923	26	46	98	22	151	571	8	1
Omitted Question	98	3	12	17	2	9	4	1	50
Had Courses	1404	34	132	259	44	277	643	11	4
Didn't Have Courses	903	86	238	279	39	102	60	99	0
Omitted Question	108	4	14	19	2	11	3	5	50
Question Very Clear	600	10	35	66	14	96	368	10	1
Clear Enough	773	20	100	188	35	165	257	7	1
A Little Confusing	556	29	133	192	26	91	70	13	2
Very Confusing	387	62	104	94	8	27	9	83	0
Omitted Question	99	3	12	17	2	11	2	2	50
Not Enough Time	135	20	26	32	6	10	7	34	0
A Little More Time	151	10	38	54	4	20	12	10	3
Right Amount of Time	1005	53	200	278	44	192	207	30	1
A Little Too Much	459	11	52	89	17	83	199	8	0
Way Too Much Time	534	25	52	77	10	73	275	22	0
Omitted Question	131	5	16	27	4	12	6	11	50

Math Question 2: Balance Beam
Counts of Subset with Multiple Choice Test
By Constructed Response Scale Score Level and Responses to Reaction Questions

Number with M.C. Test	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	2386	118	378	554	84	386	701	112	53
Difficulty Too Easy	223	5	11	15	1	32	155	3	1
Difficulty Easy	450	1	16	48	8	81	292	3	1
Difficulty Right	695	20	102	170	38	161	199	3	2
Difficulty Hard	586	46	147	215	26	80	46	26	0
Difficulty Too Hard	334	42	89	89	9	24	5	76	0
Omitted Question	98	4	13	17	2	8	4	1	49
Didn't Know Answer	651	69	184	201	25	61	16	95	0
Answer Partly Right	721	21	136	238	35	165	115	8	3
Pretty Good Answer	917	25	46	98	22	151	566	8	1
Omitted Question	97	3	12	17	2	9	4	1	49
Had Courses	1389	33	128	258	43	274	638	11	4
Didn't Have Courses	890	81	236	277	39	101	60	96	0
Omitted Question	107	4	14	19	2	11	3	5	49
Question Very Clear	595	9	35	66	14	96	364	10	1
Clear Enough	767	20	100	186	34	163	256	7	1
A Little Confusing	550	29	131	191	26	89	70	12	2
Very Confusing	376	57	100	94	8	27	9	81	0
Omitted Question	98	3	12	17	2	11	2	2	49
Not Enough Time	128	18	24	31	5	10	7	33	0
A Little More Time	150	9	38	54	4	20	12	10	3
Right Amount of Time	989	50	196	276	44	189	205	28	1
A Little Too Much	458	11	52	89	17	82	199	8	0
Way Too Much Time	531	25	52	77	10	73	272	22	0
Omitted Question	130	5	16	27	4	12	6	11	49

Math Question 2: Balance Beam
Multiple Choice Math Test Means
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test Mean	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	50.7	33.4	39.5	46.9	49.5	53.0	64.8	33.0	44.4
Difficulty Too Easy	63.4	29.9	35.1	49.7	71.8	57.6	69.6	43.5	32.3
Difficulty Easy	60.6	60.4	46.4	47.4	46.2	56.6	65.6	19.6	26.2
Difficulty Right	51.2	30.8	38.7	47.6	50.6	53.5	61.8	29.6	19.3
Difficulty Hard	45.0	33.9	40.7	47.2	46.7	49.5	56.6	33.6	*
Difficulty Too Hard	39.2	34.5	37.8	44.2	52.7	45.0	56.3	33.1	*
Omitted Question	45.2	26.0	38.5	47.0	52.6	46.0	65.8	30.5	46.1
Didn't Know Answer	40.9	34.5	38.5	44.9	49.9	44.7	59.4	33.6	*
Answer Partly Right	49.4	32.6	41.4	48.6	49.9	53.0	60.3	33.5	25.3
Pretty Good Answer	59.2	31.9	37.1	47.1	48.1	56.8	65.9	25.7	21.0
Omitted Question	45.4	25.8	41.5	47.1	52.6	44.5	60.3	27.3	46.1
Had Courses	57.1	33.8	43.3	49.8	53.3	55.8	65.2	38.5	24.3
Didn't Have Courses	41.5	33.7	37.3	44.2	45.1	46.0	59.8	32.7	*
Omitted Question	44.4	24.3	41.0	47.1	52.6	45.2	61.1	27.4	46.1
Question Very Clear	60.6	33.2	41.5	50.8	55.9	57.4	66.4	43.7	17.5
Clear Enough	53.3	35.3	42.3	48.4	47.4	53.4	64.0	31.2	26.2
A Little Confusing	45.6	29.7	38.0	46.6	50.1	50.3	59.6	27.7	26.7
Very Confusing	38.6	35.0	37.8	42.0	44.4	46.3	60.3	32.4	*
Omitted Question	45.2	25.8	39.7	47.1	52.6	45.5	63.1	43.4	46.1
Not Enough Time	40.4	30.1	35.1	47.5	53.4	49.0	61.2	34.1	*
A Little More Time	44.6	34.1	40.5	46.4	57.1	50.7	62.9	27.0	26.5
Right Amount of Time	47.9	34.3	38.6	46.7	48.0	51.2	60.4	33.8	17.5
A Little Too Much	55.0	34.9	40.9	47.5	49.1	55.2	64.5	33.2	*
Way Too Much Time	57.6	33.3	42.1	47.2	48.7	57.1	68.4	29.9	*
Omitted Question	45.5	31.8	39.8	46.7	56.4	46.9	62.8	39.4	46.1

* No data for this cell.

Math Question 2: Balance Beam
Multiple Choice Math Test Standard Deviations
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test S.D.	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	15.4	10.9	11.8	11.7	13.2	12.3	8.7	11.0	15.5
Difficulty Too Easy	14.1	8.7	12.1	13.4	0.0	12.6	6.5	20.6	0.0
Difficulty Easy	12.3	0.0	9.4	11.8	14.6	10.9	8.3	0.4	0.0
Difficulty Right	13.7	11.3	12.4	11.2	11.6	11.0	8.1	10.2	1.8
Difficulty Hard	12.9	10.5	11.7	11.3	12.4	12.8	9.4	9.5	*
Difficulty Too Hard	13.0	10.8	11.1	12.9	16.2	13.8	8.1	10.6	*
Omitted Question	14.4	1.4	9.0	11.2	14.8	14.1	5.7	0.0	14.9
Didn't Know Answer	12.9	11.0	11.3	11.9	13.9	13.2	7.2	11.0	*
Answer Partly Right	13.2	11.2	12.7	11.3	11.2	11.4	8.7	13.3	6.1
Pretty Good Answer	13.9	10.5	10.4	11.6	14.8	10.7	8.4	3.4	0.0
Omitted Question	14.1	1.6	9.7	11.4	14.8	13.9	6.5	0.0	14.9
Had Courses	13.4	12.6	12.4	11.2	11.1	10.8	8.5	16.1	5.6
Didn't Have Courses	13.2	10.2	11.0	11.4	13.8	12.9	9.7	10.2	*
Omitted Question	14.4	2.8	9.2	13.3	14.8	12.8	4.6	3.9	14.9
Question Very Clear	13.3	13.2	14.7	11.8	11.3	11.2	8.4	15.9	0.0
Clear Enough	13.4	12.1	11.1	11.3	12.2	10.7	8.2	9.7	0.0
A Little Confusing	13.8	8.7	11.7	10.8	13.3	12.9	9.6	6.0	5.7
Very Confusing	12.7	10.8	11.0	12.6	14.3	15.3	7.7	9.8	*
Omitted Question	13.9	1.6	8.5	11.4	14.8	12.9	4.4	12.8	14.9
Not Enough Time	15.1	10.3	9.9	13.2	12.7	12.5	7.2	14.2	*
A Little More Time	13.9	8.3	11.7	11.0	11.8	12.6	8.4	7.5	4.6
Right Amount of Time	13.9	10.9	11.8	11.3	13.8	12.1	9.3	9.5	0.0
A Little Too Much	14.1	10.6	12.1	12.3	10.8	11.3	8.2	6.8	*
Way Too Much Time	15.9	12.4	12.2	12.8	13.2	11.7	6.9	7.0	*
Omitted Question	13.5	8.1	8.2	10.5	11.2	13.2	7.3	10.4	14.9

* No data for this cell.

Math Question 3: Area of Figure Made of Rectangles
Counts of All Constructed Response Test Takers
By Constructed Response Scale Score Level and Responses to Reaction Questions

Sample Counts	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	2415	183	320	214	128	168	1197	112	93
Difficulty Too Easy	540	12	10	10	9	34	459	5	1
Difficulty Easy	633	13	22	31	21	48	496	2	0
Difficulty Right	496	43	79	76	54	60	178	3	3
Difficulty Hard	356	56	115	68	30	20	40	27	0
Difficulty Too Hard	270	50	83	25	11	5	21	75	0
Omitted Question	120	9	11	4	3	1	3	0	89
Didn't Know Answer	503	91	159	65	26	16	49	97	0
Answer Partly Right	490	44	101	96	52	51	134	9	3
Pretty Good Answer	1295	38	47	48	47	100	1009	5	1
Omitted Question	127	10	13	5	3	1	5	1	89
Had Courses	1782	73	163	149	97	148	1135	13	4
Didn't Have Courses	494	97	141	59	26	18	57	96	0
Omitted Question	139	13	16	6	5	2	5	3	89
Question Very Clear	1085	16	35	48	32	82	866	5	1
Clear Enough	583	46	77	73	55	53	268	9	2
A Little Confusing	317	48	99	67	27	22	42	12	0
Very Confusing	305	64	96	21	11	10	17	85	1
Omitted Question	125	9	13	5	3	1	4	1	89
Not Enough Time	125	23	29	13	6	9	15	30	0
A Little More Time	112	9	33	17	4	12	23	12	2
Right Amount of Time	815	82	147	104	73	63	308	36	2
A Little Too Much	384	22	40	31	17	27	242	5	0
Way Too Much Time	833	37	52	43	23	55	600	23	0
Omitted Question	146	10	19	6	5	2	9	6	89

Math Question 3: Area of Figure Made of Rectangles
Counts of Subset with Multiple Choice Tests
By Constructed Response Scale Score Level and Responses to Reaction Questions

Number with M.C. Test	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	2386	179	314	210	127	167	1189	109	91
Difficulty Too Easy	537	12	10	10	9	34	456	5	1
Difficulty Easy	627	13	21	31	21	47	492	2	0
Difficulty Right	491	43	77	76	53	60	177	3	2
Difficulty Hard	346	53	113	64	30	20	40	26	0
Difficulty Too Hard	266	49	82	25	11	5	21	73	0
Omitted Question	119	9	11	4	3	1	3	0	88
Didn't Know Answer	491	87	157	63	26	16	48	94	0
Answer Partly Right	481	44	98	94	52	51	131	9	2
Pretty Good Answer	1288	38	46	48	46	99	1005	5	1
Omitted Question	126	10	13	5	3	1	5	1	88
Had Courses	1765	73	160	147	96	147	1127	12	3
Didn't Have Courses	483	93	138	57	26	18	57	94	0
Omitted Question	138	13	16	6	5	2	5	3	88
Question Very Clear	1076	16	34	48	32	81	861	4	0
Clear Enough	577	46	77	72	54	53	265	8	2
A Little Confusing	310	46	96	65	27	22	42	12	0
Very Confusing	299	62	94	20	11	10	17	84	1
Omitted Question	124	9	13	5	3	1	4	1	88
Not Enough Time	118	21	28	12	6	9	14	28	0
A Little More Time	110	9	32	17	4	12	23	12	1
Right Amount of Time	802	81	144	102	72	62	304	35	2
A Little Too Much	382	22	39	31	17	27	241	5	0
Way Too Much Time	829	36	52	42	23	55	598	23	0
Omitted Question	145	10	19	6	5	2	9	6	88

Math Question 3: Area of Figure Made of Rectangles
Multiple Choice Math Test Means
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test Mean	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	50.7	33.2	38.6	41.1	45.2	51.8	60.6	33.4	45.3
Difficulty Too Easy	62.0	41.4	41.9	42.1	50.7	58.8	64.3	31.2	38.5
Difficulty Easy	57.3	31.3	38.6	43.7	49.7	52.3	60.6	20.7	*
Difficulty Right	46.0	32.2	38.3	39.4	44.2	49.0	55.1	38.0	40.2
Difficulty Hard	39.8	31.4	38.9	41.7	42.2	47.0	48.8	34.2	*
Difficulty Too Hard	37.8	34.8	38.1	40.5	44.3	51.4	48.5	33.5	*
Omitted Question	44.2	32.0	39.6	42.5	47.9	51.9	55.7	*	45.5
Didn't Know Answer	38.7	34.2	38.5	41.2	43.6	45.5	50.2	33.0	*
Answer Partly Right	43.5	31.4	38.9	41.2	41.8	47.8	52.4	36.3	40.2
Pretty Good Answer	58.5	32.6	38.0	40.5	49.7	54.9	62.2	32.2	38.5
Omitted Question	44.5	35.2	41.1	44.5	47.9	51.9	49.7	54.0	45.5
Had Courses	55.0	34.6	40.1	42.8	46.7	52.4	61.3	36.2	39.6
Didn't Have Courses	37.1	32.3	36.9	36.7	39.9	46.7	47.3	33.3	*
Omitted Question	43.0	32.0	39.1	40.9	44.1	54.4	46.7	25.5	45.5
Question **Very** Clear	59.8	34.2	38.6	45.0	49.6	55.6	62.8	38.1	*
Clear **Enough**	48.6	34.5	40.7	41.1	45.2	49.4	56.5	31.9	40.2
A Little Confusing	38.9	32.5	37.9	38.8	41.5	44.4	46.2	31.4	*
Very Confusing	36.9	32.8	37.6	39.3	40.2	50.2	52.8	33.4	38.5
Omitted Question	44.0	32.0	40.3	39.7	47.9	51.9	49.6	54.0	45.5
Not Enough Time	40.3	34.3	38.2	36.6	53.0	47.7	54.2	36.5	*
A Little More Time	41.8	31.6	36.9	42.3	38.3	46.7	56.1	30.9	38.5
Right Amount of Time	45.6	33.6	38.3	40.0	44.4	50.5	54.9	32.3	40.2
A Little Too Much	53.1	34.4	40.6	42.6	42.9	48.5	59.9	26.9	*
Way Too Much Time	58.3	31.3	39.6	42.8	49.0	56.5	64.2	32.9	*
Omitted Question	43.9	33.4	37.9	45.5	41.7	57.0	54.3	38.3	45.5

* No data for this cell.

