DOCUMENT RESUME

ED 376 198 TM 022 305

AUTHOR	Kupermintz, Haggai; And Others
TITLE	Enhancing the Validity and Usefulness of Large-Scale Educational Assessments: I. NELS:88 Mathematics Achievement.
INSTITUTION	Center for Research on the Context of Secondary School Teaching.
SPONS AGENCY	National Science Foundation, Washington, D.C.; Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	6 Jun 94
CONTRACT	G0087C0235; RED-9253068
NOTE	39p.; Version of a paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Achievement Tests; *Educational Assessment; Ethnic Groups; Factor Analysis; Grade 8; Grade 10; Longitudinal Studies; *Mathematics Achievement; National Surveys; Racial Differences; Regression (Statistics); *Scores; Secondary Education; Sex Differences; Socioeconomic Status; Student Attitudes; Testing Programs; Test Items; Test Use; *Test Validity; Thinking Skills
IDENTIFIERS	*Large Scale Assessment; *National Education Longitudinal Study 1988

ABSTRACT

This study demonstrates that the validity and usefulness of mathematics achievement tests can be improved by defining psychologically meaningful subscores that yield differential relations with student. teacher, and school variables. The National Education Longitudinal Study of 1988 (NELS:88) 8th- and 10th-grade math tests were subjected to full information item factor analysis. Math knowledge and math reasoning factors were distinguished at both grade levels. Regression analyses showed that student attitudes, instructional variables, course, and program experiences related more to knowledge, whereas gender, socioeconomic status, and some ethnic differences related more to reasoning. Teacher emphasis on higher-order thinking, student use of home computers, and early experience with advanced mathematics courses related to both dimensions. It is recommended that national educational surveys use multidimensional achievement scores, not total scores alone. One figure and eight tables illustrate the analysis. (Contains 35 references.) (Author/SLD)

CRC

Center for Research on the Context of Secondary Teaching
School of Education, CERAS Building, Stanford University, Stanford, CA 94305-3084

Enhancing the Validity and Usefulness of Large-Scale Educational Assessments: I. NELS:88 Mathematics Achievement

Haggai Kupermintz, Michele M. Ennis, Laura S. Hamilton, Joan E. Talbert, Richard E. Snow

Stanford University
June 6, 1994

Enhancing the Validity and Usefulness
of Large-Scale Educational Assessments:
I. NELS:88 Mathematics Achievement

Haggai Kupermintz, Michele M. Ennis,
Laura S. Hamilton, Joan E. Talbert,
Richard E. Snow

Stanford University
June 6, 1994

This paper is a revision of a paper presented at the American Educational Research Association meeting, New Orleans, April 4-8, 1994. This research was begun with support from the U.S. Department of Education (OERI Grant No. G0087C0235). The present analysis and report was supported by the National Science Foundation (Grant No. RED-9253068). The views and conclusions presented here are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Education, the National Science Foundation, or the U.S. Government.

The work reported here was a team effort; the authors share equally the responsibility for the research and its conclusions. Requests for reprints should be sent to Professor Richard E. Snow, School of Education, Stanford University, Stanford, CA 94305.

Enhancing the Validity and Usefulness of Large-Scale Educational Assessments:

I. NELS:88 Mathematics Achievement

Haggai Kupermintz

Michele M. Ennis

Laura S. Hamilton

Joan E. Talbert

Richard E. Snow

School of Education

Stanford University

Stanford, CA 94305

June 6, 1994

Submitted for Publication

Do not quote without permission

Abstract

This study demonstrates that the validity and usefulness of mathematics achievement tests can be improved by defining psychologically meaningful subscores that yield differential relations with student, teacher, and school variables. The NELS:88 8th and 10th grade math tests were subjected to full information item factor analysis. Math knowledge and math reasoning factors were distinguished at both grade levels. Regression analyses showed that student attitudes, instructional variables, course, and program experiences related more to knowledge, whereas gender, SES, and some ethnic differences related more to reasoning. Teacher emphasis on higher-order thinking, student use of home computers, and early experience with advanced math courses related to both dimensions. It is recommended that national educational surveys use multidimensional achievement scores, not total scores alone.

Enhancing the Validity and Usefulness of
Large-Scale Educational Assessments:
I. NELS:88 Mathematics Achievement

This study is one of a series examining the construct validity of mathematics and science achievement tests used in national survey research on school and teaching effects on educational outcomes. Our purpose was to determine whether or not psychologically meaningful subscores could be distinguished within these tests that might show differential patterns of relationship with educational variables. If so, then the usefulness of such surveys for informing educational policy and practice might be significantly enhanced.

Background

There is mounting federal and state commitment to national educational achievement tests as a way to monitor and promote the success of U.S. schools. Also, national educational surveys including achievement tests have been used increasingly in recent decades to estimate the effects of myriad societal and school variables on educational outcomes for the purpose of informing educational policy. Principal examples are the research on social inequalities in education (Coleman et. al., 1966; Jencks et. al., 1972) and on public versus private school effectiveness (Coleman, Hoffer, & Kilgore, 1982; Coleman & Hoffer, 1987; Chubb & Moe, 1990). As state-by-state comparisons using the National Assessment of Educational Progress (NAEP) are instituted (Glaser & Linn, 1992, 1993) and as one or another proposed national assessment system is developed, we can expect substantial and regularized influence of assessment data on state as well as federal educational policy. Also, as policy makers focus on different parts of the education problem, evaluations will need to address readiness to learn, opportunity to learn, gender and ethnic differences, subject matter differences, and a host of other special issues. Furthermore, the new goals for education not only specify higher standards of achievement, they emphasize deep understanding and higher-order reasoning as central educational outcomes. Assessments used to study these issues involve new psychological interpretations and thus depend on new construct validity arguments.

The survey methods and measuring instruments used in this work have been criticized, as have some of the interpretations derived from them (Alexander & Pallas, 1983; Cain & Goldberger, 1983; Haertel, 1988; Haertel, James, & Levin, 1987; Witte, 1990). Standardized achievement tests, of course, have been criticized across a broad front beyond their use in survey or policy research (e.g., Gifford, 1989a, 1989b), and there are many suggestions for improvement. Many of these proposals derive in one way or another from considering the cognitive psychology of educational achievement alongside the psychometrics of achievement assessment (see, e.g., Frederiksen, Glaser, Lesgold, & Shafto, 1990; Frederiksen, Mislevy, & Bejar, 1993; Snow & Lohman, 1989). There has also been a move to bring cognitive analysis to bear in the improvement of survey questionnaires and methods (Jabine, Straf, Tanur, & Torangeau, 1984). In both initiatives, the goal is to build a bridge between cognitive science and measurement science for the benefit of educational evaluation policy and practice.

The present research attempts to contribute to that bridge. It addresses the problem of how to improve the interpretation and use of tests and surveys that assess high school students' achievement in the diverse subject areas and classroom contexts of U.S. secondary education. It concentrates on some existing tests and questionnaires from the National Educational Longitudinal Study of 1988 (NELS:88), although the present approach could be used in building new kinds of instruments, or in reanalyzing old measures and data as well.

A schematic view of the NELS:88 survey. NELS:88 is the latest of three national longitudinal surveys conducted by the U.S. Department of Education and, compared with its predecessors, is especially designed to measure instructional practices and cognitive outcomes in four core subject areas. It began in Spring 1988 with a national survey and testing of 8th grade students. For details on design and initial analysis, see Horn, Hafner, and Owings (1992). The first follow-up of these students was conducted at 10th grade in Spring 1990, with a second follow-up at 12th grade in Spring 1992. Extensive student, parent, teacher, and school questionnaires were administered.