Math Question 3: Area of Figure Made of Rectangles
Multiple Choice Math Test Standard Deviations
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test S.D.	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	15.4	10.9	10.1	10.5	11.4	11.7	10.8	11.8	14.0
Difficulty Too Easy	12.4	15.7	13.6	8.4	13.5	11.3	10.3	10.6	0.0
Difficulty Easy	11.9	6.9	8.1	9.6	12.1	12.8	9.2	1.2	*
Difficulty Right	13.3	10.0	10.0	11.9	10.3	9.4	11.1	9.5	1.4
Difficulty Hard	11.8	11.0	10.0	9.9	11.0	10.2	11.1	12.5	*
Difficulty Too Hard	11.5	9.9	9.9	9.3	9.6	11.7	11.3	11.6	*
Omitted Question	14.0	10.2	13.2	2.6	10.6	0.0	11.3	*	14.1
Didn't Know Answer	12.3	11.5	9.9	10.1	12.1	12.0	13.8	11.5	*
Answer Partly Right	12.2	9.0	9.9	10.2	10.7	9.3	10.9	14.5	1.4
Pretty Good Answer	12.9	10.4	10.6	12.0	10.1	11.7	9.8	9.0	0.0
Omitted Question	13.9	13.5	12.7	4.6	10.6	0.0	13.0	0.0	14.1
Had Courses	13.8	11.1	10.2	10.6	10.9	11.0	10.2	15.1	1.4
Didn't Have Courses	11.8	10.8	9.4	9.2	11.7	15.4	12.4	11.4	*
Omitted Question	14.0	9.8	12.7	9.1	10.1	2.6	14.2	3.9	14.1
Question Very Clear	12.0	13.4	11.9	9.4	13.0	10.8	9.4	16.3	*
Clear Enough	13.4	11.0	9.4	11.1	9.6	10.4	11.2	12.7	1.4
A Little Confusing	11.5	9.8	9.7	10.6	11.2	9.5	13.7	8.9	*
Very Confusing	12.1	10.8	9.8	8.7	9.8	17.1	10.9	11.6	0.0
Omitted Question	13.9	10.2	12.4	6.1	10.6	0.0	14.5	0.0	14.1
Not Enough Time	14.2	12.9	11.3	9.9	15.3	10.0	12.8	13.9	*
A Little More Time	13.7	7.9	8.2	11.6	12.0	14.8	10.8	11.7	0.0
Right Amount of Time	13.4	10.4	10.0	11.0	10.0	9.8	10.7	10.0	1.4
A Little Too Much	13.8	12.4	10.1	10.0	11.0	10.7	9.8	5.9	*
Way Too Much Time	14.5	10.2	10.1	9.2	12.2	12.0	9.7	11.5	*
Omitted Question	13.8	10.5	11.8	4.7	11.2	5.2	11.0	10.7	14.1

* No data for this cell.

Math Question 4: Car Stopping Distance
Counts of All Constructed Response Test Takers
By Constructed Response Scale Score Level and Responses to Reaction Questions

Sample Counts	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	2415	69	962	307	378	0	462	112	125
Difficulty Too Easy	149	3	35	14	26	0	66	2	3
Difficulty Easy	362	2	69	41	74	0	175	0	1
Difficulty Right	702	6	294	105	142	0	151	2	2
Difficulty Hard	652	22	354	102	95	0	48	31	0
Difficulty Too Hard	340	33	160	36	24	0	10	77	0
Omitted Question	210	3	50	9	17	0	12	0	119
Didn't Know Answer	720	51	385	91	66	0	28	99	0
Answer Partly Right	841	10	382	139	174	0	125	6	5
Pretty Good Answer	641	4	146	66	122	0	296	6	1
Omitted Question	213	4	49	11	16	0	13	1	119
Had Courses	1376	20	463	190	279	0	407	13	4
Didn't Have Courses	802	43	442	106	75	0	41	95	0
Omitted Question	237	6	57	11	24	0	14	4	121
Question Very Clear	360	5	101	34	64	0	147	6	3
Clear Enough	712	8	253	103	144	0	198	4	2
A Little Confusing	744	16	377	122	116	0	96	16	1
Very Confusing	384	35	181	39	36	0	8	85	0
Omitted Question	215	5	50	9	18	0	13	1	119
Not Enough Time	140	9	70	10	12	0	4	35	0
A Little More Time	152	10	74	22	22	0	16	8	0
Right Amount of Time	990	31	456	151	163	0	155	30	4
A Little Too Much	405	5	129	64	83	0	116	7	1
Way Too Much Time	492	10	169	50	78	0	157	27	1
Omitted Question	236	4	64	10	20	0	14	5	119

Math Question 4: Car Stopping Distance
Counts of Subset with Multiple Choice Test
By Constructed Response Scale Score Level and Responses to Reaction Questions

Number with M.C. Test	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	2386	67	950	302	373	0	462	110	122
Difficulty Too Easy	147	3	35	13	25	0	66	2	3
Difficulty Easy	359	2	68	41	72	0	175	0	1
Difficulty Right	698	6	292	103	142	0	151	2	2
Difficulty Hard	641	20	348	100	94	0	48	31	0
Difficulty Too Hard	335	33	158	36	23	0	10	75	0
Omitted Question	206	3	49	9	17	0	12	0	116
Didn't Know Answer	705	49	379	88	64	0	28	97	0
Answer Partly Right	834	10	378	138	172	0	125	6	5
Pretty Good Answer	638	4	145	65	121	0	296	6	1
Omitted Question	209	4	48	11	16	0	13	1	116
Had Courses	1365	19	460	187	275	0	407	13	4
Didn't Have Courses	788	42	434	104	74	0	41	93	0
Omitted Question	233	6	56	11	24	0	14	4	118
Question Very Clear	356	5	100	33	62	0	147	6	3
Clear Enough	708	8	252	101	143	0	198	4	2
A Little Confusing	737	16	373	120	115	0	96	16	1
Very Confusing	374	33	176	39	35	0	8	83	0
Omitted Question	211	5	49	9	18	0	13	1	116
Not Enough Time	132	8	66	9	11	0	4	34	0
A Little More Time	152	10	74	22	22	0	16	8	0
Right Amount of Time	976	31	449	148	160	0	155	29	4
A Little Too Much	404	5	129	63	83	0	116	7	1
Way Too Much Time	490	9	169	50	77	0	157	27	1
Omitted Question	232	4	63	10	20	0	14	5	116

Math Question 4: Car Stopping Distance
Multiple Choice Math Test Means
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test Mean	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	50.7	33.3	43.8	50.3	59 2	*	65.6	37.2	43.9
Difficulty Too Easy	57.4	30.2	42.7	55.6	59.1	*	68.3	29.5	27.1
Difficulty Easy	61.4	35.1	48.7	57.0	62.0	*	67.6	*	31.6
Difficulty Right	51.4	24.1	43.2	48.8	57.7	*	64.5	45.1	34.8
Difficulty Hard	48.2	35.8	44.9	49.5	59.3	*	62.0	35.4	*
Difficulty Too Hard	41.8	34.5	41.0	47.6	56.9	*	52.1	38.0	*
Omitted Question	46.6	24.1	43.4	49.4	62.2	*	60.6	*	44.6
Didn't Know Answer	43.4	34.3	42.5	45.9	56.1	*	54.9	37.6	*
Answer Partly Right	51.2	34.2	44.9	51.1	58.3	*	63.1	41.4	30.2
Pretty Good Answer	59.4	29.1	44.8	54.6	61.5	*	67.9	29.3	31.6
Omitted Question	46.6	23.5	43.1	49.1	64.2	*	61.1	27.0	44.6
Had Courses	55.9	32.5	46.9	52.2	60.6	*	66.6	36.0	33.8
Didn't Have Courses	42.8	34.0	40.5	47.1	53.5	*	57.6	37.8	*
Omitted Question	46.4	31.3	44.0	47.9	59.8	*	61.0	29.0	44.3
Question Very Clear	57.6	33.2	46.7	53.6	59.4	*	67.3	41.1	27.1
Clear Enough	54.4	29.1	44.9	51.9	59.1	*	66.0	37.8	36.6
A Little Confusing	49.1	33.8	43.4	49.5	58.7	*	63.9	37.4	28.0
Very Confusing	42.6	35.1	41.8	46.6	59.5	*	54.9	37.0	*
Omitted Question	46.3	26.8	43.4	46.6	61.5	*	59.7	31.0	44.6
Not Enough Time	44.3	32.5	43.9	48.6	59.7	*	55.5	40.7	*
A Little More Time	47.5	25.1	45.4	53.0	54.2	*	61.1	33.1	*
Right Amount of Time	49.5	36.7	43.3	49.2	57.6	*	64.2	39.7	33.8
A Little Too Much	55.0	40.6	44.4	53.0	61.0	*	65.9	28.3	21.2
Way Too Much Time	54.2	31.1	44.3	49.8	61.4	*	68.0	33.1	26.1
Omitted Question	46.2	25.6	43.0	47.3	61.1	*	60.0	40.9	44.6

* No data for this cell.

Math Question 4: Car Stopping Distance
Multiple Choice Math Test Standard Deviations
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test S.D.	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	15.4	11.4	13.1	12.2	11.0	*	9.4	12.3	14.2
Difficulty Too Easy	16.5	7.2	16.0	11.0	13.0	*	6.9	6.8	5.3
Difficulty Easy	12.8	7.1	13.9	10.9	10.4	*	8.3	*	0.0
Difficulty Right	14.7	4.6	13.1	11.4	10.8	*	9.4	14.8	6.8
Difficulty Hard	14.1	10.6	12.8	12.7	10.3	*	10.5	11.4	*
Difficulty Too Hard	13.1	12.5	11.6	11.0	11.1	*	12.6	12.5	*
Omitted Question	15.1	0.8	13.4	14.2	12.1	*	10.5	*	14.2
Didn't Know Answer	13.4	12.3	12.4	10.9	11.7	*	13.3	12.4	*
Answer Partly Right	13.9	7.6	12.9	12.1	10.4	*	9.3	11.8	7.1
Pretty Good Answer	14.7	6.8	14.6	12.1	10.9	*	7.8	6.9	0.0
Omitted Question	15.1	1.2	13.4	13.0	9.4	*	10.2	0.0	14.2
Had Courses	14.2	12.3	13.0	12.3	10.0	*	8.3	10.5	5.0
Didn't Have Courses	13.6	11.1	12.2	11.6	12.5	*	14.2	12.7	*
Omitted Question	15.0	10.1	14.0	9.9	11.9	*	9.9	2.0	14.3
Question Very Clear	14.9	16.5	14.3	9.9	12.4	*	8.0	14.0	5.3
Clear Enough	14.4	9.3	12.9	11.8	10.8	*	8.1	13.2	5.0
A Little Confusing	14.6	8.2	12.7	12.8	10.0	*	11.8	9.9	0.0
Very Confusing	14.2	12.4	12.8	12.3	11.4	*	14.5	12.6	*
Omitted Question	14.9	3.3	13.4	10.5	12.1	*	10.5	0.0	14.2
Not Enough Time	14.9	13.2	14.4	10.4	6.5	*	16.5	14.1	*
A Little More Time	15.0	4.2	13.0	13.3	12.3	*	12.0	8.9	*
Right Amount of Time	14.4	12.9	12.6	11.5	11.0	*	9.1	12.2	5.0
A Little Too Much	14.7	3.2	12.6	13.3	8.8	*	9.6	4.9	0.0
Way Too Much Time	16.6	6.7	14.1	12.3	12.1	*	8.0	9.8	0.0
Omitted Question	14.6	2.6	13.0	10.1	11.7	*	10.1	9.8	14.2

* No data for this cell.

Science Question 1: Nuclear vs. Fossil Fuels
Counts of All Constructed Response Test Takers
By Constructed Response Scale Score Level and Responses to Reaction Questions

Sample Counts	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	2239	555	672	319	211	161	75	206	40
Difficulty Too Easy	71	9	23	12	8	10	5	3	1
Difficulty Easy	296	29	64	64	51	55	25	4	4
Difficulty Right	765	123	262	138	107	73	33	21	8
Difficulty Hard	677	210	238	90	38	22	12	67	0
Difficulty Too Hard	370	168	76	11	6	0	0	109	0
Omitted Question	60	16	9	4	1	1	0	2	27
Didn't Know Answer	855	349	224	60	31	6	4	181	0
Answer Partly Right	694	134	261	132	79	55	21	9	3
Pretty Good Answer	624	58	180	121	99	99	50	10	7
Omitted Question	66	14	7	6	2	1	0	6	30
Had Courses	1094	200	349	189	135	108	50	53	10
Didn't Have Courses	1072	338	311	127	72	51	25	148	0
Omitted Question	73	17	12	3	4	2	0	5	30
Question Very Clear	724	107	196	136	109	94	50	27	5
Clear Enough	812	174	306	128	81	58	24	38	3
A Little Confusing	455	186	136	49	17	8	1	55	3
Very Confusing	190	77	27	3	3	0	0	80	0
Omitted Question	58	11	7	3	1	1	0	6	29
Not Enough Time	97	34	16	4	6	4	1	31	1
A Little More Time	232	58	66	31	22	23	14	17	1
Right Amount of Time	1057	247	352	158	109	79	28	76	8
A Little Too Much	371	80	125	53	47	39	20	7	0
Way Too Much Time	383	112	102	66	25	13	12	52	1
Omitted Question	99	24	11	7	2	3	0	23	29

Science Question 1: Nuclear vs. Fossil Fuels
Counts of Subset with Multiple Choice Tests
By Constructed Response Scale Score Level and Responses to Reaction Questions

Number with M.C. Test	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	2200	544	661	315	209	158	75	200	38
Difficulty Too Easy	69	7	23	12	8	10	5	3	1
Difficulty Easy	290	29	62	62	51	53	25	4	4
Difficulty Right	754	121	256	137	106	72	33	21	8
Difficulty Hard	669	209	236	89	37	22	12	64	0
Difficulty Too Hard	363	164	75	11	6	0	0	107	0
Omitted Question	55	14	9	4	1	1	0	1	25
Didn't Know Answer	836	342	221	59	29	6	4	175	0
Answer Partly Right	689	133	258	132	79	54	21	9	3
Pretty Good Answer	613	57	175	118	99	97	50	10	7
Omitted Question	62	12	7	6	2	1	0	6	28
Had Courses	1079	199	342	186	134	106	50	52	10
Didn't Have Courses	1052	329	307	126	71	50	25	144	0
Omitted Question	69	16	12	3	4	2	0	4	28
Question Very Clear	713	106	191	134	109	92	50	26	5
Clear Enough	801	172	302	126	79	57	24	38	3
A Little Confusing	447	181	134	49	17	8	1	54	3
Very Confusing	184	75	27	3	3	0	0	76	0
Omitted Question	55	10	7	3	1	1	0	6	27
Not Enough Time	93	33	16	3	5	4	1	30	1
A Little More Time	229	57	65	31	22	23	14	16	1
Right Amount of Time	1044	243	348	155	109	78	28	75	8
A Little Too Much	364	79	122	53	46	37	20	7	0
Way Too Much Time	376	109	99	66	25	13	12	51	1
Omitted Question	94	23	11	7	2	3	0	21	27

Science Question 1: Nuclear vs. Fossil Fuels
Multiple Choice Science Test Means
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test Mean	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	23.5	19.3	23.7	26.3	27.7	29.6	31.3	18.4	20.2
Difficulty Too Easy	26.1	18.3	24.7	26.5	28.6	31.5	32.0	17.5	30.1
Difficulty Easy	27.1	20.7	23.9	27.7	29.9	30.5	31.5	13.6	19.6
Difficulty Right	24.7	19.7	24.0	26.4	27.0	28.9	31.5	17.9	19.8
Difficulty Hard	22.0	18.7	23.2	25.2	26.6	28.7	30.3	17.3	*
Difficulty Too Hard	20.5	19.3	23.5	23.5	27.1	*	*	19.5	*
Omitted Question	23.0	22.8	28.1	31.1	26.3	30.3	*	13.2	20.0
Didn't Know Answer	20.5	18.6	22.7	23.8	26.4	27.5	30.7	18.6	*
Answer Partly Right	24.4	20.3	23.8	26.2	27.1	29.0	30.4	13.0	12.5
Pretty Good Answer	26.7	20.6	24.6	27.4	28.5	30.0	31.8	18.3	21.0
Omitted Question	23.3	21.2	29.3	29.6	29.3	30.3	*	22.8	20.8
Had Courses	24.7	19.3	23.9	26.7	28.3	30.2	30.9	18.1	21.0
Didn't Have Courses	22.3	19.2	23.4	25.5	26.7	28.2	32.2	18.5	*
Omitted Question	22.9	22.1	26.6	32.6	25.4	31.6	*	21.9	19.9
Question **Very** Clear	26.2	20.8	24.9	27.2	28.3	30.0	31.5	22.0	20.8
Clear Enough	24.3	21.0	23.7	26.3	27.5	28.8	31.0	19.6	18.3
A Little Confusing	20.2	17.8	22.5	23.5	25.4	30.2	31.5	16.5	19.6
Very Confusing	17.9	16.6	19.7	23.8	24.0	*	*	18.0	*
Omitted Question	22.4	21.6	29.3	32.6	26.3	30.3	*	18.1	20.4
Not Enough Time	20.1	18.1	22.3	25.5	30.4	24.6	34.0	17.5	24.0
A Little More Time	24.2	19.0	24.3	25.4	30.7	29.7	32.7	16.9	11.9
Right Amount of Time	23.2	18.9	23.1	25.8	26.4	29.5	30.7	18.7	19.8
A **Little** Too Much	25.4	20.6	24.5	27.3	28.6	30.0	30.8	18.7	*
Way Too Much Time	23.3	19.3	24.6	26.6	28.8	29.5	31.7	19.1	30.1
Omitted Question	21.9	21.5	25.9	30.4	23.1	31.2	*	18.4	20.1

* No data for this cell.

Science Question 1: Nuclear vs. Fossil Fuels
Multiple Choice Science Test Standard Deviations
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test S.D.	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	6.6	5.5	5.7	5.3	5.3	4.6	3.9	5.7	7.1
Difficulty Too Easy	7.3	6.2	7.2	6.1	7.1	2.4	2.7	4.7	0.0
Difficulty Easy	6.2	6.1	6.1	4.7	3.9	3.9	4.0	1.0	4.6
Difficulty Right	6.4	5.7	5.5	5.2	5.4	5.2	4.0	5.5	8.0
Difficulty Hard	6.1	4.9	5.3	5.4	5.3	4.6	3.3	5.5	*
Difficulty Too Hard	6.1	5.7	6.1	6.1	4.6	*	*	5.8	*
Omitted Question	7.4	6.9	4.8	3.5	0.0	0.0	*	0.0	6.9
Didn't Know Answer	5.9	5.3	5.5	5.3	5.0	4.7	1.5	5.6	*
Answer Partly Right	6.2	5.6	5.6	5.2	5.1	4.2	4.5	2.1	1.4
Pretty Good Answer	6.2	5.6	5.8	5.1	5.4	4.8	3.6	4.8	7.6
Omitted Question	7.3	7.4	3.9	4.1	3.0	0.0	*	7.0	6.8
Had Courses	6.5	5.3	5.8	5.1	5.1	4.4	4.3	4.4	8.1
Didn't Have Courses	6.5	5.5	5.5	5.5	5.6	4.9	2.7	6.1	*
Omitted Question	7.2	7.1	5.7	2.5	3.3	1.3	*	6.0	6.6
Question Very Clear	6.2	5.5	5.5	5.3	5.1	4.9	4.1	7.4	6.9
Clear Enough	6.0	5.6	5.6	5.0	5.2	4.5	3.4	5.6	8.2
A Little Confusing	6.0	4.8	5.6	5.6	6.3	1.7	0.0	4.3	7.0
Very Confusing	5.0	3.9	5.3	0.7	4.1	*	*	5.3	*
Omitted Question	7.7	8.3	3.9	2.5	0.0	0.0	*	5.2	6.9
Not Enough Time	6.3	5.1	6.2	1.3	3.5	7.3	0.0	4.5	0.0
A Little More Time	7.1	5.1	6.0	5.4	3.6	5.5	2.5	6.4	0.0
Right Amount of Time	6.5	5.6	5.5	5.3	5.7	4.2	4.0	5.8	8.0
A Little Too Much	6.3	5.8	5.6	5.5	4.2	4.4	4.3	4.3	*
Way Too Much Time	6.5	5.1	5.5	5.1	4.6	4.4	3.6	6.0	0.0
Omitted Question	7.0	6.1	6.3	4.6	3.2	1.6	*	5.9	6.7

* No data for this cell.

Science Question 2: Eclipses
Counts of All Constructed Response Test Takers
By Constructed Response Scale Score Level and Responses to Reaction Questions

Sample Counts	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	2239	260	368	738	444	50	223	106	50
Difficulty Too Easy	75	3	7	10	26	1	27	1	0
Difficulty Easy	220	10	20	40	58	7	81	1	3
Difficulty Right	649	49	130	214	143	26	75	11	1
Difficulty Hard	824	108	151	321	161	12	37	34	0
Difficulty Too Hard	388	80	58	139	46	4	3	58	0
Omitted Question	83	10	2	14	10	0	0	1	46
Didn't Know Answer	1037	186	171	404	149	6	21	100	0
Answer Partly Right	711	43	148	246	180	24	68	1	1
Pretty Good Answer	395	19	46	69	102	20	134	3	2
Omitted Question	96	12	3	19	13	0	0	2	47
Had Courses	1015	87	169	302	233	32	168	22	2
Didn't Have Courses	1122	161	193	416	198	18	55	81	0
Omitted Question	102	12	6	20	13	0	0	3	48
Question Very Clear	828	75	119	247	193	24	154	16	0
Clear Enough	773	75	151	273	170	23	56	23	2
A Little Confusing	378	59	78	146	59	2	12	21	1
Very Confusing	170	40	18	54	14	1	0	43	0
Omitted Question	90	11	2	18	8	0	1	3	47
Not Enough Time	85	17	10	28	13	2	3	12	0
A Little More Time	126	19	19	56	22	2	3	5	0
Right Amount of Time	1048	120	194	373	207	18	93	41	2
A Little Too Much	406	31	83	113	91	19	64	4	1
Way Too Much Time	447	56	57	141	92	9	59	33	0
Omitted Question	127	17	5	27	19	0	1	11	47

Science Question 2: Eclipses
Counts of Subset with Multiple Choice Test
By Constructed Response Scale Score Level and Responses to Reaction Questions

Number with M.C. Test	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	**2200**	**254**	**365**	**723**	**440**	**50**	**218**	**104**	**46**
Difficulty Too Easy	70	2	7	9	25	1	25	1	0
Difficulty Easy	218	10	20	40	58	7	79	1	3
Difficulty Right	641	47	130	211	141	26	74	11	1
Difficulty Hard	810	105	148	313	161	12	37	34	0
Difficulty Too Hard	384	80	58	137	46	4	3	56	0
Omitted Question	77	10	2	13	9	0	0	1	42
Didn't Know Answer	1019	182	168	397	147	6	21	98	0
Answer Partly Right	702	41	148	240	180	24	67	1	1
Pretty Good Answer	390	19	46	68	102	20	130	3	2
Omitted Question	89	12	3	18	11	0	0	2	43
Had Courses	1001	84	169	296	232	32	164	22	2
Didn't Have Courses	1103	158	190	408	196	18	54	79	0
Omitted Question	96	12	6	19	12	0	0	3	44
Question **Very** Clear	816	73	119	245	190	24	149	16	0
Clear Enough	763	73	150	266	170	23	56	23	2
A Little Confusing	370	57	76	143	59	2	12	20	1
Very Confusing	167	40	18	52	14	1	0	42	0
Omitted Question	84	11	2	17	7	0	1	3	43
Not Enough Time	83	16	10	27	13	2	3	12	0
A Little More Time	123	18	19	54	22	2	3	5	0
Right **Amount** of Time	1035	117	194	365	206	18	92	41	2
A Little Too Much	402	30	81	113	91	19	63	4	1
Way Too Much Time	436	56	56	138	90	9	56	31	0
Omitted Question	121	17	5	26	18	0	1	11	43

Science Question 2: Eclipses
Multiple Choice Science Test Means
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test Mean	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	23.5	20.0	21.8	22.7	25.7	29.6	30.5	17.5	20.9
Difficulty Too Easy	26.7	22.2	19.3	21.8	28.3	30.8	29.6	12.3	*
Difficulty Easy	26.3	17.4	20.9	23.7	26.6	28.7	30.3	14.0	15.9
Difficulty Right	23.9	19.0	21.6	22.1	26.0	30.3	31.0	16.5	14.9
Difficulty Hard	23.3	20.3	22.4	23.1	25.4	30.1	30.0	18.0	*
Difficulty Too Hard	21.6	20.2	21.9	22.8	24.2	24.5	33.4	17.3	*
Omitted Question	21.5	21.4	15.5	22.0	22.1	*	*	30.9	21.4
Didn't Know Answer	22.2	20.2	21.5	23.0	25.0	29.4	29.6	17.4	*
Answer Partly Right	24.0	18.7	22.7	22.0	25.9	29.1	31.0	12.5	14.9
Pretty Good Answer	26.4	19.1	20.7	23.6	26.7	30.2	30.4	21.5	18.3
Omitted Question	21.9	22.5	20.4	22.8	24.0	*	*	18.9	21.1
Had Courses	24.2	19.7	21.5	22.2	25.7	29.4	30.4	17.7	20.3
Didn't Have Courses	23.1	20.1	22.1	23.2	25.9	30.0	30.7	17.5	*
Omitted Question	21.2	21.0	22.5	21.6	22.6	*	*	14.7	20.9
Question Very Clear	26.1	22.5	23.1	25.0	27.1	30.5	30.8	17.8	*
Clear Enough	23.5	20.3	21.9	22.4	25.6	29.6	30.0	20.4	20.3
A Little Confusing	20.7	18.2	20.3	20.9	23.0	27.5	28.3	15.2	10.9
Very Confusing	18.5	16.9	19.7	19.4	24.3	13.2	*	16.6	*
Omitted Question	21.1	21.8	15.5	21.1	19.3	*	32.7	20.3	21.1
Not Enough Time	19.7	17.1	17.2	20.5	21.5	30.1	32.4	16.7	*
A Little More Time	22.3	20.2	21.0	22.2	25.0	23.4	29.8	18.2	*
Right Amount of Time	23.4	20.1	21.6	22.9	25.6	29.5	30.6	17.4	12.9
A Little Too Much	24.8	19.2	23.2	22.6	26.0	28.9	31.2	15.8	25.7
Way Too Much Time	24.0	20.4	22.1	23.1	26.8	32.5	29.5	18.2	*
Omitted Question	21.8	21.4	21.2	22.6	25.4	*	27.6	17.0	21.1

* No data for this cell

Science Question 2: Eclipses
Multiple Choice Science Test Standard Deviations
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test S.D.	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	6.6	5.8	5.9	6.2	6.0	4.3	4.2	5.0	6.5
Difficulty Too Easy	6.5	6.2	5.4	6.9	5.3	0.0	4.2	0.0	*
Difficulty Easy	6.7	4.5	5.6	6.0	5.9	4.9	4.6	0.0	7.0
Difficulty Right	6.7	5.4	5.7	6.3	6.1	3.1	3.6	3.2	0.0
Difficulty Hard	6.3	6.0	6.1	5.8	5.7	3.8	4.7	5.8	*
Difficulty Too Hard	6.3	5.7	6.1	6.3	5.8	7.0	1.3	4.4	*
Omitted Question	6.8	6.1	0.7	8.0	7.9	*	*	0.0	6.3
Didn't Know Answer	6.3	5.7	5.8	6.1	5.7	7.8	4.2	4.9	*
Answer Partly Right	6.4	5.7	5.7	6.0	5.6	3.2	4.5	0.0	0.0
Pretty Good Answer	6.8	5.8	6.7	6.4	6.4	3.8	4.1	8.1	7.4
Omitted Question	6.9	6.5	6.9	7.2	8.3	*	*	0.3	6.4
Had Courses	6.6	5.3	5.8	6.1	5.7	3.5	4.3	3.9	5.4
Didn't Have Courses	6.7	6.1	6.0	6.2	6.1	5.3	4.0	5.3	*
Omitted Question	6.6	5.3	6.4	6.4	7.9	*	*	3.1	6.5
Question Very Clear	6.1	5.7	5.3	5.7	5.8	3.6	3.6	4.4	*
Clear Enough	6.4	5.6	6.1	6.2	5.6	3.6	4.7	5.3	5.4
A Little Confusing	6.2	5.4	5.8	5.9	5.7	0.9	7.3	4.0	0.0
Very Confusing	5.4	4.7	6.6	4.6	6.1	0.0	*	4.3	*
Omitted Question	6.6	5.7	0.7	6.8	6.8	*	0.0	8.0	6.4
Not Enough Time	6.4	4.9	4.3	6.4	5.8	4.3	2.8	4.5	*
A Little More Time	6.3	6.5	4.9	5.9	5.8	10.3	2.3	7.0	*
Right Amount of Time	6.5	5.9	5.6	6.2	5.5	3.9	4.4	5.1	2.0
A Little Too Much	6.8	5.4	6.5	6.0	6.7	3.0	3.6	4.8	0.0
Way Too Much Time	6.6	5.8	6.0	6.1	5.5	2.7	4.6	5.1	*
Omitted Question	6.7	5.3	8.4	6.5	7.8	*	0.0	3.6	6.4

* No data for this cell.