Figure 1 gives a schematic framework showing the main categories of variables available in the NELS:88 data structure and identifies with arrows the relationships studied and reported in the present paper. We are concerned here only with the analyses of mathematics tests into subscores at 8th and 10th grades and their prediction from student and teaching variables at these grades. A following paper provides the parallel analyses of the science tests. Data and analyses for 12th grade math and science will be added as our research continues. The project combines analyses of national NELS:88 data with our own small-scale studies of the same tests and questionnaires. Technical reports showing our exploratory large-scale and small scale work on the 8th grade math and science tests are available (see Ennis, Kerkhoven, & Snow, 1993; Snow & Ennis, in press). Another report, on technical issues and comparative methodology, is forthcoming (Kupermintz

& Snow, in preparation). Further reports on our analyses in other regions of the data structure of Figure 1 are also planned.

Figure 1 here

The NELS:88 tests. Rock, Pollack, Owings, and Hafner (1990) provided a detailed psychometric report for the NELS:88 base year test battery. In brief summary, they produced four multiple-choice tests, covering reading, history, math, and science, to fit into 1-1/2 hours of testing time and yet be sufficiently reliable to justify IRT scoring. The tests would thus allow adaptive testing at 10th and 12th grades, vertical scaling to study individual student gain across the three testings, and cross-sectional trend comparisons with the 10th-to-12th grade gains obtained in 1980-82 in the High School and Beyond study. The tests were also shown to be relatively unspeeded and free of gender and ethnic bias. In addition, the test developers paid attention to the need for educational and psychological diagnosis, as far as was possible within practical limits. They formed content testlets to allow subscores for specific content areas within subject-matter domains and, for reading and math, they designed proficiency level scores to provide somewhat richer cognitive interpretation than is usually available from standardized tests.

However, the test design also required that three forms of the 10th grade math test be administered, for students who scored in the low 25% (Form L), middle 50% (Form M), or high 25% (Form H) of the 8th grade distribution. Only 20 items were common to the 8th grade and all three 10th grade forms. Table 1 provides the math item identification numbers used in the 8th grade form and in each 10th grade form, keyed to the master item numbers we use to report our analyses. A verbal description of each item is also provided, reproduced here from Rock, et.al. (1990); we are prohibited by government rules from providing more detailed item descriptions. The 10th grade reading comprehension test was also designed and administered in two forms for two levels of 8th grade performance. However, for our math item analysis we used only the 8th grade total reading IRT scale score as a reference variable.

Table 1 here

It is not our intent here to criticize the NELS:88 psychometric work. Indeed, we think it an excellent example of how to use modern measurement theory and practice to make high quality, functional assessment instruments that are useful within the purposes and constraints of study designs such as those represented by NELS:88.

Nonetheless, it does seem worthwhile also to move outside the boundaries of conventional psychometric theory and practice to see whether more elaborated, richer cognitive interpretations might be gotten from the tests, and also the questionnaires. NELS:88 cost millions of dollars to conduct; millions more will be invested in data analysis. We think psychological interpretations of student achievement scores in these analyses can and should go beyond the typically vague, molar constructs of "amount of science knowledge possessed" or "level of mathematical ability reached." Traditionally, such statements offer the only interpretation one can attach to total scores from conventional achievement tests. In current cognitive instructional psychology, by contrast, "science knowledge" and "mathematical ability" are highly differentiated theoretical constructs. To whatever extent possible, these differentiations ought to be represented explicitly in educational assessments. The NELS:88 content and proficiency level subscores developed by Rock, et.al. (1990) are useful steps toward more detailed construct interpretations; the present research tests whether or not it is possible to go still farther in this direction.

Overview of Project Methodology

Emphasis on construct validity argument. Most achievement tests are still evaluated for validity only on content sampling considerations -- what the test measures is simply represented by the categories and tasks identified in the test specification tables, along with evidence that the test is without significant content bias. But validity theory has progressed far beyond the simple operationalism of a generation ago (see, e.g., Cronbach, 1988, 1989; Messick, 1989). Test users are entitled to expect evidence justifying recommended interpretations and ruling out rival alternatives. And these interpretations always involve hypothesized cognitive processes and structures, not just content distinctions. Indeed, most achievement tests are built using test specification tables that explicitly include process as well as content distinctions, though these distinctions are almost never validated empirically.

Some might argue that the prime issue, at least for educational tests such as NAEP, is indeed content sampling not psychological constructs. NAEP tests are simply designed to show the proportions of population groups who do or do not respond correctly to given problems, and the results are often reported at the item level, with the item itself in view. Even here however, as soon as one moves to interpretations about proficiency levels, or relations with other person or school characteristics, or explanations for changes in trend lines across years, psychological constructs enter the discussion. Furthermore, when tests such as those of NELS:88 become criteria for psychological and sociological hypothesis testing and modelling in the service of policy decisions, interpretations rest primarily on construct validity arguments.

Our approach thus puts the validity standard foremost. We use two kinds of studies in tandem: large-scale statistical analyses based on the item level test and questionnaire data in the national sample; and small-scale interview studies of the same tests and questionnaires using local student samples.

Large-scale studies. In support of the validity concern, a second emphasis reflects Tukey's (1969) view of data analysis as detective work. In conventional psychometrics, one usually chooses a strong measurement model, such as IRT, checks that the data can be made to fit it, and then proceeds with application. Similarly, a particular statistical model is often fit without regard for alternatives. But interest should also attach to features of the data that the chosen model leaves out. It is not usually appreciated that all models throw some kinds of information away; there are always tradeoffs. Therefore, our approach in the large-scale studies uses alternative statistical analyses to see if different methods lead to similar or different interpretations. We study various item statistics, different kinds of item intercorrelations and scatterplots, different component and factor analyses of items, different rotations of axes, and different hierarchical modelling techniques. Nonmetric multidimensional scaling adds a useful alternative to the metric methods. Our main aim is to test whether meaningfully distinguishable and interpretable subscores are possible within each test. If so, then the achievement construct represented by the test has been substantially elaborated.

Small-scale studies. The small-scale studies also emphasize detective work with multiple methods to reach understandable and usable subscores. Here we obtain small samples of local students, administer the tests and questionnaires to them under standard conditions similar to those used in the national study, and then interview them individually about their responses to each item and their knowledge and attitudes in each subject-matter domain. Retrospective reports about thought processes during the test are obtained. Detailed coding schemes are used to score these interviews. Also included for the science domain is a computer-based depth interview technique. In the plan for future studies are additional think-aloud, teach-back, and other performance tasks designed to assess student ability, domain knowledge, and thinking style. We can administer reference tests to locate students on national norms, and choose student participants to represent different categories of expertise and experience in the subject matter domains. Computing subscores for each student using formulae derived from the large-scale analyses allows individual profiles on the large-scale reference dimensions to be compared. The evidence on items and students from the small-scale studies is also used to corroborate or elaborate the large-scale analyses. There is thus a two-way street; the large-scale analyses can help direct the small-scale work and vice versa.

Given space limitations, the present report focuses on what we consider the best large-scale results. Details of supporting analyses of both large-scale and small-scale studies are merely summarized. Also, of course, validation is a continuing process. The derived subscores for each test and questionnaire are provisional; their meaning will be elaborated as multiple analyses continue across the 8th, 10th, and 12th grade data. The aim of all these methods is to reach richer substantive descriptions of what each item might represent psychologically for different individuals.

Samples. The NELS:88 base year sample consisted of 24,600 8th graders from 1052 schools. For some large-scale analyses, we used random student subsamples to allow for cross validation. Analyses of test item data were conducted initially using one-sixteenth random samples. Given comparable results in different samples, these analyses were then merged to form approximate quarter-samples and then half-samples of the 8th and 10th grade data. This allowed us to bring teacher questionnaire variables into the analyses in various combinations. In the NELS:88 sampling design, only certain pairs of teachers were surveyed for each student; no student had both math and science teachers reporting within a grade, and some students who had a math or a science teacher reporting in 8th grade may not have had a same-subject teacher reporting in 10th grade (see Horn, Hafner, & Owings, 1992). There was also substantial student attrition between grades. Thus, sample size varies across analyses. For example, one original subsample we chose for preliminary analyses consisted of 8th graders who had both math and English teachers reporting; it contained 6022 cases. This number was reduced to 5823 8th graders for test item dimensional analyses due to missing scores, reduced further to 4059 cases for analyses in which both 8th and 10th grade math scores were required, and reduced still further to 3044 cases for whom a math teacher reported at 10th grade. Since about half of the original subsample was lost due to these restrictions, we constructed subsamples that maximized our test and teacher data in the subject domains. For math, our student sample included all students for whom cognitive tests and math teacher responses were available at 8th and 10th grade; it consisted of 5460 cases. Attrition between grades is not random (see Ingels, et al., 1992), so we can expect substantive differences in results for analyses on different subsamples.

The samples for small-scale studies noted in this report included 50 8th and 10th graders. They were recruited as paid volunteers from local schools.

Scoring. For the present analyses, the NELS:88 achievement tests were scored at the item level simply as right versus wrong. The math proficiency level scores were also used in some analyses (Rock, et al., 1990). The IRT total score for Reading Comprehension was used to represent general verbal comprehension and reading ability. The IRT total scores for math and science were used for comparison purposes. Some of the NELS:88 teacher and student questionnaire items were rescored for our purposes using our own analyses of national data.

Analyses and Results

Our report of analyses and results is organized as follows. We first report the dimensional analyses that lead us to define psychologically different dimensions and subscores of mathematics achievement. Next we relate these subscores to the categories of the test specification table originally used to select items. We then form categories of predictor variables representing student, course, program, and instructional variables for use in regression models. Finally, we present regression results for main effects of these variables on the 10th grade math achievement subscores, as well as on math total scores.

Dimensional analysis. In previous exploratory work with the 8th grade data alone, five component subscales were identified for the mathematics test. We used conventional factor and principal component analysis on both Pearson and tetrachoric item correlation matrices, as well as nonmetric multidimensional scaling for this purpose. A small-scale study using local 8th graders also provided interviews concerning approaches and reactions to particular items. The five-component solution with varimax rotation seemed most satisfactory based on eigenvalues, the number of salient loadings on each dimension, and their interpretability using all available information. A unidimensional model could also be rejected using Stout's (1987) test. The upper panel of Table 2 shows these 8th grade mathematics subscales. Details of the analysis are given by Ennis, Kerkhoven, and Snow (1993).

Table 2 here

Although justifiable on both statistical and substantive grounds in the 8th grade data, these distinctions were regarded as provisional. The last two components were small; the last, particularly, was of doubtful reliability. And it is well-known that item-level correlational analyses of this sort can be distorted by distributional anomalies. When the test data for the 10th grade year became available, it was possible to analyze both grade levels in parallel and to include a cross validation by comparing factor loadings obtained in random subsamples. We also investigated full-information factor analysis (Bock, Gibbons, & Muraki, 1988) as a new and improved approach to this problem. This new method is now implemented in the TESTFACT computer program (Wilson, Woods, & Gibbons, 1991). It is based on specifying a multi-dimensional item response model for the test items. Thurstonian factor structure is employed for modelling the latent ability dimensions that affect the probability of correct responses. Combining the models for ability structure and item response provides a strong tool for estimating item loadings on distinct abilities. The statistical analysis of the test data is based on maximum likelihood estimates and iterative computation, where a principal factor analysis of the tetrachoric correlation matrix provides reasonable starting points for the iterative process. The procedure allows for combining information from different test forms, omitted responses, and adjustment for guessing. Factorial solutions can be rotated using orthogonal or oblique techniques. An explicit Chi-square statistic for improvement in fit is used to determine the number of statistically significant dimensions. Once a factor structure is decided upon, a Bayes estimate (average of the posterior ability distribution) generates scores on each ability dimension.

Along with exploratory use of conventional factor analyses applied to the 20 math items common to 8th grade and 10th grade test forms, and to the low, middle, and high 10th grade test forms separately, we also applied the full information factor method to the 8th grade and 10th grade test items, and to the combined data set for both grades. Since mathematical abilities were expected to be correlated, promax rotation was employed. Results proved highly satisfactory. We were able to obtain factor solutions that apparently provided the same two major dimensions in both grades. At the 8th grade level, the full-information procedure again identified the inferential reasoning factor shown in Table 2 and a knowledge-computation factor that combined the advanced and basic facts knowledge and computation factors of Table 2. It also isolated the specific counterexample reasoning items as a factor. At the 10th grade level, the full-information procedure again provided the inferential reasoning and knowledge-computation factors, along with a small factor different from that obtained at the 8th grade level. With the 8th and 10th grade levels analyzed together, the number of dimensions was decided using change in Chi-squared statistics for the fitted models. This indicated significant results for adding a second (Chi-squared change = 739.34, df = 57) and a third factor (Chi-squared change = 155.38, df = 56).

The lower panel of Table 2 defines the two major factors that appear common to both grade levels regardless of test form. Table 3 gives the factor loadings for each item at the two levels. In both years, items that were highly loaded on the first factor, symbolized as MR, required mainly inferential reasoning. Typically, no direct computation was called for in answering these items; rather, the correct solution was derived more from a logical argument. Although some knowledge of formal mathematical facts facilitated the solution, it alone was not sufficient to arrive at the correct answer. Most of the items in this factor introduced some kind of a scenario, not just mathematical expressions. On the other hand, items that were highly loaded on the knowledge-computation factor, symbolized here as MK, required mainly a straightforward computation or the use of mathematical knowledge. Most of these items were composed of a mathematical expression for which a one-step solution was possible. New items that were added to the 10th grade test had the same characteristics as 8th grade test items in each factor; of the 19 common items in the 10th grade reasoning factor, 16 were also classified as reasoning items in the 8th grade. The correlation between the two factor scores (over persons) was .72 and .75 at 8th and 10th grade, respectively. The correlation (over items) of the reasoning factor loadings from the 10th grade with the 8th grade reasoning factor loadings was .69; the correlation was -.74 with the 8th grade knowledge-computation factor. New items added to the knowledge-computation dimension appeared to be in the form of one-step solution mathematical expressions; 11 of the 18 items in this factor were classified in the 8th grade test as knowledge-computation. The correlation of 10th grade factor loadings was .68 with the 8th grade knowledge-computation factor and -.73 with the 8th grade reasoning factor.

Table 3 here

A noted, the full information factor analysis indicated the existence of a third factor in each grade (X and Y in Table 3). In the 8th grade, this small third factor was not easily identified as a distinct mathematical ability; in the earlier exploratory work, it was interpreted as counterexample reasoning (MC4 in Table 2). The three defining items were the only comparison items for which the correct answer was "the relationship cannot be determined from the information given", so item format may also be the key here. Respondents, when answering incorrectly, also tended to prefer one particular option (resulting in high values for the guessing parameters). In the 10th grade, the third factor seemed to capture a technical aspect of functional notation that was present in both of the two highest loading items. These items were also loaded substantially on the knowledge computation factor.

For present purposes, we have not retained these third factors, or the distinction between basic and more advanced knowledge and computation that seemed apparent in the 8th grade exploratory analyses. This does not mean that such distinctions can never be important. For example, counterexample or none-of-the-above reasoning could be usefully distinguished if represented by a more substantial range of items. Also, in other work on gender differences in the NELS:88 data, the subscore for basic facts computation favored females over males, on average, whereas an advanced knowledge subscore did not (see Snow & Ennis, in press). Furthermore, factors that appear at one grade level and not another, or item factor loadings that change between levels, should not be discounted. For example, variance arising from one source, e.g. reasoning, at one grade level could be replaced by variance from another source, e.g. knowledge, at a later grade level due to the effects of intervening instruction. We adopt the common two-factor representation as most expeditious for present purposes while recognizing that some finer distinctions may prove more useful for other purposes.

The common two-factor solution for the math tests at both 8th and 10th grade levels permitted computation of separate factor scores at each level expressed on a common psychometric scale, despite the involvement of different items at each level. This in turn allowed us to examine some alternative measures of achievement gain or growth. If one assumes that like factors at the two grade levels are indeed the same ability dimension, then simple gains or residual gains might be computed separately for knowledge and reasoning, or a three-parameter growth model might be considered (with parameters reflecting average baseline, average growth, and differential growth in knowledge versus reasoning). Clearly, a two-dimensional representation of learning gain would be valuable theoretically, even though the growth model approach is extremely limited when only two points in time are available. When 12th grade data can be added, however, the growth model approach may prove uniquely useful.