Science Question 3: Rabbit and Wolf Populations
Counts of All Constructed Response Test Takers
By Constructed Response Scale Score Level and Responses to Reaction Questions

Sample Counts	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	2239	478	515	576	170	117	94	202	87
Difficulty Too Easy	147	24	36	53	8	14	3	4	5
Difficulty Easy	419	61	76	161	37	43	34	5	2
Difficulty Right	807	172	204	258	62	48	38	17	8
Difficulty Hard	471	123	141	77	42	10	19	59	0
Difficulty Too Hard	297	87	52	22	17	2	0	117	0
Omitted Question	98	11	6	5	4	0	0	0	72
Didn't Know Answer	641	185	147	75	43	8	3	180	0
Answer Partly Right	736	175	212	212	54	38	34	5	6
Pretty Good Answer	753	106	149	283	69	71	57	12	6
Omitted Question	109	12	7	6	4	0	0	5	75
Had Courses	943	150	203	323	85	75	70	25	12
Didn't Have Courses	1180	315	303	246	80	41	24	171	0
Omitted Question	116	13	9	7	5	1	0	6	75
Question Very Clear	732	111	143	282	66	64	44	15	7
Clear Enough	698	154	182	200	55	42	40	22	3
A Little Confusing	431	130	128	74	30	8	8	50	3
Very Confusing	275	72	57	15	16	3	2	110	0
Omitted Question	103	11	5	5	3	0	0	5	74
Not Enough Time	81	26	13	6	4	1	1	29	1
A Little More Time	108	23	31	22	6	3	9	12	2
Right Amount of Time	1007	233	258	261	90	48	37	74	6
A Little Too Much	445	80	103	143	37	38	27	16	1
Way Too Much Time	460	98	102	132	30	26	17	52	3
Omitted Question	138	18	8	12	3	1	3	19	74

Science Question 3: Rabbit and Wolf Populations
Counts of Subset with Multiple Choice Tests
By Constructed Response Scale Score Level and Responses to Reaction Questions

Number with **M.C.** Test	Total	**0**	**1**	**2**	**3**	**4**	**5**	Imputed **0**	Blank
All Test Takers	2200	466	509	569	169	115	93	197	82
Difficulty Too Easy	143	23	35	52	8	14	3	3	5
Difficulty Easy	416	60	76	160	37	42	34	5	2
Difficulty Right	801	170	202	256	62	48	38	17	8
Difficulty Hard	456	117	139	75	41	10	18	56	0
Difficulty Too Hard	291	85	51	21	17	1	0	116	0
Omitted Question	93	11	6	5	4	0	0	0	67
Didn't Know Answer	625	181	143	73	42	8	3	175	0
Answer Partly Right	725	169	211	210	54	37	33	5	6
Pretty Good Answer	746	104	148	280	69	70	57	12	6
Omitted Question	104	12	7	6	4	0	0	5	70
Had Courses	933	149	201	320	85	73	69	24	12
Didn't Have Courses	1156	304	299	242	79	41	24	167	0
Omitted Question	111	13	9	7	5	1	0	6	70
Question Very Clear	723	111	141	278	66	62	44	14	7
Clear Enough	690	148	181	199	55	42	40	22	3
A Little Confusing	421	127	125	73	29	8	7	49	3
Very Confusing	268	69	57	14	16	3	2	107	0
Omitted Question	98	11	5	5	3	0	0	5	69
Not Enough Time	79	26	12	6	4	1	1	28	1
A Little More Time	105	22	31	22	6	3	8	11	2
Right Amount of Time	997	230	257	258	89	47	37	73	6
A Little Too Much	438	76	101	143	37	37	27	16	1
Way Too Much Time	450	95	100	128	30	26	17	51	3
Omitted Question	131	17	8	12	3	1	3	18	69

Science Question 3: Rabbit and Wolf Populations
Multiple Choice Science Test Means
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test Mean	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	23.5	20.0	22.6	26.2	26.7	29.3	30.7	18.1	20.3
Difficulty Too Easy	26.2	22.7	25.5	26.8	31.3	31.5	31.9	17.4	18.4
Difficulty Easy	26.4	21.4	24.6	27.2	29.9	29.0	30.1	17.6	17.9
Difficulty Right	24.1	20.0	22.6	25.9	27.3	28.6	31.9	16.8	16.7
Difficulty Hard	21.9	19.7	21.7	25.3	24.3	30.4	29.2	17.2	*
Difficulty Too Hard	19.4	18.9	19.6	22.4	21.1	31.4	*	18.8	*
Omitted Question	21.7	19.4	26.6	28.3	26.0	*	*	*	20.9
Didn't Know Answer	19.9	19.0	20.2	23.3	20.6	29.0	28.6	18.2	*
Answer Partly Right	23.9	19.9	22.8	26.0	27.7	28.6	30.1	16.1	17.2
Pretty Good Answer	26.4	21.9	24.7	27.0	29.7	29.7	31.2	18.1	17.7
Omitted Question	21.4	20.4	24.7	28.7	23.9	*	*	17.0	20.7
Had Courses	25.7	20.9	24.4	26.9	28.8	29.8	31.2	16.8	17.5
Didn't Have Courses	21.9	19.7	21.4	25.2	24.3	28.3	29.4	18.4	*
Omitted Question	21.2	18.6	23.9	26.8	27.7	33.5	*	14.9	20.7
Question Very Clear	26.2	22.4	25.1	26.6	29.6	29.1	31.4	19.2	18.6
Clear Enough	24.2	19.8	22.8	26.2	26.8	30.5	31.1	18.5	13.6
A Little Confusing	20.9	19.2	20.8	25.3	23.3	25.1	25.1	16.6	17.9
Very Confusing	19.2	18.4	19.5	21.4	21.2	26.1	28.9	18.6	*
Omitted Question	21.3	19.3	26.9	28.3	23.5	*	*	17.7	20.8
Not Enough Time	19.0	18.6	20.8	26.8	16.1	26.4	31.2	16.9	11.1
A Little More Time	22.5	18.5	21.9	27.0	27.0	29.2	30.1	15.2	12.8
Right Amount of Time	23.3	19.6	22.2	25.7	26.6	28.6	30.3	18.9	19.3
A Little Too Much	25.4	21.2	23.4	27.0	27.3	29.9	31.3	18.9	15.1
Way Too Much Time	23.8	21.0	23.1	26.0	27.8	29.5	30.5	18.0	19.0
Omitted Question	21.6	19.2	24.0	27.7	23.5	32.3	33.1	18.6	20.8

* No data for this cell.

Science Question 3: Rabbit and Wolf Populations
Multiple Choice Science Test Standard Deviations
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test S.D.	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	6.6	5.5	5.9	5.4	6.9	4.7	4.7	5.3	6.8
Difficulty Too Easy	6.2	6.4	5.8	5.1	2.5	2.5	1.3	3.4	6.4
Difficulty Easy	6.1	5.8	4.8	5.3	4.8	5.4	5.3	8.2	6.7
Difficulty Right	6.4	5.1	5.9	5.5	6.1	4.7	2.9	5.6	5.3
Difficulty Hard	6.6	5.6	6.0	5.1	7.8	2.9	6.2	4.5	*
Difficulty Too Hard	5.3	5.0	4.8	4.7	6.5	0.0	*	5.4	*
Omitted Question	7.1	5.3	6.6	5.4	7.8	*	*	*	6.9
Didn't Know Answer	5.7	5.2	5.3	5.2	6.8	5.4	3.8	5.2	*
Answer Partly Right	6.3	5.1	5.8	5.2	6.0	4.8	4.9	6.9	5.9
Pretty Good Answer	6.1	5.9	5.7	5.3	4.9	4.6	4.6	6.3	7.1
Omitted Question	7.0	6.7	7.3	5.0	6.6	*	*	3.2	6.8
Had Courses	6.4	5.7	5.7	5.4	5.6	4.7	4.7	4.2	6.6
Didn't Have Courses	6.3	5.3	5.7	5.3	7.2	4.7	4.6	5.4	*
Omitted Question	7.1	5.8	7.2	5.2	7.7	0.0	*	3.2	6.8
Question Very Clear	5.9	5.8	5.4	5.3	4.8	4.9	3.5	5.6	7.5
Clear Enough	6.5	5.1	5.7	5.4	6.6	3.6	4.9	5.4	1.2
A Little Confusing	6.1	5.2	5.8	5.2	7.2	5.5	6.4	4.7	4.7
Very Confusing	5.4	4.8	4.8	5.6	8.0	4.7	4.1	5.3	*
Omitted Question	7.1	6.1	7.2	5.4	7.6	*	*	7.3	6.8
Not Enough Time	6.2	4.8	8.1	5.6	2.8	0.0	0.0	4.6	0.0
A Little More Time	6.7	4.6	5.1	5.7	5.6	5.1	3.6	3.3	0.6
Right Amount of Time	6.5	5.4	5.7	5.5	6.8	5.4	5.0	5.4	5.9
A Little Too Much	6.4	5.7	6.0	4.9	7.0	4.1	4.9	5.8	0.0
Way Too Much Time	6.5	5.4	5.9	5.6	6.2	4.3	4.7	5.0	8.0
Omitted Question	7.1	5.5	6.9	5.0	7.6	0.0	0.6	6.1	6.8

* No data for this cell.

Science Question 4: Heating Curve
Counts of All Constructed Response Test Takers
By Constructed Response Scale Score Level and Responses to Reaction Questions

Sample Counts	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	2239	182	878	346	417	61	37	199	119
Difficulty Too Easy	201	6	67	30	73	7	8	9	1
Difficulty Easy	498	15	208	104	132	25	11	3	0
Difficulty Right	814	44	372	156	180	24	17	15	6
Difficulty Hard	340	49	175	42	25	5	1	43	0
Difficulty Too Hard	240	57	43	7	6	0	0	127	0
Omitted Question	146	11	13	7	1	0	0	2	112
Didn't Know Answer	475	96	151	30	26	2	0	170	0
Answer Partly Right	727	41	395	136	124	18	4	6	3
Pretty Good Answer	877	33	318	171	264	41	33	14	3
Omitted Question	160	12	14	9	3	0	0	9	113
Had Courses	1533	75	650	284	379	60	37	41	7
Didn't Have Courses	547	95	211	52	36	1	0	152	0
Omitted Question	159	12	17	10	2	0	0	6	112
Question Very Clear	878	32	329	182	256	32	26	20	1
Clear Enough	709	43	354	122	129	26	11	18	6
A Little Confusing	298	56	141	29	24	1	0	47	0
Very Confusing	195	41	37	6	6	0	0	105	0
Omitted Question	159	10	17	7	2	2	0	9	112
Not Enough Time	81	16	20	5	3	0	0	37	0
A Little More Time	76	11	26	10	16	4	1	8	0
Right Amount of Time	922	78	413	170	152	25	14	65	5
A Little Too Much	417	27	184	59	109	14	10	13	1
Way Too Much Time	557	35	211	89	133	16	12	60	1
Omitted Question	186	15	24	13	4	2	0	16	112

Science Question 4: Heating Curve
Counts of Subset with Multiple Choice Test
By Constructed Response Scale Score Level and Responses to Reaction Questions

Number with **M.C.** Test	Total	**0**	**1**	**2**	**3**	**4**	**5**	**Imputed 0**	Blank
All Test Takers	2200	180	861	345	413	60	36	192	113
Difficultv Too Easy	196	6	65	30	73	7	7	7	1
Difficulty Easy	492	15	205	103	130	25	11	3	0
Difficultv Right	802	43	364	156	179	23	17	14	6
Difficultv Hard	336	48	172	42	25	5	1	43	0
Difficultv Too Hard	234	57	42	7	5	0	0	123	0
Omitted Question	140	11	13	7	1	0	0	2	106
Didn't Know Answer	461	95	147	30	25	1	0	163	0
Answer Partly Right	719	40	388	136	124	18	4	6	3
Pretty Good Answer	866	33	312	170	261	41	32	14	3
Omitted Question	154	12	14	9	3	0	0	9	107
Had Courses	1513	75	639	283	376	59	36	38	7
Didn't Have Courses	534	93	205	52	35	1	0	148	0
Omitted Question	153	12	17	10	2	0	0	6	106
Question Verv Clear	865	31	324	181	254	32	25	17	1
Clear Enough	698	43	346	122	128	25	11	17	6
A Little Confusing	294	55	138	29	24	1	0	47	0
Very Confusing	190	41	36	6	5	0	0	102	0
Omitted Question	153	10	17	7	2	2	0	9	106
Not Enough Time	79	16	20	5	3	0	0	35	0
A Little More Time	73	11	26	10	15	4	1	6	0
Right Amount of Time	910	77	405	169	151	24	14	65	5
A Little Too Much	413	26	181	59	109	14	10	13	1
Way Too Much Time	545	35	205	89	131	16	11	57	1
Omitted Question	180	15	24	13	4	2	0	16	106

Science Question 4: Heating Curve
Multiple Choice Science Test Means
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test Mean	Total	0	1	2	3	4	5	Imputed 0	Blank
AU Test Takers	23.5	18.8	22.7	24.9	27.2	29.8	32.6	18.4	21.8
Difficulty Too Easy	26.5	19.2	24.5	26.9	28.4	31.1	32.6	21.7	11.1
Difficulty Easy	25.8	20.6	24.1	26.2	27.5	30.6	33.1	14.2	*
Difficulty Right	23.9	19.5	22.5	24.4	26.7	28.6	32.4	19.8	16.4
Difficulty Hard	21.0	17.3	21.4	22.3	25.9	29.9	31.0	18.1	*
Difficulty Too Hard	19.0	18.6	20.3	22.1	27.4	*	*	18.3	*
Omitted Question	22.2	21.1	20.5	25.8	32.5	*	*	16.9	22.3
Didn't Know Answer	19.4	17.9	20.1	21.9	24.0	30.1	*	18.3	*
Answer Partly Right	23.2	20.0	22.3	23.8	26.1	27.3	32.5	13.7	17.6
Pretty Good Answer	26.2	19.8	24.5	26.2	28.0	30.9	32.6	19.4	13.7
Omitted Question	22.2	19.5	20.1	27.3	29.1	*	*	21.5	22.2
Had Courses	25.1	19.9	23.5	25.6	27.7	30.0	32.6	19.4	15.6
Didn't Have Courses	19.4	17.8	20.3	20.7	21.9	19.6	*	18.1	*
Omitted Question	22.1	20.1	19.5	26.2	29.8	*	*	20.6	22.3
Question Very Clear	26.1	21.0	24.2	26.4	28.0	30.4	32.8	20.0	11.1
Clear Enough	23.7	19.5	22.8	23.8	26.3	28.6	32.3	19.9	16.4
A Little Confusing	19.6	17.3	19.8	20.8	23.8	32.0	*	18.5	*
Very Confusing	18.2	17.8	19.2	19.6	25.6	*	*	17.7	*
Omitted Question	22.3	21.5	20.8	25.8	25.7	33.7	*	20.8	22.3
Not Enough Time	18.2	18.0	20.2	22.0	17.3	*	*	16.8	*
A Little More Time	22.3	15.7	22.1	22.0	25.1	32.4	30.1	20.0	*
Right Amount of Time	22.9	17.9	22.2	23.8	26.5	28.9	31.1	19.1	16.6
A Little Ton Much	25.1	20.0	23.6	26.4	27.4	30.2	34.6	17.4	15.4
Way Too Much Time	24.6	20.2	23.3	26.4	28.3	29.7	33.0	18.2	11.1
Omitted Question	22.2	21.1	20.2	25.4	27.5	33.3	*	20.2	22.3

* **No** data for this cell.

Science Question 4: Heating Curve
Multiple Choice Science Test Standard Deviations
By Constructed Response Scale Score Level and Responses to Reaction Questions

M.C. Test S.D.	Total	0	1	2	3	4	5	Imputed 0	Blank
All Test Takers	6.6	5.5	6.0	6.1	5.5	4.7	3.1	5.6	7.0
Difficulty Too Easy	6.2	4.1	5.5	5.9	5.4	2.7	2.5	8.2	0.0
Difficulty Easy	6.2	6.0	6.0	5.5	5.8	4.3	3.1	3.5	*
Difficulty Right	6.4	5.9	6.1	6.2	5.4	5.5	3.3	5.9	1.3
Difficulty Hard	6.1	4.7	5.7	6.3	4.3	3.0	0.0	5.8	*
Difficulty Too Hard	5.5	5.0	5.7	6.2	6.7	*	*	5.2	*
Omitted Question	7.0	7.5	5.9	5.5	0.0	*	*	5.5	7.1
Didn't Know Answer	5.6	5.0	5.3	6.1	5.9	0.0	*	5.4	*
Answer Partly Right	6.1	5.5	5.8	6.1	5.4	4.4	2.0	3.3	0.3
Pretty Good Answer	6.3	5.8	6.1	5.8	5.4	4.4	3.2	6.3	1.9
Omitted Question	7.0	7.5	5.8	5.6	3.4	*	*	6.1	7.1
Had Courses	6.3	5.6	6.0	5.9	5.4	4.5	3.1	6.1	2.2
Didn't Have Courses	5.5	5.0	5.5	5.7	4.7	0.0	*	5.4	*
Omitted Question	7.0	7.1	5.5	5.8	2.7	*	*	5.7	7.1
Question Very Clear	6.0	5.4	5.7	5.6	5.1	3.9	2.5	6.6	0.0
Clear Enough	6.3	5.6	5.9	6.1	5.9	5.4	4.1	5.9	1.3
A Little Confusing	5.9	4.9	5.8	5.8	5.2	0.0	*	5.7	*
Very Confusing	5.2	4.6	5.6	5.4	5.1	*	*	4.9	*
Omitted Question	7.1	7.8	5.7	5.5	6.8	0.3	*	6.9	7.1
Not Enough Time	5.6	5.6	5.2	7.7	4.8	*	*	4.8	*
A Little More Time	6.9	3.0	5.9	7.7	6.1	2.6	0.0	6.0	*
Right Amount of Time	6.4	5.1	6.1	5.8	5.4	5.9	3.9	5.7	1.3
A Little Too Much	6.4	6.0	5.8	6.2	5.4	3.0	0.8	5.8	0.0
Way Too Much Time	6.6	5.1	6.1	5.7	5.3	3.9	2.1	5.2	0.0
Omitted Question	6.9	6.7	6.0	5.7	5.2	0.1	*	6.1	7.1

* No data for this cell.

Appendix C

- Reader Reliability Statistics, Analytic and Scale Scores

Agreement of Analytic Scores: Math Question 1

A: Latest Train						
	Second Reader	**0**	**1**	**2**	**3**	**4**
Counts of Scores Given by First Reader	0	0	0	0	0	0
0	0	1	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	13	1
4	0	0	0	0	2	274
288 Scores Agree Out of **291=99%**						

B: Least Amount **of** Time								
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0
0	0	3	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	79	0	4	0
4	0	0	0	0	1	4	9	0
5	0	0	0	0	10	2	26	1
6	0	0	0	0	1	0	0	151
263 Scores Agree Out of **291 = 90%**								

Agreement of Analytic Scores: Math Question 1

C: Latest Leave Home

	Second Reader	0	1	2	3	4	5	6	7
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0
0	0	9	0	1	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0
	0	0	1	0	0	0	0	0	0
3	0	0	0	0	38	3	0	13	1
4	0	0	0	0	1	4	8	0	0
	0	0	0	0	0	3	2	2	0
6	0	0	0	0	3	1	0	31	2
7	0	0	0	0	0	2	0	0	165

250 Scores Agree Out of **291=86%**

D: Formula for Winter

	Second Reader	0	1	2	3	4	5	6	7	8	9	A
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0	0	0	0
0	0	55	0	2	2	0	0	0	0	0	0	0
1	0	0	2	1	0	0	0	0	0	0	0	0
2	0	3	0	4	0	0	0	0	0	0	0	0
3	0	1	2	3	86	4	16	0	0	0	0	1
4	0	0	0	0	7	0	1	0	0	0	0	0
5	0	0	0	0	16	0	36	0	1	1	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	4	0	11	0	0	0
8	0	0	0	0	0	0	3	0	0	4	0	0
9	0	0	0	0	1	0	0	0	0	0	4	0
A	0	0	0	0	0	0	0	0	0	0	0	20

222 Scores Agree Out of **291=76%**

Agreement of Analytic Scores: Math Question 2

A: Two **9-Pound** Weights											
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Counts of Scores Given **by** First Reader	0	0	0	0	0	0	0	0	0	0	0
0	0	18	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	9	0	0	0	0	0	4
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0
8	0	1	0	0	4	0	0	0	0	4	2
9	0	0	0	0	2	0	0	0	0	3	193
224 Scores Agree **Out** of **241 = 93%**											

B: One **4-Pound Weight**											
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Counts of Scores Given **by** First Reader	0	0	0	0	0	0	0	0	0	0	0
0	0	20	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	0	0	0
3	0	0	0	0	48	0	0	2	2	2	5
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	5	1	4	0
7	0	0	1	0	2	0	0	1	3	0	0
8	0	0	0	0	0	0	0	0	0	3	0
9	0	1	0	0	3	0	0	0	0	2	134
213 Scores **Agree** Out of **241 = 88%**											

Agreement of Analytic Scores: Math Question 2

C: Additional Weight											
	Second Reader	0	1	2	3	4	5	6	7	8	9
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0	0	0
0	0	40	2	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	2	0	0	0	0	0	0
3	0	1	0	0	65	0	1	5	0	0	3
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	2	0	2	0	3	0	0
6	0	0	0	0	2	0	1	13	1	0	0
7	0	0	0	0	0	0	2	0	4	0	0
8	0	0	0	0	0	0	1	0	1	1	1
9	0	0	0	0	1	0	0	0	0	0	87
212 Scores Agree Out of 241 = 88%											

D: Balance Equation											
	Second Reader	0	1	2	3	4	5	6	7	8	9
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0	0	0
0	0	69	0	1	3	0	0	0	0	0	0
1	0	0	2	1	0	0	0	0	0	0	0
2	0	1	0	7	1	0	0	0	0	0	0
3	0	1	1	1	29	10	0	0	1	0	0
4	0	0	0	1	15	13	2	0	1	1	1
5	0	0	0	0	2	0	3	0	0	0	1
6	0	0	0	0	0	1	0	0	0	0	0
7	0	0	0	0	4	2	1	0	4	0	0
8	0	0	0	0	0	0	1	0	0	3	0
9	0	0	0	0	0	0	1	0	0	0	56
186 Scores Agree Out of 241 = 77%											

Agreement of Analytic Scores: Math Question 3

A: I aw lines							
	Second Reader	0	1	2	3	8	9
Counts of Scores Given by First Reader	0	0	0	0	0	0	0
0	0	44	0	0	2	0	7
1	0	1	0	0	0	0	0
2	0	0	1	1	0	0	0
3	0	1	0	0	1	0	3
8	0	0	0	0	0	4	0
9	0	3	0	0	3	1	199
249 Scores Agree Out of **271=92%**							

B: Area of Figure A										
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Counts of Scores Given by First Reader	0	0	0	0	1	0	0	0	0	0	0
0	0	36	0	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	2	0	0	0	0	0	0	0
3	0	0	0	0	33	0	5	0	4	1	1
4	0	0	0	0	0	4	5	0	0	0	0
5	0	0	0	0	0	1	21	1	12	0	1
6	0	0	0	0	0	0	2	1	1	1	0
7	0	0	0	0	0	0	4	0	32	2	0
8	0	0	0	0	0	0	0	0	1	0	0
9	0	0	0	0	2	1	1	0	0	0	93
222 Scores Agree Out of **271=82%**											

Agreement of Analytic Scores: Math Question 3

C: First Graph								
	Second Reader	0	1	2	3	4	5	9
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0
0	0	35	0	0	0	1	1	0
1	0	0	0	0	0	0	0	0
2	0	0	0	3	0	0	0	0
3	0	0	0	0	22	1	0	2
4	0	0	0	0	0	8	1	10
5	0	0	0	0	0	0	0	0
9	0	0	0	0	0	8	0	179
247 Scores Agree Out of **271=91%**								

C1: Small Rectangle					
	Second Reader	0	1	8	9
Counts of Scores Given by First Reader	71	6	2	0	3
0	6	26	1	0	1
1	3	2	1	0	1
8	0	0	0	0	1
9	0	0	0	1	146
244 Scores Agree Out of **271=90%**					

Agreement of Analytic Scores: Math Question 3

C2: Large Figure					
	Second Reader	**0**	**1**	**8**	9
Counts of Scores Given by First Reader	72	0	11	0	0
0	0	3	2	0	0
1	8	0	38	3	1
8	0	0	2	3	2
9	0	0	0	0	126
242 Scores **Agree** Out of **271 = 89%**					

Math Question 3C3: Subtract					
	Second Reader	**0**	**1**	**8**	**9**
Counts of Scores Given by First Reader	72	6	0	1	4
0	6	29	1	0	2
1	1	4	0	0	4
8	0	0	0	4	1
9	1	0	1	3	131
236 Scores Agree Out of **271 = 87%**					

Agreement of Analytic Scores: Math Question 3 (Continued)

D: Second Graph								
	Second Reader	0	1	2	3	4	5	9
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0
0	0	48	1	0	0	1	0	1
1	0	0	0	0	0	0	0	0
2	0	0	0	6	0	0	0	0
3	0	0	0	0	30	4	0	0
4	0	0	0	0	1	11	1	9
5	0	0	0	0	2	1	0	0
9	0	0	0	0	1	6	0	148
243 Scores Agree Out of **271 = 90%**								

D 1: Small Rectangle					
	Second Reader	0	1	8	9
Counts of Scores Given by First Reader	106	4	3	0	4
0	2	13	2	0	5
1	2	2	1	0	1
8	0	0	1	0	0
9	3	1	0	0	121
241 Scores Agree Out of **271 = 89%**					

Agreement of Analytic Scores: Math Question 3

D2: Large Figure					
	Second Reader	**0**	**1**	**8**	**9**
Counts of Scores Given **by First Reader**	107	0	7	2	1
0	1	1	2	1	0
1	4	0	22	14	2
8	0	0	2	10	0
9	2	0	1	1	91
231 Scores Agree Out of **271=85%**					

D3: Subtract					
	Second **Reader**	**0**	**1**	**8**	**9**
Counts of Scores Given by First Reader	107	3	0	1	6
0	4	15	1	0	5
1	0	2	0	0	4
8	0	0	0	0	2
9	3	1	0	1	116
238 Scores Agree Out of **271=88%**					

Agreement of Analytic Scores: Math Question 4

A1: Reaction Distance							
	Second Reader	**0**	**1**	**2**	**3**	**6**	**9**
Counts of Scores Given by First Reader	0	0	0	0	0	0	0
0	0	34	0	0	0	0	0
1	0	0	0	1	0	0	0
2	0	0	0	3	0	0	0
3	0	0	1	0	2	0	1
6	0	0	0	0	0	22	0
9	0	0	0	0	0	1	183
244 Scores Agree Out of 248 = 98%							

A2: Braking Distance							
	Second Reader	**0**	**1**	**2**	**3**	**6**	**9**
Counts of Scores Given by First Reader	**0**	**0**	**0**	**0**	**0**	**0**	**0**
0	0	35	0	0	0	0	0
1	0	2	0	0	0	0	0
2	0	0	0	3	0	0	0
3	0	0	0	0	1	4	0
6	0	0	0	0	0	62	2
9	0	0	0	0	0	0	139
240 Scores Agree Out of **248 = 97%**							

Agreement of Analytic Scores: Math Question 4

B: How Far to Stop										
	Second Reader	0	1	2	3	4	5	6	8	9
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0	0
0	0	49	0	0	0	0	0	0	0	0
1	0	0	1	1	0	0	0	0	0	0
2	0	0	1	2	0	0	0	0	0	0
3	0	0	0	0	16	0	0	3	0	0
4	0	0	0	0	0	40	0	0	0	0
5	0	0	0	0	0	1	13	0	0	0
6	0	0	0	0	0	1	7	51	1	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	3	0	0	0	58
230 Scores Agree Out of **248 = 93%**										

Math Question 4C: How Close to Collision									
	Second Reader	0	1	2	3	4	6	8	9
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0
0	0	79	0	0	0	1	0	0	0
1	0	1	0	1	0	0	0	0	0
2	0	0	1	10	0	0	1	0	0
3	0	0	0	0	37	0	14	0	0
4	0	0	1	0	3	25	0	0	0
6	0	0	0	2	4	3	33	0	2
8	0	0	0	0	0	0	0	1	1
9	0	0	0	0	0	0	0	0	28
213 Scores **Agree** Out of **248 = 86%**									

Agreement of Analytic Scores: Math Question 4

Math Question 4D: Explain							
	Second Reader	**0**	**1**	**2**	**3**	**6**	**9**
Counts of Scores Given by First Reader	0	0	0	0	0	0	0
0	0	123	0	0	0	0	0
1	0	0	4	0	0	1	0
2	0	0	0	9	0	0	0
6	0	0	0	1	0	29	31
9	0	0	0	0	0	12	38
203 Scores Agree Out of 248 = 82%							

Agreement of Scale Scores

Math Question 1: Scale Score

	0	**1**	**2**	**3**	**4**	**5**	**Blank Imputed**	**Blank Missing**
0	4	1	0	1	0	0	0	0
1	0	31	13	0	0	0	0	0
2	1	3	65	2	0	0	0	0
3	0	2	0	122	2	1	0	0
4	0	0	0	5	19	0	0	0
5	0	0	0	0	0	18	0	0
Blank-Imputed	0	0	0	0	0	0	1	0
Blank-Missing	0	0	0	0	0	0	0	0

260 Scores Agree Out of 291 = 89%
26 Scores Off By 1 Point = 9%
286 Scores Are Within 1 Point = 98%

Math Question 2: Scale Score

	0	**1**	**2**	**3**	**4**	**5**	Blank Imputed	Blank Missing
0	9	3	4	0	1	0	0	0
1	1	31	7	2	0	0	0	0
2	2	3	45	3	2	1	0	0
3	0	1	4	5	1	1	0	0
4	0	1	0	1	42	0	0	0
5	0	0	0	0	1	56	0	0
Blank-Imputed	0	0	0	0	0	0	5	0
Blank-Missing	0	0	0	0	0	0	0	9

202 Scores Agree Out of 241 = 84%
24 Scores Off By 1 Point = 10%
226 Scores Are Within 1 Point = 94%

Agreement of Scale Scores **(Continued)**

Math Question **3:** Scale Score								
	0	**1**	**2**	**3**	**4**	**5**	Blank Imputed	Blank Missing
0	18	3	2	0	0	0	0	0
1	3	26	8	3	0	1	0	0
2	0	2	27	2	1	2	0	0
3	0	1	2	16	1	1	0	0
4	0	0	0	1	16	8	0	0
5	1	0	1	0	2	98	0	0
Blank-Imputed	0	1	0	0	0	0	13	0
Blank-Missing	0	0	0	0	0	0	0	11

225 Scores Agree Out of 271 = 83%
32 Scores Off By 1 Point = 12%
257 Scores Are Within 1 Point = 95%

Math Question **4:** Scale Score								
	0	**1**	**2**	**3**	**4**	**5**	Blank Imputed	Blank Missing
0	6	0	0	0	0	0	0	0
1	0	99	7	0	0	1	0	0
2	0	4	34	0	0	0	0	0
3	0	3	0	30	0	1	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	30	0	0
Blank-Imputed	1*	0	0	0	0	0	7	0
Blank-Missing	0	0	0	0	0	0	0	25

232 Scores Agree Out of 248 = 94%
11 Scores Off By 1 Point = 4%
243 Scores Are Within 1 Point = 98%

Zero scores vs. blanks resulting in imputed-zeroes are counted as agreement.