For the present, we think it preferable to consider like factors across grade levels as similar but not vertically equated dimensions. Regression analyses can simply treat 8th grade factors as predictors of 10th grade factors, with no special status in the spectrum of other predictors. No assumption of 8th to 10th grade equivalence is then needed. (see Kupermintz & Snow, forthcoming, for technical discussion of this issue, as well as comparative and cross validational analyses).

Our subscale interpretations were aided considerably by small-scale interview studies of local 8th and 10th graders concerning their approaches and reactions in trying to solve particular items. Their responses for math items included: guessing, eliminating some alternatives then guessing, computation, estimation, reasoning the item out, and immediately knowing the correct response. Apparent sources of incorrect responses were also identified, including: computational errors, carelessness, lack of knowledge, incorrect knowledge, failure to reason, and incorrect reasoning. Instances in which examinees arrived at the correct answer using incorrect reasoning or knowledge were also noted. It did appear that MR items called more on reasoning strategies and produced errors due to incorrect reasoning as well as lack of knowledge. On the other hand, MK items more often involved computation or estimation strategies, with errors more often arising from computational mistakes as well as faulty knowledge (see Ennis, Kerkhoven, & Snow, 1993).

In addition to assisting with the interpretation of our math subscores, the interviews identified some problem items. Of particular concern are several items for which the correct answer can be obtained using incorrect reasoning or knowledge. Examples are given in Ennis, Kerkhoven, & Snow (1993). We decided to leave these in the large-scale analyses, but to keep track of them in further analyses and interpretation. Of the four items that might be questioned on this basis, none was crucial to defining a factor at either grade level. Three of the four had low loadings on all factors at one or both levels. Thus it seems that the results were not negatively influenced by retaining these items.

Relation to test specifications and proficiency scores. As noted earlier, the NELS:88 math test was developed using a typical process-by-content test specification table. Separate content scores were computed. Examinees were also assigned to proficiency levels according to their performance on three four-item subsets. An important question concerns how our proposed math subscores relate to these proficiency levels and to the process and content dimensions of the test specification table.

Table 4 shows the items as given in the original 8th grade specification table (see Rock, et al., 1990) and indicates the subscore to which we assigned each item. Items that mark one of the three proficiency levels are also designated (1), (2), or (3). Our subscores do not represent any of the original process or content distinctions cleanly. Although there is a preponderance of MK items in the skills-knowledge row and of MR items in the understanding-comprehension and problem solving rows, each subscore spans both process and content categories. Comparing subscores and proficiency levels, each proficiency level appears to be a mixture of knowledge and reasoning at the 8th grade level. We do not yet have a test specification table for the 10th grade items. In the 10th grade factor analysis, however, Proficiency Level 1 items load on the MK or the Y factor, Proficiency Level 2 items all load on MK, and all but one Proficiency Level 3 item load on MK. Also, many of the proficiency items show relatively low factor loadings at 8th grade, as though the factor solution did not represent them well. It would appear that the psychology of the proficiency levels is not well understood and may change across grades, with reasoning variance tending to become knowledge variance.

Table 4 here

Student, course, program, and instructional variables as predictors. In addition to the student achievement factors from 8th grade, several other categories of variables at one or both grade levels provided predictors of achievement factors at 10th grade. As shown in Figure 1, 8th grade student reading ability was one index of general prior achievement. Gender, ethnicity, and socio-economic status indices were obvious additions. Further, the student survey included questions on learning opportunities outside of school, such as visits to museums, computer availability, parental help with homework, and amount of TV watching, as well as formal courses taken in school during present and past years. In addition to courses taken, we have included an index for academic program, contrasting advanced, academic, and general-vocational tracks. A separate index for those students in programs for the gifted and talented was also included.

Both the 8th and 10th grade student questionnaires also contained locus of control and self esteem scales, as well as items concerned with anxiety about asking questions in class and attitude toward different subject areas. Our own factor analysis of the 8th grade national sample data (not reproduced here) produced component scores for positive and negative statements about self esteem and a separate score reflecting attributions of success to luck versus hard work (called here Pos Self 8, Neg Self 8, and Luck 8).

The variable labeled "positive self-esteem" represents a factor reflecting positively worded statements (such as "I am a person of worth"); the "negative self-esteem" factor represents negatively-worded statements (such as "at times, I think I am no good at all"). We also produced separate scales for math anxiety and positive attitude toward math (called Math Anx and Math Att 8). We expect that these student affective indicators will be important criterion measures along with the cognitive subscores in subsequent longitudinal analyses; however, we have used them only as predictors in the present work.

From the 10th grade teacher questionnaire, we have used teacher report on student absence and track level. We also subjected teacher reports of instructional practices to a series of principal component analyses with varimax rotation (not reproduced here) to identify distinct instructional treatment dimensions. These are defined in Table 5. Of particular interest here was the teaching dimension called "emphasis on higher-order thinking", including emphasis on conceptual structure and problem solving. In the present work and in related studies (e.g., Raudenbush, Rowan, & Cheong, 1993; Talbert & DeAngelis, forthcoming) we have investigated this variable as an indicator of teaching for understanding as opposed to memorization and computation. As a complement to this index, we also included as a predictor variable student report of teacher emphasis on understanding from the 10th grade student questionnaire.

Table 5 here

Regression analyses for main effects. To explore the degree to which predictors of academic success might be differentially important for different components of math achievement, regression analyses were carried out at the subscore level as well as for the total math IRT scores.

Four regression models were computed for each of the two major math subscores in 10th grade. Prior achievements represented by 8th grade math and reading scores were entered at the first step in each model. The student model then included student SES, gender, ethnicity, absenteeism, and the 8th grade affective factors. A second model examined course and program variables. A third included all the teacher and instructional factors. Finally, a fourth model included indicators of opportunity to learn outside of school.

All models were fitted by ordinary least squares on random halves of the sample and compared. The present report shows combined results only for the total sample. A fixed alpha level of .01 (corresponding to a t-ratio of about 2.5 in this sample) was used to designate statistical significance. All variables (except dummy indicators) were standardized prior to model fitting.

Table 6 here

Table 6 presents the regression results for the MK factor as dependent variable. As expected, 10th grade math knowledge was predicted by 8th grade reading and math reasoning as well as math knowledge. Higher achievement on MK was also associated with higher SES and Asian ethnicity. It is interesting to note that no gender or other ethnicity effects were found for this factor. As for the affective variables, higher math knowledge was associated with a positive attitude toward math and, to a lesser extent, with emphasizing hard work over luck and reporting a less negative self esteem. There was also a negative effect due to student absence.

Strong effects were found for the course and program indicators. Students who scored higher on math knowledge were more likely to have taken algebra (I or II) and geometry courses, while lower achievement was strongly related to having taken general math courses. Students in the advanced track scored somewhat higher on math knowledge, whereas students in the general-vocational track scored much lower, compared with students in the academic track. Also important was having taken algebra in 8th grade.

The best instructional treatment predictor of math knowledge was teacher emphasis on higher order thinking. Student report of teacher emphasis on understanding was also significant. Higher student scores on math knowledge were also related to teacher reports of more use of traditional instruction, less use of individualized instruction, and more time assigned to homework. A negative effect was associated with emphasis on math applications.

Finally, students who reported visiting science museums and having a computer at home for their educational use showed higher math knowledge scores. Students with lower scores received more help from parents on homework.

Table 7 here

Table 7 presents the regression results for the MR factor. Again as expected, 8th grade reading and math knowledge as well as math reasoning predicted 10th grade math reasoning. Note however that prior math reasoning appears to be a stronger predictor of later math knowledge than prior math knowledge is of later math reasoning. Also as before, students with higher SES performed better on reasoning. Large gender and ethnicity effects were found; males and non-black students showed higher math reasoning scores on average. No affective variables showed significant relation to reasoning. These patterns stand in contrast to the findings for math knowledge.