Agreement of Analytic Scores: Science Question 1

Any Answer					
	Second Reader	0	1	2	3
Counts of Scores Given by First Reader	0	0	0	0	0
0	0	15	0	2	0
1	0	0	4	0	0
2	0	0	0	13	2
3	1	0	0	1	206
238 Scores Agree Out of 244 = 98%					

# Nuclear Advantages								
	Second Reader	0	1	2	3	4	5	6
counts of Scores Given by First Reader	25	54*	5	0	0	0	0	0
0	47*	25	22	4	0	0	0	0
1	2	10	33	6	0	0	0	0
2	1	2	5	2	0	0	0	0
3	0	1	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
186 Scores Agree Out of 244 = 76%								

*Counted as Agreement.

Agreement of Analytic Scores: Science Question 1

#] ıclear Disadvantages								
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Counts of Scores Given by First Reader	25	45*	7	0	0	0	0	0
0	41*	30	17	0	1	0	0	0
1	2	16	36	11	1	0	0	0
2	0	1	6	5	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
182 Scores Agree Out of 244 = 75%								

# Fossil Advantages								
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Counts of Scores Given by First Reader	25	56*	3	0	0	0	0	0
0	49*	28	19	3	0	0	0	0
1	4	13	33	4	1	0	0	0
2	0	0	1	5	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
196 Scores Agree Out of 244 = 80%								

*Counted as Agreement.

Agreement of Analytic Scores: Science Question 1

# ossil D, adva, ges								
	Second Reader	0	1	2	3	4	5	6
Counts of Scores Given by First Reader	25	28*	7	2	0	0	0	0
0	25*	14	13	1	0	0	0	0
1	8	12	72	9	1	0	0	0
2	0	0	6	16	0	0	0	0
3	0	0	0	3	2	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
182 Scores Agree Out of 244 =75%								

# Incorrect Statements								
	Second Reader	0	1	2	3	4	5	6
Counts of Scores Given by First Reader	25	25*	9	2	1	0	0	0
0	22*	14	16	7	1	0	0	0
1	9	20	27	14	3	0	3	0
2	3	6	10	9	5	0	0	0
3	1	0	2	5	1	0	1	0
4	0	0	0	3	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
123 Scores **Agree** Out of **244 =50%**								

*Counted as Agreement

Agreement of Analytic Scores: Science Question 1

Any Social Issues			
	Second Reader	**0**	**1**
Counts of Scores Given by First Reader	23	21*	4
0	22*	145	6
1	4	8	11
222 Scores Agree Out of 244 = 91%			

Any Alternative Energy Source			
	Second Reader	**0**	**1**
Counts of Scores Given by First Reader	23	23*	1
0	27*	159	4
1	0	1	6
238 Scores Agree Out of 244 = 98%			

*Counted as Agreement.

Agreement of Analytic Scores: Science Question 2

Solar Eclipse Diagram									
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0
0	0	31	0	1	0	0	0	0	0
1	0	0	0	0	1	0	0	0	1
2	0	1	0	2	0	0	0	0	0
3	0	0	0	0	8	0	1	0	3
4	0	0	0	0	1	20	0	2	3
5	0	0	0	0	1	1	24	3	1
6	0	0	0	0	2	2	1	6	3
7	0	0	0	0	3	2	3	1	195
286 Scores Agree Out of 323 = 89%									

Lunar Eclipse Diagram									
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0
0	0	48	0	1	0	0	0	0	0
1	0	0	1	0	1	0	1	0	0
2	0	1	1	4	0	0	0	0	0
3	0	0	0	0	9	2	2	3	1
4	0	0	0	0	3	47	3	1	0
5	0	0	0	0	1	7	67	1	3
6	0	0	0	0	0	2	1	4	1
7	0	0	0	0	3	1	2	1	100
280 Scores Agree Out of 323 = 87%									

Agreement of Analytic Scores: Science Question 2

Explanation										
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**	7	**8**
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0	0
0	0	55	2	1	0	0	0	0	0	0
1	0	0	10	1	26	1	0	1	0	1
2	0	0	4	33	1	0	0	0	0	0
3	0	0	43	2	80	0	1	5	12	3
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	1	0	0
6	0	0	1	0	3	0	1	3	2	0
7	0	0	0	0	5	0	0	3	19	0
8	0	0	1	0	2	0	0	0	0	0
200 Scores Agree Out of 323 = 62%										

Agreement of Analytic Scores: Science Question 3

Any Drawing					
	Second Reader	**0**	**1**	**2**	**3**
Counts **of Scores Given by First Reader**	0	0	0	0	0
0	0	49	0	0	0
1	0	0	1	0	3
2	0	1	0	0	0
3	0	0	2	0	237
287 Scores Agree Out of 293 = 98%					

Ph se of W olf Cu e						
	Second Reade	**0**	**1**	**2**	**3**	**4**
Counts of Scores Given by First Reader	45	7	0	0	0	0
0	1	23	1	5	5	2
1	0	1	7	5	1	1
?	1	6	2	95	3	1
?	0	6	1	15	34	1
A	0	3	0	0	3	18
222 Scores Agree Out of 293 = 76%						

Agreement of Analytic Scores: Science Question 3

Height of Wolf Curve					
	Second Reader	**0**	**1**	**2**	**3**
Counts of Scores Given by First Reader	45	7	0	0	0
0	2	14	3	7	6
1	0	4	17	5	0
2	0	3	2	15	4
3	0	6	0	5	148
239 Scores Agree Out of 293 = 82%					

Any Explanation					
	Second Reader	**0**	**1**	**2**	**3**
Counts of Scores Given by First Reader	0	0	0	0	1
0	2	39	0	0	1
1	0	2	0	0	2
2	0	0	0	7	0
3	0	0	5	1	233
279 Scores Agree Out of 293 = 95%					

Explain Lower Amplitude				
	Second Reader	**0**	**1**	**2**
Counts of Scores Given by First Reader	43	10*	0	0
0	5*	186	7	4
1	0	6	3	1
257 Scores Agree Out of 293 = 88%				

*Counted as Agreement.

Agreement of Analytic Scores: Science Question 3

Explain Wolf Lag				
	Second Reader	**0**	**1**	**2**
Counts of Scores Given by First Reader	43	10*	0	0
0	5*	212	1	6
1	0	3	0	1
2	0	10	0	2
272 Scores Agree Out of 293 = 93%				

Explain Rabbit Causes Wolf				
	Second Reader	**0**	**1**	**2**
Counts of Scores Given by First Reader	43	10*	0	0
0	5*	30	17	4
1	0	22	16	23
2	0	11	23	89
193 Scores Agree Out of 293 = 66%				

Explain Wolf Causes Rabbit				
	Second Reader	**0**	**1**	**2**
Counts of Scores Given by First Reader	43	9*	0	1
0	5*	191	6	16
1	0	2	0	2
2	0	10	3	5
253 Scores Agree Out of 293 = 86%				

*Counted as Agreement

Agreement of Analytic Scores: Science Question 4

Section A: Constant											
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0	0	0
0	1	53	0	0	0	0	0	0	0	0	0
1	0	0	8	0	3	0	0	0	0	0	0
2	0	0	0	16	0	0	0	0	0	0	0
3	0	0	3	0	21	35	7	9	0	0	0
4	0	0	0	0	16	82	10	11	0	0	0
5	0	0	0	0	2	6	47	8	8	0	0
6	0	0	1	0	1	6	3	8	4	0	0
7	0	0	0	0	0	3	3	4	14	0	0
8	0	0	0	0	0	0	0	0	0	0	1
					0	1	0	0	0	0	0

249 Scores Agree Out of 395 = 63%

Section B: Slope Up											
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0	0	0
0	1	55	0	0	0	0	0	0	0	0	0
1	0	0	6	0	4	0	0	0	0	0	0
2	0	1	0	10	0	0	0	0	0	0	0
3	0	0	2	0	17	17	3	9	0	0	0
4	0	0	1	0	12	137	16	15	14	0	0
5	0	0	0	0	3	11	5	3	2	0	0
6	0	0	0	0	1	9	1	20	2	0	0
7	0	0	0	0	0	10	1	1	5	0	0
8	0	0	0	0	0	0	0	1	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0
255 Scores Agree Out of 395 = 65%											

Agreement of Analytic Scores: Science Question 4

Section C: Constant											
	Second Reader	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Counts of Scores Given by First Reader	0	0	0	0	0	0	0	0	0	0	0
0	1	61	0	0	0	0	0	0	0	0	0
1	0	0	3	1	2	1	0	0	0	0	0
2	0	1	0	11	0	0	0	0	0	0	0
3	0	0	1	0	20	23	4	3	0	0	0
4	0	0	1	0	6	60	9	7	0	0	0
5	0	0	0	0	2	9	119	10	7	0	0
6	0	0	0	0	0	3	5	5	1	0	0
7	0	0	0	0	0	0	5	2	11	0	0
8	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	1	0	0	0
290 Scores Agree Out of 395 = 73%											

Agreement of Scale Scores

Science Question 1: Scale Score

	0	1	2	3	4	5	Blank Imputed	Blank Missing
0	48	17	1	1	0	0	0	0
1	12	42	14	7	2	1	0	0
2	0	8	11	7	4	0	0	0
3	0	3	3	6	8	2	0	0
4	0	3	2	5	5	2	0	0
5	0	0	0	1	1	7	0	0
Blank-Imputed	2*	0	0	0	0	0	16	0
Blank-Missing	0	0	0	0	0	0	1	2

139 Scores Agree Out of 244 = 57%
77 Scores Off By 1 Point = 32%
216 Scores Are Within 1 Point = 89%

Science Question 2: Scale Score

	0	1	2	3	4	5	Blank Imputed	Blank Missing
0	20	9	4	2	0	0	0	0
1	13	27	7	0	0	0	0	0
2	6	5	95	4	0	0	0	0
3	1	0	3	48	6	12	0	0
4	0	0	1	2	1	2	0	0
5	0	0	0	5	3	19	0	0
Blank-Imputed	1*	0	1	0	0	0	19	0
Blank-Missing	0	0	0	0	0	0	0	7

237 Scores Agree Out of 323 = 73%
54 Scores Off By 1 Point = 17%
291 Scores Are Within 1 Point = 90%

*Zero scores vs. blanks resulting in imputed-zeroes are counted as agreement.

Agreement of Scale Scores

Science Question 3: Scale Score								
	0	**1**	**2**	**3**	**4**	**5**	**Blank Imputed**	**Blank Missing**
0	33	3	12	0	0	0	3*	1
1	7	41	11	3	2	1	0	0
2	7	22	49	4	3	1	0	0
3	0	6	7	8	1	2	0	0
4	0	0	6	3	6	1	0	0
5	1	1	3	2	2	4	0	0
Blank-Imputed	1*	0	0	0	0	0	28	0
Blank-Missing	1	0	0	0	0	0	0	7

183 Scores Agree Out of 293 = 62%
61 Scores Off By 1 Point = 21%
244 Scores Are Within 1 Point = 83%

Science Question 4: Scale Score								
	0	**1**	**2**	**3**	**4**	**5**	**Blank Imputed**	**Blank Missing**
0	31	13	3	1	0	0	1*	0
1	12	115	28	12	1	0	1	0
2	0	11	13	5	2	0	1	0
3	1	10	8	49	3	5	0	0
4	0	1	4	2	0	1	0	0
5	0	1	0	3	0	0	0	0
Blank-Imputed	3*	0	0	0	0	0	33	0
Blank-Missing	0	0	0	0	0	0	0	21

267 Scores Agree Out of 395 = 68%
84 Scores Off By 1 Point = 21%
351 Scores Are Within 1 Point = 89%

*Zero scores vs. blanks resulting in imputed-zeroes are counted as agreement.

Appendix D

- Percentage of Multiple Choice and Constructed Response Items Omitted By Gender and Racial/Ethnic Group
  Mathematics and Science

- Mean Multiple Choice and Constructed Response Scores, By Gender and Racial/Ethnic Group
  Mathematics and Science

- Correlations of Constructed Response Scores and Omit Rates with Multiple Choice Test Scores and Background Variables
  Mathematics and Science

- Student Reaction Questions By Gender and Racial/Ethnic Group
  Mathematics and Science

Percentage of Multiple Choice and Constructed Response Items Omitted, By Gender and Racial/Ethnic Group: Mathematics

	Total	Male	Female	Asian	Hispanic	Black	White
Multiple Choice Test							
# Cases	2386	1235	1151	253	378	318	1404
% Omits	3.5	3.1	3.9	3.3	4.1	4.0	3.3
Constructed Response Test							
# Cases	2415	1250	1165	256	387	326	1412
Total % Omits							
Question 1	0.8	1.1	0.4	1.6	1.8	0.9	0.4
Question 2	7.0	8.2	5.7	5.1	10.1	16.0	4.5
Question 3	8.5	10.2	6.7	8.2	12.9	13.5	6.1
Question 4	9.8	10.8	8.8	9.0	14.5	18.7	6.6
All Questions	6.5	7.6	5.4	6.0	9.8	12.3	4.4
% Unresolved Omits							
Question 1	0.2	0.2	0.3	0.4	0.8	0.0	0.1
Question 2	2.2	3.0	1.5	2.0	3.1	5.8	1.2
Question 3	3.9	4.7	2.9	4.3	4.1	7.1	3.0
Question 4	5.2	6.2	4.1	4.3	5.9	11.0	3.8
All Questions	2.9	3.5	2.2	2.7	3.5	6.0	2.0

Percentage of Multiple Choice and Constructed Response Items Omitted,
By Gender and **Racial/Ethnic Group:**
Science

	Total	Male	Female	Asian	Hispanic	Black	White
Multiple Choice Test							
# Cases	2200	1103	1097	230	347	300	1302
% Omits	2.5	2.2	2.9	1.7	2.0	5.9	2.0
Constructed Response Test							
# Cases	2239	1125	1114	232	356	305	1321
Total % Omits							
Question 1	11.0	11.2	10.8	5.6	14.0	23.3	8.1
Question 2	7.0	5.5	8.4	4.3	9.6	16.4	4.5
Question 3	12.9	13.0	12.8	10.8	16.3	28.5	8.5
Question 4	14.2	15.4	13.0	9.1	16.9	27.5	11.1
All Questions	11.3	11.3	11.3	7.4	14.2	23.9	8.0
% Unresolved Omits							
Question 1	1.8	2.0	1.5	1.3	3.4	3.6	0.9
Question 2	2.2	2.4	2.1	1.7	3.1	5.9	1.1
Question 3	3.9	4.7	3.1	4.7	5.3	9.8	1.7
Question 4	5.3	6.1	4.5	4.7	5.3	10.8	3.9
All Questions	3.3	3.8	2.8	3.1	4.3	7.5	1.9

Mean Multiple Choice and Constructed Response Scores, By Gender and Racial/Ethnic Group: Mathematics

	Sample N	Mean	S.D.	S.E.	Effect Size
Multiple Choice Test					
Total	2232	51.17	15.33	0.32	
Male	1141	52.66	15.83	0.47	
Female	1091	49.61	14.63	0.44	-19.89
Asian	236	59.25	13.73	0.89	31.76
Hispanic	347	42.16	14.06	0.75	-79.69
Black	276	40.99	14.09	0.85	-87.33
White	1341	54.38	13.97	0.38	
Constructed Response Total Score					
Total	2232	11.32	5.34	0.11	
Male	1141	11.87	5.54	0.16	
Female	1091	10.74	5.06	0.15	-21.18
Asian	236	13.81	4.97	0.32	27.89
Hispanic	347	8.97	4.85	0.26	-62.70
Black	276	7.60	4.67	0.28	-88.41
White	1341	12.32	5.04	0.14	

NOTES:

1) Only test takers with multiple choice scores and responses to all four constructed response questions are included in this table.