Course effects were less pronounced on math reasoning compared with math knowledge. Again, general math courses were associated with lower scores, whereas Algebra II and Geometry courses were associated with higher scores. It is also noteworthy that taking Algebra I was associated with lower math reasoning scores; this effect became significant in the overall model and is analyzed further below. No significant track effects were found.

As with math knowledge, student achievement on math reasoning was associated with teacher reports of more emphasis on higher order thinking and less emphasis on individualized instruction, though both effects appeared to be weaker. Here also student reports of teacher emphasis on understanding were not significant.

Having a computer in the home for educational use appeared to have a marked relation to reasoning. Again, parent help with homework showed a negative relation. Museum visits had little relation to reasoning.

The next stage of analysis consisted of fitting two overall regression models, for MK and MR separately, to include all of the significant predictors from the previous analyses. Table 8 presents these overall models for the 10th grade MK and MR factor scores, along with comparable results for the total math IRT score.

Table 8 here

For the knowledge factor, after taking other significant factors into account, the student variables reflecting SES, Asian ethnicity, positive self esteem, and emphasis on hard work over luck fell out of statistical significance. The coefficient for Black ethnicity became significantly negative. The effects of museum visits and home computers were reduced. All other effects previously noted remained significant in the overall model.

The overall model for the reasoning factor was consistent with the partial models. The exception, previously noted, was that taking Algebra I in the 9th or 10th grade became a statistically significant negative predictor. Further analysis revealed that this result arose from a quadratic pattern in the distribution of Algebra I course taking along the two math achievement scales; the likelihood of having taken that course in 9th or 10th grade was considerably higher in the middle achievement region in comparison to the upper and lower regions. However, since lower MR scores in particular were more associated with Algebra I than higher MR scores, on average, negative relation appeared for MR, not for MK.

It is also worth noting that when comparing the magnitude of the coefficients between the overall and partial models, almost no changes were observed for MR. On the other hand, overall model coefficients for MK were generally much smaller when compared with the coefficients from the partial models.

Table 8 also shows that using the total math IRT score as a criterion often seems to average the effects found for the two math factors studied separately, and important effects are thus missed. The gender effect was significant on reasoning, nonsignificant but opposite in sign on knowledge. Yet the total score analysis taken alone would have dismissed gender differences as unimportant in general. Also, the Black ethnicity effect was much stronger on reasoning than on knowledge. On the other hand, the knowledge factor showed stronger effects for student math attitude, and for all significant course and instructional treatment variables. Student report of teacher emphasis on understanding was important only for the knowledge score, not for reasoning, yet the total score analysis would support a general conclusion. Track showed no relation to reasoning, whereas home computer availability was associated more strongly with reasoning. These differences demonstrate our point that total score analyses may misconstrue some effects and miss some effects entirely; psychological construct interpretations and policy considerations may both be helped by differentiating total scores into psychologically and educationally meaningful subscores.

Discussion and Conclusions

The analyses reported herein examine only a few of the many important questions that can be addressed with the NELS:88 data. And the results in hand so far must be considered provisional. Our plans for further analyses include studies of a variety of student and instructional treatment interaction hypotheses and much more detailed analysis of gender, ethnic, and other personal effects. Nonetheless, we believe our research to date supports several important conclusions and implications. These provide guidelines for further NELS:88 data analyses. They also suggest a new approach to future survey research on educational achievement, with particular reference to readiness and opportunity to learn, and the evaluation of educational programs.

A first conclusion is that the NELS:88 mathematics test is multidimensional and should be treated as such. At least two dimensions, representing separate scores for knowledge and reasoning, can be distinguished at both 8th and 10th grade levels. These two dimensions differ psychologically and thus statistically in their relations to other student, instructional, and school characteristics. They should therefore be distinguished in research seeking to improve our understanding of student readiness to learn, of teacher, course, and program effects on opportunity to learn, and of effective instructional design in general.

A second set of conclusions derives from this distinction. Math knowledge and reasoning, which can be distinguished in the laboratory, can also be distinguished in survey-level multiple choice tests. Of course, the labels "knowledge" versus "reasoning" offer only a simplistic shorthand. Both are complex constructs intimately related in cognitive learning and performance. But it is reasonable to think of these two aspects of mathematical achievement as having different weight in influencing student performance on particular tasks. Some tasks emphasize the application of concepts and computational skill. Although reasoning may be involved in deciding what is applicable and what to do when, differences in student success or failure on an item arise mainly from the presence or absence of the relevant declarative and procedural knowledge. Other tasks emphasize perceptive analysis and sequencing of steps to find solutions to problems embedded in scenes. Knowledge of concepts and computation is involved, but here student success or failure arises more from the ability to decontextualize the key mathematical aspects of a problem and interrelate them in a system. The contrast in the NELS:88 math test may be akin to the more general distinction between crystallized and fluid intelligence (see, e.g., Carroll, 1993; Snow, 1982). It is known that many other math achievement tests display these two aspects of ability. Although growth in crystallized knowledge and skill and in fluid analytic reasoning in mathematics are both promoted by educational experiences, it is to be expected that they will be differentially affected by specific instructional practices and learning opportunities.

Our results suggest some of these differentials. Average differences favoring males occur on reasoning but not on knowledge. Average differences favoring Asian-Americans occur on knowledge, but not on reasoning, while average differences reflecting African-American disadvantage are more pronounced on reasoning than on knowledge. Research aiming to understand such differences will be aided by knowing where to look more specifically, and thus perhaps what to look at in the spectrum of readiness and opportunity to learn variables. Student affective variables such as positive attitude toward math seem more relevant to knowledge than to reasoning. Instructional, course, and program variables also show more pronounced relation to knowledge than to reasoning, as does time spent on homework. On the other hand, home computing and the advantages of SES in general seem more strongly related to reasoning. All these patterns seem consistent with the hypothesis that crystallized knowledge growth is more influenced by formal educational factors and by personal factors such as attitude, homework, and attendance, whereas growth in fluid reasoning ability is more a function of informal learning experience promoted by home and family advantages, as well as school advantages, over the long haul.

With respect to opportunity-to-learn objectives, it is clear that taking Algebra early, and taking advanced courses by 10th grade, are positive factors in promoting both knowledge and reasoning development. Several instructional factors, such as emphasis on traditional instruction and homework, also promote knowledge growth specifically. But beyond course taking patterns, only teacher emphasis on higher order thinking and parental provision of home computers for educational use seem associated with both knowledge and reasoning development in mathematics.

The notion of opportunity to learn should consider the cognitive dimensions of student learning and differences in the instructional environments conducive to developing different kinds of cognitive aptitudes. Indeed, the thrust of current reforms in math education aim precisely to enhance U.S. students' math reasoning aptitude, which our analysis suggests is not strongly related to in-school learning opportunities at present. Large-scale assessments of student learning and educational progress should certainly aim to represent the cognitive and educational distinctions being made by cognitive psychologists, math educators, and by the nation's education goals.

A final conclusion is a reemphasis and recommendation that further research aimed at theory or policy not use total math IRT scores as a lone criterion. We believe our analysis shows that mathematics achievement is multidimensional and that much is to be gained by recognizing this fact. Research that relies on total score criteria misrepresents some important issues and misses others altogether.

References

Alexander, K. L. & Pallas, A. M. (1983). Private schools and public policy: New evidence on cognitive achievement in public and private schools. Sociology of Education, 56, 170-181.

Bock, D., Gibbons, R., & Muraki, E (1988). Full-information item factor analysis. Applied Psychological Measurement, 12, 261-280.

Cain, G. C. & Goldberger, A. S. (1983). Public and private schools revisited. Sociology of Education, 56, 208-218.

Carroll, J. B. (1993) Human Cognitive Abilities New York: Cambridge University Press.

Chubb, J. E. & Moe, T. M. (1990) Politics, markets, and America's schools. Washington, D.C.: The Brookings Institution.

Coleman, J. S. et.al. (1966) Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office.

Coleman, J. S., & Hoffer, T. (1987). Public and private high schools: The impact of communities. New York: Basic Books.

Coleman, J. S., Hoffer, T., & Kilgore, S. D. (1982). High school achievement: Public, private, and Catholic schools compared. New York: Basic Books.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), Intelligence: Measurement, theory, and public policy (pp. 147-171). Urbana and Chicago: University of Illinois Press.

Ennis, M., Kerkhoven, J.I.M., & Snow, R.E. (1993) Enhancing the validity and usefulness of large-scale educational assessments. Report P 93-151, Center for Research on the Context of Secondary School Teaching, Stanford University, Stanford, CA, January.

Frederiksen, N., Glaser, R., Lesgold, A., & Shafto, M. (Eds.) (1990). Diagnostic monitoring of skill and knowledge acquisition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Frederiksen, N., Mislevy, R., & Bejar, I. (Eds.) (1993). Test theory for a new generation of tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gifford, B.R. (Ed.)(1989a). Test policy and test performance: Education, language, and culture. Boston: Kluwer.

Gifford, B.R. (Ed.)(1989b). Test policy and the politics of opportunity allocation: The workplace and the law. Boston: Kluwer.

Glaser, R. & Linn, R. (1992). Assessing student achievement in the states. Stanford, CA: National Academy of Education.

Glaser, R. & Linn, R. (1993) The trial state assessment: Prospects and realities. Stanford, CA: National Academy of Education.

Haertel, E. H. (1988). A technical review of the National Assessment of Educational Progress: Introduction and overview. International Journal of Educational Research, 12, 673-677.

Haertel, E. H., James, T., & Levin, H. M. (Eds.) (1987). Comparing public and private schools, Vol. 2: School achievement. New York: Falmer Press.

Horn, L., Hafner, A., & Owings, J. (1992). A profile of American eighth-grade mathematics and science instruction. National Center for Educational Statistics NCES 92-486. Washington, D.C.: U.S. Department of Education.

Ingels, S.J., Scott, L.A., Lindmark, J.T., Frankel, M.R., Myers, S.L., & Wu, S. (1992). NELS:88 First follow-up drop-out component data file user's guide. National Center for Educational Statistics NCES 92-083. Washington, DC: U.S. Department of Education.

Jabine, T.B., Straf, M.L., Tanur, J.M., & Tourangeau, R. (1984). Cognitive aspects of survey methodology: Building a bridge between disciplines. Wash, DC: National Academy Press.

Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heynes, B., & Michelson, S. (1972). Inequality. New York: Basic Books.

Kupermintz, H. & Snow, R.E. (in preparation). Methodological issues in research on longitudinal educational achievement: The example of NELS:88. Center for Research on the Context of Teaching, Stanford University.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan.

Raudenbush, S.W., Rowan, B., & Cheong, Y.F. (1993) Higher-order instructional goals in secondary schools: Class, teacher, and school influences. American Educational Research Journal, 30, 523-553.

Rock, D.A., Pollack, J.M., Owings, J., & Hafner, A. (1990). Psychometric report for the NELS:88 base year test battery. National Center for Educational Statistics NCES 91-467. Washington, D.C.: U.S. Department of Education.

Snow, R.E. (1982). Education and intelligence. In R.J. Sternberg (Ed.), Handbook of Human Intelligence. New York: Cambridge University Press.

Snow, R.E. and Ennis, M. (in press) Correlates of high mathematical ability in a national sample of eighth graders. In C.P. Benbow and D. Lubinski (Eds) From psychometrics to giftedness. Baltimore, MD: Johns Hopkins University Press.

Snow, R. E. & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 262-331). New York: Macmillan.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Talbert, J.E. & DeAngelis, K. (forthcoming) School context effects on teaching for understanding in high school mathematics. Center for Research on the Context of Teaching, Stanford University, Stanford, CA.

Tukey, J.W. (1969). Analyzing data: Sanctification or detective work. American Psychologist, 24, 83-91.

Wilson, D., Wood, R., & Gibbons, R.D. (1991). TESTFACT: Test scoring, item statistics, and item factor analysis. Chicago, IL: Scientific Software Inc.

Witte, J. F. (1990). Understanding high school achievement: After a decade of resear' do we have any confident policy recommendations? Paper presented to the American Political Science Association, San Francisco.

Figure Caption

Fig 1. A schematic representation of categories of variables in the NELS:88 survey indicating the main relationships studied in the present report.

Table 1

NELS:88 Mathematics Items and Descriptions

Master Item Number	8th Grade	10th Grade Form L	10th Grade Form M	10th Grade Form H	Description
M01	1	1	1		Compare 2 algebraic expressions, given values of variables
M02	2		2	1	Compare 2 numbers read from a graph
M03	3		3	2	Read two numbers from a graph and perform an operation with them
M04	4	2	4	3	Compare two algebraic expressions, given a relationship
M05	5	3	5	4	Perform an arithmetic operation and compare result with a number
M06	6		6	5	Determine coordinates of points on a graph, perform an operation
M07	7		7	6	Compare two algebraic expressions
M08	8		8	7	Perform an arithmetic operation, compare result with a number
M09	9	4	9	8	Perform an arithmetic operation, compare result with a number
M10	10	5	10	9	Compare statements about locations on two number lines
M11	11	10	11	10	Compare length of line segments illustrated in a diagram
M12	12	11	12	11	Compare expressions involving multiplication and division of integers
M13	13	12	13	12	Compare an integer with an expression using division of decimals
M14	14	13	14	13	Compare expressions, given information containing exponents
M15	15	14		14	Compare expressions, requiring solution of simple equations
M16	16	15	15		Compare two quantities of money expressed differently
M17	17	16	16		Compare two simple arithmetic expressions involving division
M18	18	17	17	15	Compare two simple arithmetic expressions involving division
M19	19	18	18		Compare two simple arithmetic expressions involving multiplication
M20	20	21	19		Set up a simple equation that is the solution of a word problem
M21	21	22		16	Estimate a probability that is the solution of a word problem
M22	22	23			Determine the greatest of 4 decimal numbers
M23	23	24	20		Determine the smallest of 4 fractions in a word problem
M24	24	25	21	17	Choose verbal description of a problem that doesn't match diagram
M25	25	26			Determine the length of a line segment in a diagram
M26	26	27	22	18	Evaluate a relationship given statements about the variables
M27	27	28			Find an algebraic expression odd or even given fact about variables
M28	28	29	23	19	Solve a word problem requiring logical inference
M29	29	30	24	20	Solve a word problem whose answer is an algebraic expression
M30	30	31	25	21	Solve a word problem using multiplication or factoring
M31	31	32	26		Choose which decimal number is between two other numbers
M32	32	33			Choose points on a number line that include a specified decimal
M33	33		27	22	Estimate a number using a percentage indicated in a diagram
M34	34	34	28	23	Solve a simple algebraic equation
M35	35	35	29	24	Evaluate statements inferred from a word problem with a fraction
M36	36	36	30	25	Choose which expression is different from a specified percentage
M37	37		31	26	Solve a word problem requiring logical inference
M38	38	37	32	27	Evaluate statements referring to area and diagonal of a diagram
M39	39	38	33	28	Supply number that completes an algebraic equation correctly
M40	40	39	34	29	Simplify an algebraic expression
M41		6			Perform an arithmetic operation, compare result with number
M42		7			Compare two numbers containing exponents
M43		8			Compare two numbers involving multiplication and division of fractions
M44		9			Compare two expressions involving addition and multiplication of integers
M45		19			Compare two expressions involving addition and subtraction of a variable
M46		20			Perform an arithmetic operation involving decimals, compare with number
M47		40			Identify parallel line segments
M48			35	34	Determine distance between points in a diagram
M49			36	36	Solve a long division problem
M50			37	37	Determine length of side of figure given area
M51			38	38	Determine least odd integer from set of expressions
M52			39	39	Solve an algebraic inequality
M53			40	40	Determine which of a set of expressions represents a positive number
M54				30	Compute a factorial
M55				31	Solve a word problem involving area and dimensions
M56				32	Determine highest score given lowest score, mean, and range
M57				33	Solve an equation involving function notation
M58				35	Solve an equation involving function notation and exponents

Source: Rock, Pollack, Owings, & Hafner (1990)

Table 2

Proposed Math Subscales and Interpretations for Preliminary 8th Grade and Revised 8th and 10th Grade Analyses

Components from Preliminary 8th Grade Analysis

MC1: Advanced Knowledge - Computation (items 5, 8, 9, 11, 12, 13, 14, 18, 34, 39, 40)
Computational items that require knowledge of roots, exponents, negative numbers, algebra, multiplication or division of decimals.