2) Standard errors are computed using actual sample sizes.

3) Effect sizes are differences from a reference group (females compared with males; Asian, Hispanic, and black students compared with whites) in total group standard deviation units.

Mean Multiple Choice and Constructed Response Scores, By Gender and Racial/Ethnic Group: Mathematics (Continued)

	Sample N	Mean	S.D.	S.E.	Effect Size
Constructed Response Question 1					
Total	2232	2.72	1.19	0.03	
Male	1141	2.81	1.29	0.04	
Female	1091	2.64	1.07	0.03	-14.73
Asian	236	3.15	1.26	0.08	19.75
Hispanic	347	2.27	1.06	0.06	-54.74
Black	276	2.04	1.07	0.06	-73.96
White	1341	2.92	1.14	0.03	
Constructed Response Question 2					
Total	2232	2.96	1.76	0.04	
Male	1141	3.05	1.78	0.05	
Female	1091	2.88	1.74	0.05	-9.72
Asian	236	3.76	1.62	0.11	30.96
Hispanic	347	2.37	1.69	0.09	-47.63
Black	276	1.94	1.69	0.10	-72.40
White	1341	3.21	1.68	0.05	
Constructed Response Question 3					
Total	2232	3.41	1.93	0.04	
Male	1141	3.57	1.91	0.06	
Female	1091	3.24	1.93	0.06	-17.19
Asian	236	4.19	1.52	0.10	24.15
Hispanic	347	2.71	2.00	0.11	-52.38
Black	276	2.21	1.93	0.12	-78.23
White	1341	3.72	1.80	0.05	

Mean Multiple Choice and Constructed Response Scores, By Gender and Racial/Ethnic Group: Mathematics (Continued)

	Sample N	Mean	S.D.	S.E.	Effect Size
Constructed Response Question 4					
Total	2232	2.22	1.64	0.03	
Male	1141	2.44	1.74	0.05	
Female	1091	1.99	1.50	0.05	-27.62
Asian	236	2.71	1.67	0.11	14.83
Hispanic	347	1.62	1.38	0.07	-51.67
Black	276	1.41	1.19	0.07	-64.47
White	1341	2.47	1.68	0.05	

Mean Multiple Choice and Constructed Response Scores, By Gender and Racial/Ethnic Group: Science

	Sample N	Mean	S.D.	S.E.	Effect Size
Multiple Choice Test					
Total	2033	23.68	6.59	0.15	
Male	1011	24.37	6.58	0.21	
Female	1022	23.01	6.53	0.20	-20.56
Asian	215	24.98	6.44	0.44	-4.41
Hispanic	316	20.40	6.05	0.34	-73.79
Black	248	19.16	5.52	0.35	-92.58
White	1237	25.27	6.20	0.18	
Constructed Response Total Score					
Total	2033	6.47	3.93	0.09	
Male	1011	7.22	4.10	0.13	
Female	1022	5.72	3.60	0.11	-38.15
Asian	215	6.84	3.67	0.25	-10.83
Hispanic	316	4.96	3.10	0.17	-58.65
Black	248	4.21	3.26	0.21	-77.62
White	1237	7.26	4.01	0.11	

NOTES:

1) Only test takers with multiple choice scores and responses to all four constructed response questions are included in this table.

2) Standard errors are computed using actual sample sizes.

3) Effect sizes are differences from a reference group (females compared with males; Asian, Hispanic, and black students compared with whites) in total group standard deviation units.

Mean Multiple Choice and Constructed Response Scores, By Gender and Racial/Ethnic Group: Science (Continued)

	Sample N	Mean	S.D.	S.E.	Effect Size
Constructed Response Question 1					
Total	2033	1.37	1.40	0.03	
Male	1011	1.61	1.46	0.05	
Female	1022	1.13	1.30	0.04	-34.21
Asian	215	1.54	1.40	0.10	-5.56
Hispanic	316	0.88	1.09	0.06	-52.80
Black	248	0.66	1.01	0.06	-68.43
White	1237	1.62	1.46	0.04	
Constructed **Response Question 2**					
Total	2033	2.08	1.43	0.03	
Male	1011	2.46	1.48	0.05	
Female	1022	1.70	1.27	0.04	-53.30
Asian	215	2.13	1.47	0.10	-11.57
Hispanic	316	1.73	1.25	0.07	-39.15
Black	248	1.47	1.24	0.08	-57.28
White	1237	2.29	1.45	0.04	
Constructed Response Question 3					
Total	2033	1.47	1.38	0.03	
Male	1011	1.61	1.42	0.04	
Female	1022	1.33	1.32	0.04	-20.08
Asian	215	1.58	1.42	0.10	-5.09
Hispanic	316	1.09	1.22	0.07	-40.41
Black	248	0.96	1.18	0.07	-50.52
White	1237	1.65	1.40	0.04	

Mean Multiple Choice and Constructed Response Scores, By Gender and Racial/Ethnic Group: Science (Continued)

	Sample N	Mean	S.D.	S.E.	Effect Size
Constructed Response Question 4					
Total	2033	1.55	1.17	0.03	
Male	1011	1.54	1.20	0.04	
Female	1022	1.56	1.14	0.04	1.76
Asian	215	1.59	1.08	0.07	-9.52
Hispanic	316	1.26	0.97	0.05	-38.01
Black	248	1.13	1.09	0.07	-48.85
White	1237	1.70	1.22	0.03	

Correlations of Constructed Response Scores and Omit Rates with Multiple Choice Test Scores and Background Variables

Definition of Variables

QUEST 1-QUEST 4:	**Constructed Response Scale** Score, **Questions** 1-4
CR TOTAL:	**Constructed Response** Total Score (complete data cases only)
MC READ:	Multiple Choice Reading Test Score
MC MATH:	Multiple Choice Mathematics Test Score
MC SCI:	Multiple Choice Science Test Score
MC HIST:	Multiple Choice History/Citizenship/Geography Test Score
LEVEL 1-LEVEL 5:	Proficiency Scores Derived from Multiple Choice Test, Levels 1-5 (Math), Levels 1-3 (Science)
MC OMITS:	Number of Omitted Items on Corresponding Multiple Choice Test
MALE:	Male coded 1; Female coded O
ASIAN:	Asian coded 1; all other racial/ethnic groups coded O
HISPANIC:	Hispanic coded 1; all other racial/ethnic groups coded O
BLACK:	Black coded 1; all other racial/ethnic groups coded O
WHITE:	White coded 1; all other racial/ethnic groups coded O
SES:	Socioeconomic Status, continuous variable
PUBLIC:	Public Schools = 1; Catholic and NAIS Private Schools = O
URBAN:	Urban coded 1; Suburban coded O
OMIT Q1-OMIT Q4:	Constructed Response Question Omitted = 1; Answered = O
# OMITS:	Constructed Response Total Number of Omitted Items

NOTE: Correlation coefficients for continuous vs. dichotomous variables are r-biserial correlations; coefficients for two dichotomous variables are tetrachoric correlations.

Correlations of Constructed Response Scores with Multiple Choice Test Scores and Background Variables Mathematics

	QUEST 1	QUEST 2	QUEST 3	QUEST 4	CR TOTAL
MC READ	0.55	0.57	0.55	0.51	0.66
MC MATH	0.66	0.68	0.70	0.62	0.82
MC SCI	0.60	0.62	0.63	0.61	0.75
MC HIST	0.57	0.58	0.56	0.56	0.69
LEVEL 1	0.37	0.37	0.43	0.27	0.44
LEVEL 2	0.49	0.53	0.61	0.42	0.64
LEVEL 3	0.54	0.61	0.67	0.51	0.72
LEVEL 4	0.61	0.64	0.62	0.63	0.77
LEVEL 5	0.42	0.35	0.27	0.41	0.43
MC OMITS	0.01	0.02	0.02	0.02	0.02
MALE	0.07	0.05	0.09	0.17	0.13
ASIAN	0.08	0.15	0.13	0.06	0.16
HISPANIC	-0.31	-0.28	-0.30	-0.30	-0.38
BLACK	-0.42	-0.42	-0.46	-0.36	-0.51
WHITE	0.26	0.24	0.26	0.25	0.30
SES	0.33	0.32	0.33	0.32	0.40
PUBLIC	-0.32	-0.27	-0.25	-0.27	-0.34
URBAN	-0.09	-0.11	-0.14	-0.08	-0.12

Correlations of Constructed Response Omit Rates with Multiple Choice Test Scores and Background Variables Mathematics

	OMIT Q1	OMIT Q2	OMIT Q3	OMIT Q4	# OMITS
MC MATH	-0.38	-0.47	-0.42	-0.36	-0.28
MC OMITS	0.08	0.07	0.03	0.06	0.04
MALE	0.21	0.12	0.14	0.07	0.08
ASIAN	0.30	0.03	0.09	0.09	0.06
HISPANIC	0.35	0.25	0.24	0.26	0.19
BLACK	0.19	0.40	0.25	0.35	0.26
WHITE	-0.29	-0.28	-0.22	-0.27.	-0.18
SES	-0.22	-0.16	-0.18	-0.16	-0.12
PUBLIC	0.09	0.02	-0.01	-0.02	0.00
URBAN	0.22	0.25	0.26	0.29	0.17

Correlations of Constructed Response Scores with Multiple Choice Test Scores and Background Variables Science

	QUEST 1	QUEST 2	QUEST 3	QUEST 4	CR TOTAL
MC READ	0.48	0.33	0.43	0.42	0.57
MC MATH	0.46	0.40	0.47	0.46	0.61
MC SCI	0.55	0.48	0.51	0.48	0.70
MC HIST	0.52	0.40	0.43	0.41	0.61
LEVEL 1	0.34	0.28	0.28	0.31	0.41
LEVEL 2	0.51	0.45	0.47	0.44	0.65
LEVEL 3	0.49	0.43	0.48	0.41	0.63
MC OMITS	-0.10	-0.14	-0.12	-0.11	-0.16
MALE	0.21	0.33	0.12	-0.01	0.24
ASIAN	-0.02	-0.04	-0.02	-0.04	-0.06
HISPANIC	-0.29	-0.23	-0.23	-0.23	-0.33
BLACK	-0.38	-0.33	-0.29	-0.26	-0.42
WHITE	0.29	0.24	0.22	0.21	0.32
SES	0.34	0.26	0.28	0.26	0.39
PUBLIC	-0.18	-0.12	-0.10	-0.12	-0.18
URBAN	-0.05	-0.03	-0.02	0.01	-0.04

Correlations of Constructed Response Omit Rates with Multiple Choice Test Scores and Background Variables
Science

	OMIT Q1	OMIT Q2	OMIT Q3	OMIT Q4	# OMITS
MC SCI	-0.42	-0.39	-0.43	-0.36	-0.33
MC OMITS	0.27	0.32	0.25	0.23	0.21
MALE	0.01	-0.14	0.00	0.07	0.00
ASIAN	-0.10	-0.01	0.07	-0.06	-0.02
HISPANIC	0.19	0.23	0.23	0.15	0.17
BLACK	0.38	0.41	0.45	0.35	0.40
WHITE	-0.23	-0.27	-0.31	-0.20	-0.21
SES	-0.25	-0.25	-0.25	-0.16	-0.18
PUBLIC	0.15	0.04	0.03	-0.01	0.04
URBAN	0.13	0.19	0.12	0.09	0.10

Student Reaction Questions,
By Gender and Racial/Ethnic Group
Mathematics Question 1: Train Schedule

	Total	Male	Female	Asian	Hispanic	Black	White
Sample Size	2415	1250	1165	256	387	326	1412
How hard was the question?							
Too easy	6%	9%	3%	9%	3%	3%	7%
Easy	18%	21%	15%	25%	10%	9%	21%
About right	32%	30%	35%	33%	29%	35%	33%
Hard	32%	28%	35%	25%	40%	38%	29%
Too hard	9%	8%	10%	5%	12%	9%	9%
No response	3%	4%	3%	3%	6%	6%	1%
How good was your answer?							
Didn't know answer	24%	20%	27%	20%	34%	30%	20%
Partly right	36%	34%	39%	29%	36%	36%	38%
Pretty good answer	36%	42%	31%	48%	23%	27%	40%
No response	4%	4%	3%	4%	7%	6%	2%
Have you taken courses needed for question?							
Yes, enough background	73%	75%	72%	80%	56%	60%	80%
Have not taken course	22%	21%	24%	15%	36%	30%	18%
No response	4%	4%	5%	4%	9%	9%	2%
Did you understand the question?							
Very clear	19%	22%	14%	21%	10%	14%	22%
Clear enough	30%	33%	28%	37%	24%	23%	33%
A little confusing	39%	32%	45%	34%	46%	48%	35%
Very confusing	9%	8%	9%	5%	13%	8%	8%
No response	4%	4%	4%	4%	7%	8%	2%
Did you have enough time?							
Not enough time	4%	5%	3%	5%	7%	7%	3%
Needed a little more	9%	8%	10%	9%	12%	10%	7%
About right	46%	41%	51%	39%	48%	52%	46%
A little too much	19%	19%	18%	21%	12%	14%	21%
Way too much	18%	22%	14%	21%	12%	10%	21%
No response	4%	5%	4%	5%	9%	8%	2%

Student Reaction Questions,
By Gender and Racial/Ethnic Group
Mathematics Question 2: Balance Beam

	Total	Male	Female	Asian	Hispanic	Black	White
Sample Size	2415	1250	1165	256	387	326	1412
How hard was the question?							
Too easy	9%	13%	5%	19%	5%	5%	10%
Easy	19%	21%	17%	29%	13%	10%	21%
About right	29%	26%	32%	26%	32%	26%	30%
Hard	25%	22%	27%	14%	29%	30%	24%
Too hard	14%	12%	16%	8%	14%	20%	13%
No response	4%	5%	3%	4%	6%	9%	2%
How good was your answer?							
Didn't know answer	27%	23%	33%	18%	38%	36%	24%
Partly right	30%	30%	31%	23%	31%	30%	31%
Pretty good answer	38%	43%	33%	55%	25%	25%	42%
No response	4%	5%	3%	4%	6%	9%	2%
Have you taken courses needed for question?							
Yes, enough background	58%	62%	54%	73%	41%	40%	65%
Have not taken course	37%	33%	42%	23%	52%	50%	33%
No response	4%	5%	4%	4%	7%	10%	3%
Did you understand the question?							
very clear	25%	29%	20%	38%	15%	14%	28%
Clear enough	32%	32%	32%	36%	29%	27%	33%
A little confusing	23%	20%	26%	13%	31%	24%	22%
Very confusing	16%	14%	18%	9%	19%	25%	14%
No response	4%	5%	3%	4%	6%	10%	2%
Did you have enough time?							
Not enough time	6%	6%	5%	4%	7%	10%	5%
Needed a little more	6%	5%	8%	5%	9%	9%	5%
About right	42%	38%	46%	35%	48%	46%	40%
A little too much	19%	19%	19%	18%	14%	13%	22%
Way too much	22%	26%	18%	32%	14%	12%	25%
No response	5%	6%	5%	7%	8%	10%	3%

Student Reaction Questions, By Gender and Racial/Ethnic Group
Mathematics Question 3: Area of Figure Made of Rectangles

	Total	Male	Female	Asian	Hispanic	Black	White
Sample Size	2415	1250	1165	256	387	326	1412
How hard was the question?							
Too easy	22%	29%	16%	37%	12%	9%	26%
Easy	26%	26%	26%	30%	22%	14%	30%
About right	21%	17%	24%	13%	25%	27%	19%
Hard	15%	12%	18%	9%	21%	26%	12%
Too hard	11%	10%	12%	6%	13%	15%	10%
No response	5%	6%	4%	5%	7%	10%	3%
How good was your answer?							
Didn't know answer	21%	18%	24%	13%	31%	27%	18%
Partly right	20%	17%	24%	16%	20%	33%	18%
Pretty good answer	54%	59%	48%	66%	41%	30%	60%
No response	5%	6%	4%	5%	7%	10%	4%
Have you taken courses needed for question?							
Yes, enough background	74%	74%	74%	84%	59%	56%	80%
Have not taken course	20%	20%	21%	11%	32%	32%	16%
No response	6%	6%	5%	5%	9%	12%	4%
Did you understand the question?							
Very clear	45%	49%	40%	63%	27%	25%	51%
Clear enough	24%	21%	28%	20%	29%	25%	24%
A little confusing	13%	12%	14%	7%	20%	21%	10%
Very confusing	13%	12%	13%	6%	17%	18%	11%
No response	5%	6%	4%	5%	7%	10%	4%
Did you have enough time?							
Not enough time	5%	7%	4%	4%	5%	9%	5%
Needed a little more	5%	5%	4%	4%	6%	8%	4%
About right	34%	29%	39%	25%	46%	44%	30%
A little too much	16%	14%	18%	14%	15%	13%	17%
Way too much	34%	39%	30%	47%	19%	17%	40%
No response	6%	6%	6%	7%	9%	10%	4%

Student Reaction Questions,
By Gender and Racial/Ethnic Group
Mathematics Question 4: Car Stopping Distance

	Total	Male	Female	Asian	Hispanic	Black	White
Sample Size	2415	1250	1165	256	387	326	1412
How hard was the question?							
Too easy	6%	10%	2%	6%	3%	4%	8%
Easy	15%	19%	11%	18%	9%	7%	18%
About right	29%	27%	31%	30%	28%	30%	29%
Hard	27%	22%	33%	27%	33%	26%	26%
Too hard	14%	12%	16%	9%	17%	19%	13%
No response	9%	10%	7%	10%	10%	14%	7%
How good was your answer?							
Didn't know answer	30%	23%	37%	31%	40%	35%	26%
Partly right	35%	33%	37%	31%	35%	32%	36%
Pretty good answer	27%	33%	19%	28%	16%	19%	31%
No response	9%	10%	7%	10%	10%	14%	7%
Have you taken courses needed for question?							
Yes, enough background	57%	59%	54%	64%	42%	41%	64%
Have not taken course	33%	30%	37%	24%	48%	43%	28%
No response	10%	10%	9%	11%	10%	16%	8%
Did you understand the question?							
Very clear	15%	20%	10%	12%	8%	10%	18%
Clear enough	29%	30%	29%	31%	25%	26%	31%
A little confusing	31%	26%	35%	34%	36%	29%	29%
Very confirming	16%	14%	18%	14%	20%	20%	14%
No response	9%	10%	8%	10%	10%	14%	7%
Did you have enough time?							
Not enough time	6%	6%	5%	5%	7%	10%	5%
Needed a little more	6%	7%	6%	9%	7%	7%	5%
About right	41%	36%	47%	40%	46%	44%	39%
A little too much	17%	16%	17%	16%	14%	11%	19%
Way too much	20%	25%	16%	20%	13%	13%	24%
No response	10%	10%	9%	11%	11%	14%	8%

Student Reaction Questions,
By Gender and Racial/Ethnic Group
Science Question 1: Nuclear vs. Fossil Fuels

	Total	Male	Female	Asian	Hispanic	Black	White
Sample Size	2239	1125	1114	232	356	305	1321
How hard was the question?							
Too easy	3%	5%	1%	3%	3%	2%	4%
Easy	13%	18%	8%	10%	6%	9%	17%
About right	34%	39%	30%	41%	34%	24%	35%
Hard	30%	21%	39%	34%	37%	37%	26%
Too hard	17%	13%	20%	9%	17%	23%	16%
No response	3%	3%	2%	3%	3%	4%	2%
How good was your answer?							
Didn't know answer	38%	28%	49%	31%	50%	49%	34%
Partly right	31%	30%	32%	38%	29%	25%	32%
Pretty good answer	28%	38%	17%	28%	18%	21%	32%
No response	3%	3%	3%	3%	3%	5%	2%
Have you taken courses needed for question?							
Yes, enough background	49%	52%	46%	47%	44%	47%	51%
Have not taken course	48%	45%	51%	50%	51%	49%	47%
No response	3%	3%	3%	3%	5%	4%	2%
Did **you understand the question?**							
Very clear	32%	38%	26%	32%	23%	21%	38%
Clear enough	36%	35%	38%	44%	34%	32%	36%
A little confusing	20%	17%	24%	16%	29%	28%	17%
Very confusing	8%	7%	10%	5%	10%	13%	8%
No response	3%	3%	2%	3%	3%	5%	2%
Did **you have enough time?**							
Not enough time	4%	5%	4%	6%	5%	8%	3%
Needed a little more	10%	10%	10%	17%	9%	10%	9%
About right	47%	45%	50%	48%	53%	52%	45%
A little too much	17%	17%	16%	13%	11%	10%	21%
Way too much	17%	19%	16%	12%	17%	13%	19%
No response	4%	4%	5%	4%	6%	8%	3%

Student Reaction Questions,
By Gender and Racial/Ethnic Group
Science Question 2: Eclipses

	Total	Male	Female	Asian	Hispanic	Black	White
Sample Size	2239	1125	1114	232	356	305	1321
How hard was the question?							
Too easy	3%	6%	1%	3%	4%	1%	4%
Easy	10%	13%	6%	9%	7%	7%	12%
About right	29%	30%	28%	34%	31%	34%	26%
Hard	37%	33%	41%	36%	41%	29%	38%
Too hard	17%	14%	21%	14%	12%	21%	19%
No response	4%	4%	3%	3%	5%	8%	2%
How good was your answer?							
Didn't know answer	46%	38%	55%	47%	48%	47%	46%
Partly right	32%	33%	30%	34%	31%	30%	32%
Pretty good answer	18%	24%	11%	15%	15%	13%	20%
No response	4%	5%	4%	4%	6%	10%	2%
Have you taken courses needed for question?							
Yes, enough background	45%	48%	43%	44%	45%	43%	46%
Have not taken course	50%	48%	53%	52%	48%	46%	51%
No response	5%	5%	4%	4%	6%	10%	3%
Did you understand the question?							
Very clear	37%	40%	34%	39%	28%	24%	42%
Clear enough	35%	34%	35%	35%	35%	35%	35%
A little confusing	17%	15%	18%	16%	24%	22%	14%
Very confusing	8%	7%	9%	6%	6%	10%	8%
No response	4%	5%	3%	3%	6%	9%	2%
Did you have enough time?							
Not enough time	4%	5%	3%	5%	5%	5%	3%
Needed a little more	6%	5%	6%	9%	4%	7%	5%
About right	47%	45%	48%	50%	49%	50%	45%
A little too much	18%	17%	19%	15%	15%	12%	21%
Way too much	20%	22%	18%	16%	19%	14%	22%
No response	6%	5%	6%	5%	8%	12%	4%

Student Reaction Questions, By Gender and Racial/Ethnic Group Science Question 3: Rabbit and Wolf Populations

	Total	Male	Female	Asian	Hispanic	Black	White
Sample Size	2239	1125	1114	232	356	305	1321
How hard was the question?							
Too easy	7%	10%	3%	6%	6%	4%	7%
Easy	19%	22%	16%	18%	13%	14%	21%
About right	36%	32%	40%	39%	34%	36%	36%
Hard	21%	20%	22%	22%	26%	23%	19%
Too hard	13%	11%	16%	9%	15%	15%	13%
No response	4%	5%	4%	6%	6%	9%	2%
How good was your answer?							
Didn't know answer	29%	24%	33%	26%	39%	34%	25%
Partly right	33%	31%	35%	35%	29%	29%	34%
Pretty good answer	34%	39%	28%	33%	26%	26%	38%
No response	5%	6%	4%	6%	6%	11%	3%
Have you taken courses needed for question?							
Yes, enough background	42%	44%	41%	41%	33%	37%	46%
Have not taken course	53%	50%	55%	52%	60%	52%	51%
No response	5%	6%	4%	7%	7%	11%	3%
Did you understand the question?							
Very clear	33%	35%	30%	35%	24%	25%	37%
Clear enough	31%	30%	32%	35%	32%	26%	31%
A little confusing	19%	18%	21%	15%	23%	24%	18%
Very confusing	12%	11%	13%	9%	15%	14%	12%
No response	5%	5%	4%	6%	6%	11%	2%
Did you have enough time?							
Not enough time	4%	4%	3%	2%	4%	6%	3%
Needed a little more	5%	5%	5%	8%	4%	3%	5%
About right	45%	42%	48%	49%	49%	51%	42%
A little too much	20%	20%	20%	19%	15%	10%	24%
Way too much	21%	23%	18%	15%	19%	16%	23%
No response	6%	6%	6%	7%	8%	13%	4%

Student Reaction Questions,
By Gender and Racial/Ethnic Group
Science Question 4: Heating Curve

	Total	Male	Female	Asian	Hispanic	Black	White
Sample Size	2239	1125	1114	232	356	305	1321
How hard was the question?							
Too easy	9%	13%	5%	11%	8%	7%	9%
Easy	22%	24%	20%	24%	19%	15%	25%
About right	36%	31%	41%	41%	32%	34%	37%
Hard	15%	14%	17%	12%	23%	18%	13%
Too hard	11%	10%	12%	7%	12%	13%	11%
No response	7%	8%	5%	6%	6%	13	5%
How good was your answer?							
Didn't know answer	21%	19%	24%	14%	31%	27%	19%
Partly right	32%	29%	36%	36%	33%	27%	33%
Pretty good answer	39%	43%	35%	44%	30%	32%	43%
No response	7%	9%	6%	6%	6%	14%	6%
Have you taken courses needed for question?							
Yes, enough background	68%	66%	71%	79%	60%	58%	71%
Have not taken course	24%	25%	24%	15%	34%	29%	23%
No response	7%	9%	6%	6%	6%	14%	6%
Did you understand the question?							
Very clear	39%	40%	38%	45%	31%	30%	43%
Clear enough	32%	29%	34%	38%	32%	27%	31%
A little confusing	13%	13%	14%	7%	19%	17%	12%
Very confusing	9%	9%	8%	5%	11%	12%	8%
No response	7%	8%	6%	6%	6%	13%	6%
Did you have enough time?							
Not enough time	4%	4%	3%	1%	5%	8%	3%
Needed a little more	3%	4%	3%	6%	3%	2%	3%
About right	41%	38%	44%	46%	49%	45%	37%
A little too much	19%	17%	20%	19%	13%	12%	22%
Way too much	25%	28%	22%	21%	22%	20%	28%
No response	8%	10%	7%	7%	8%	14%	7%

Appendix E

- Description of Data File

Description of Data File

This report has focussed primarily on the rationale, design, score development, reliability, omit rates and score results for the HSES constructed response tests. A data file of test scores is available to researchers interested in exploring other issues and relationships. For example, the database would permit analysis of individual features of student responses and their relationship to student background characteristics, course-taking history, and school variables; alternative methods of constructing score scales from analytic scores; in-depth subgroup analyses; and comparisons of constructed response performance with selected subsets of multiple choice questions. The HSES constructed response data file can be linked to other files containing student questionnaires, demographic data, multiple choice test results, transcripts, and school information. The variables in the constructed response test file are:

- *Analytic Scores: the* individual features of student responses identified by the test readers. These scores are categorical; the codes do *not* represent a scale of increasing quality of response. Definitions of the codes for each analytic score can be found in Appendix A.

- *Scale Scores:* composites of the analytic scores that represent a continuum of performance. The algorithms used for constructing scales from the analytic scores are in Appendix A, Zero scores include imputations derived from the student response questions as described earlier. A total scale score is present only if there is no unresolved missing data on any of the four questions The following descriptions apply to the scale score points:

  0 = no understanding of the math/science concepts revolved

  1 = shows limited or rudimentary understanding: makes an attempt related to the problem

  2 = shows understanding of some parts of the problem, but with major error(s) or omissions

  3 = shows understanding of significant part of the problem, but answer is incomplete or includes incorrect information

  4 = successfully completes all but the most advanced part of problem

  5 = full understanding of the math/science involved in all parts (but may contain minor errors)

- *Student Reaction Questions*: the test takers' self report of the difficulty, clarity and timing of the questions, as well as their perceptions of their performance.

- *Second Reader Scores:* analytic and scale scores for the 10 percent of test questions that were scored by a second reader for the purpose of evaluating reader reliability.

Constructed response data is available for the 2415 students who took the mathematics test, and for 2239 science test takers. For the reasons discussed earlier, the sample weights available for the whole HSES sample do not apply to the subset of students who took the constructed response tests.

The analytic scores in the database have been edited to ensure that the readers' scores were recorded correctly, and for the correct test taker. They have intentionally *not* been edited to remove the relatively small numbers of inconsistencies or errors made by the readers. For example, users may find codes in the data that are

not within the range of codes specified in the scoring protocols. Or the reader may record a code "0" at the beginning of a problem, indicating that the entire question was blank, but then go on to score individual features of the response. Conversely, the reader may indicate that the question *was* answered, but then *not* record codes for the other analytic scores. This raw data has not been changed for two reasons. First, it is not practical (or perhaps even possible) to go back and determine which of two contradictory indications was the intended one. Second, the database was intended to serve as an *experiment in* constructed response testing as well as a measurement of student performance. As such, it is important for researchers to have access to the various types of human errors that may appear in order to design procedures that minimize these problems, and to be able to explore the costs and consequences of different ways of resolving them.

The data will be released as part of a full High School Effectiveness Study CD with electronic codebook (ECB), which is scheduled for release in 1997. The CD will contain two waves of student, school and teacher data (1990 and 1992), one wave of parent data (1992), plus high school transcript data and course offerings data. A data file user's manual describing the High School Effectiveness Study research and sample design will accompany the HSES CD. The dataset is a restricted-use dataset; as such, researchers will need to contact Cynthia Barton at NCES (202) 219-2199 to obtain a user license.

*U.S. Government Printing Office: 1997 - 426-599/70330