MC2: Inferential Reasoning (items 21, 22, 23, 25, 27, 31, 35, 37)
"If--then" items that require examinee to draw conclusions about possible outcomes, given a particular scenario, and probability items; these items do not require (or invite) computation.

MC3: Basic Facts - computation (items 15, 16, 17, 19, 20)
Items that require basic mathematics knowledge, with answers readily computed by adding, subtracting, multiplying, or dividing whole numbers.

MC4: Counterexample Reasoning (items 4, 7, 10)
Items that invite examinees to devise their own concrete examples (a typical item involves the comparison of two unspecified real numbers) to eliminate alternatives.

MC5: Multiple Steps with Figures (items 3, 33, 38)
Items that require interpretation of graphical or figural information as well as several computational steps.

Factors from Revised 8th and 10th Grade Analysis

MR: Mathematical Inferential Reasoning
Items requiring inferential reasoning as in MC2 above, usually involving a scenario not just math expressions. Minimal computation needed. Multiple steps required. Mathematical knowledge necessary but not sufficient.

MK: Mathematical Knowledge and Computation
Items requiring computation and knowledge as in MC1 and MC3 above, usually involving solution of a mathematical expression in one or two steps. Minimal inferential reasoning required.

Table 3

Factor Loadings From Full Information Factor Analysis of NELS:88 Math Items in 8th and 10th Grades, After Promax Rotation (N=4059)

	8th Grade			10th Grade		
Factor	MR	MK	X	MR	MK	X
M54				99*	-32	03
M55				94*	-27	15
M25	69*	-09	08	81*	04	-14
M37	57*	-06	18	81*	01	-06
M22	92*	-12	-07	80*	03	-03
M56				78*	01	-04
M50				75*	11	08
M38	16	25*	10	72*	12	-06
M48				68*	04	01
M31	70*	02	03	66*	-00	14
M36	39*	10	29	59*	21	-03
M21	58*	02	-05	57*	20	-13
M52				56*	19	01
M35	58*	05	-05	50*	09	09
M27	48*	14	23	48*	34	11
M28	40*	05	17	48*	12	13
M10	17	-10	71*	45*	36	-01
M26	36*	27	16	45*	31	-00
M51				44*	38	04
M53				43*	22	08
M33	36*	00	06	41*	15	02
M30	26*	16	14	39*	11	14
M06	18	26	32*	37*	35	09
M24	38*	20	-01	36*	16	11
M29	34*	17	22	36*	26	09
M47				36*	-02	22
M02	40*	06	20	34*	31	-04
M49				29*	20	26
M32	48*	07	08	32*	04	09
M45				30	85*	36
M42				11	72*	08
M15	-06	84*	-11	11	70*	-03
M44				05	70*	10
M41				22	68*	16
M05	-05	52*	34	11	67*	12
M19	08	63*	-09	09	62*	23
M13	26	36*	14	13	61*	00
M12	12	38*	22	12	61*	03
M18	36*	15	30	38	60*	-15
M14	28	39*	11	26	59*	-03
M39	20	37*	35	34	57*	-02
M11	30*	29	13	28	57*	-10
M09	14	43*	28	24	53*	00
M01	10	31	38*	00	53*	29
M07	11	-05	75*	36	52*	-12
M34	11	51*	19	14	52*	15
M43				08	49*	41
M40	-18	60*	40	15	46*	20
M04	01	06	73*	42	44*	-01
M03	23	36*	09	27	43*	02
M08	36*	28	13	19	42*	10
M46				05	42*	40
M17	31*	17	01	14	26*	08
M57				14	42	72*
M58				01	39	53*
M20	39*	26	-18	21	04	45*
M23	54*	-02	-09	28	-10	33*
M16	16	49*	-17	07	20	28*

Notes.
Decimal points omitted
* indicates highest loading for each item

Table 4

NELS:88 8th Grade Math Items with Process, Content, Subscale, and Proficiency Level Designations

	Content				
Process	Arithmetic	Algebra	Geometry	Data / Probability	Advanced topics
Skills / knowledge	MK 5 (2)	X 1	MR 25	MK 3	X 6
	MR 8	MK 15			
	MK 9	MK 34			
	MK 12	MK 40 (3)			
	MK 13 (2)				
	MK 16 (1)				
	MR 17 (1)				
	MR 18 (2)				
	MK 19 (1)				
	MR 22				
Understanding / comprehension	X 10	X 4	MR 11 (3)	MR 2	
	MR 20 (1)	X 7	MR 37	MR 21	
	MR 31	MK 14 (2)	MK 38	MR 24	
	MR 32	MR 26			
	MR 33	MR 27			
	MR 36 (3)	MR 29			
		MK 39 (3)			
Problem solving	MR 23				MR 35
	MR 28				
	MR 30				

Table 5

Instructional Factors with Corresponding Items from the NELS:88 10th Grade Mathematics Teacher Questionnaire

Factor and Item Identifier[a]	Item Description
TRADITIONAL INSTRUCTION	
F1T2_18D	Use of oral question response
F1T2_18A	Use of lecture
F1T2_12A	Use of textbooks
F1T2_16A	Time spent instructing whole class
ADMINISTRATIVE TASKS	
F1T2_16F	Time spent on administrative tasks
F1T2_16D	Time spent maintaining order
F1T2_16E	Time spent administering test/quizzes
F1T2_16B	Time spent instructing small groups
DISCUSSION	
F1T2_18E	Use of student-led discussions
F1T2_18C	Use of whole-group discussion
F1T2_18H	Use of oral reports
INDIVIDUALIZATION	
F1T2_18G	Use of written assignments
F1T2_18F	Use of working in small groups
F1T2_16C	Time spent instructing individuals
MATERIALS/AUDIO-VISUALS	
F1T2_18B	Use of film
F1T2_12C	Use of audio-visual materials
F1T2_12B	Use of other reading materials
F1T2_16G	Time spent conducting lab periods
TEACHER CONTROL	
F1T2_17C	Control over teaching techniques
F1T2_17E	Control over amount of homework
F1T2_17D	Control over disciplining
F1T2_17B	Control over content taught
F1T2_17A	Control over texts/materials
EMPHASIS ON MATH APPLICATIONS	
F1T2M19F	Emphasis on importance of math
F1T2M19K	Emphasis on math in business
F1T2M19I	Emphasis on math in science
F1T2M19D	Emphasis on interest in math
F1T2M19L	Emphasis on q's about math
EMPHASIS ON HIGHER ORDER THINKING	
F1T2M19J	Emphasis on math concepts
F1T2M19A	Emphasis on logical structure
F1T2M19B	Emphasis on nature of proof
F1T2M19G	Emphasis on problem solution
EMPHASIS ON KNOWLEDGE/COMPUTATION	
F1T2M19C	Emphasis on memorizing facts
F1T2M19H	Emphasis on speedy computation

Note.

[a] Item identifiers are variable names on NELS:88 public data release

Table 6

Regression Results for 10th Grade Mathematical Knowledge (N=5460)

VARIABLE	STUDENT	COURSE/PRG	INSTRUCT	OUTSIDE
INTERCEPT	42	15	39	35
PRIOR ACHIEVEMENT				
READING 8	16 *	15 *	16 *	18 *
MATH KNOWLEDGE 8	41 *	37 *	40 *	43 *
MATH REASONING 8	34 *	31 *	35 *	36 *
STUDENT				
ABSENT	-09 *			
GENDER	-05			
SES	07 *			
ASIAN	11 *			
BLACK	-02			
HISPANIC	04			
MATH ANX	-01			
LUCK 8	-04 *			
NEG SELF 8	-03 *			
POS SELF 8	-01			
MATH ATT 8	06 *			
COURSE/PROGRAM				
ADVANCED TRACK		06		
GENERAL TRACK		-08 *		
GIFTED PRG		05		
GENERAL MATH TAKEN		-34 *		
ALGEBRA I TAKEN		18 *		
ALGEBRA II TAKEN		17 *		
GEOMETRY TAKEN		19 *		
ALGEBRA IN 8TH		09 *		
INSTRUCTIONAL				
TRADITIONAL INSTRUCT			06 *	
ADMINISTRATIVE TASKS			-03	
DISCUSSION			00	
INDIVIDUALIZATION			-07 *	
MATERIALS/AV			-01	
TEACHER CONTROL			-02	
EMPH. MATH APPLICATIONS			-05 *	
HIGHER-ORDER THINKING			13 *	
KNOWLEDGE/COMPUTATION			-02	
UNDERSTANDING			05 *	
HOMEWORK DISCUSS			02	
TIME ON HOMEWORK			05 *	
OUTSIDE				
SCIENCE MUSEUMS				08 *
HOURS TV				-00
COMPUTER IN HOME				08 *
COMPUTER CLASS				02
HELP WITH HOMEWORK				-03 *
R-SQUARED	63	66	64	62

Note
Decimal points omitted
*p<.01

Table 7

Regression Results for 10th Grade Mathematical Reasoning (N=5460)

VARIABLE	STUDENT	COURSE/PRG	INSTRUCT	OUTSIDE
INTERCEPT	42	38	45	40
PRIOR ACHIEVEMENT				
READING 8	16 *	15 *	15 *	16 *
MATH KNOWLEDGE 8	28 *	24 *	27 *	29 *
MATH REASONING 8	47 *	50 *	51 *	51 *
STUDENT				
ABSENT	-05 *			
GENDER	13 *			
SES	08 *			
ASIAN	03			
BLACK	-20 *			
HISPANIC	-04			
MATH ANX	-01			
LUCK 8	-01			
NEG SELF 8	-03			
POS SELF 8	-00			
MATH ATT 8	02			
COURSE/PROGRAM				
ADVANCED TRACK		02		
GENERAL TRACK		00		
GIFTED PRG		06		
GENERAL MATH TAKEN		-15 *		
ALGEBRA I TAKEN		-05		
ALGEBRA II TAKEN		11 *		
GEOMETRY TAKEN		12 *		
ALGEBRA IN 8TH		06 *		
INSTRUCTIONAL				
TRADITIONAL INSTRUCT			03	
ADMINISTRATIVE TASKS			-01	
DISCUSSION			00	
INDIVIDUALIZATION			-05 *	
MATERIALS/AV			-01	
TEACHER CONTROL			-01	
EMPH. MATH APPLICATIONS			-01	
HIGHER-ORDER THINKING			06 *	
KNOWLEDGE/COMPUTATION			-02	
UNDERSTANDING			01	
HOMEWORK DISCUSS			00	
TIME ON HOMEWORK			02	
OUTSIDE				
SCIENCE MUSEUMS				03
HOURS TV				-01
COMPUTER IN HOME				15 *
COMPUTER CLASS				04
HELP WITH HOMEWORK				-03 *
R-SQUARED	66	65	65	65

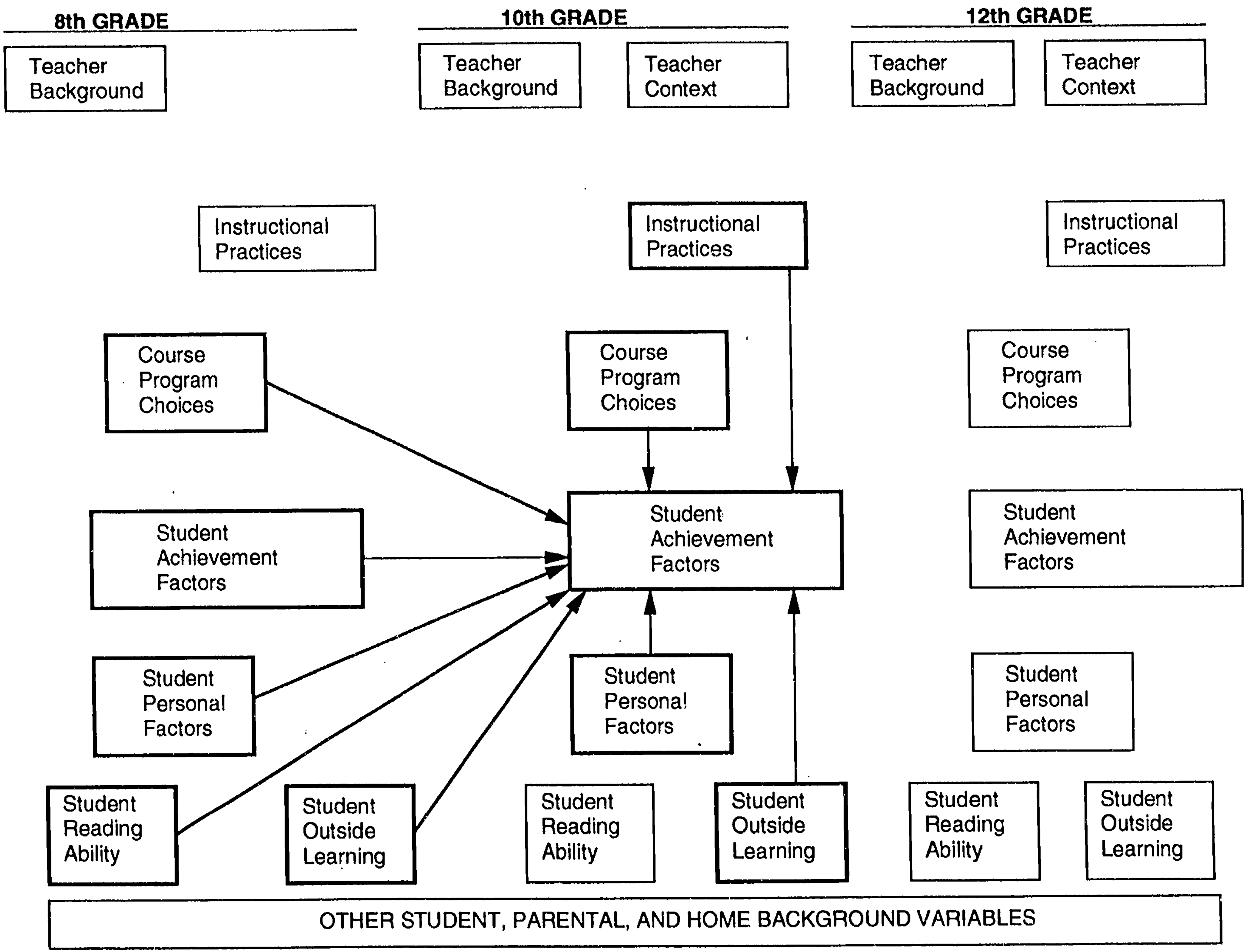
Note
Decimal points omitted
*p<.01

Table 8

Regression Results for Overall Models for 10th Grade Mathematical Knowledge, Mathematical Reasoning, and Total Math IRT Score (N=5460)

VARIABLE	MK	MR	TOTAL IRT
INTERCEPT	19	33	12
PRIOR ACHIEVEMENT			
READING 8	13 *	14 *	13 *
MATH KNOWLEDGE 8	33 *	21 *	24 *
MATH REASONING 8	29 *	44 *	33 *
STUDENT			
ABSENT	-06 *	-04 *	-05 *
GENDER	-03	12 *	01
SES	03	06 *	03 *
ASIAN	06	-00	00
BLACK	-08 *	-24 *	-14 *
HISPANIC	01	-06	-01
LUCK 8	-02	-01	-01
NEG SELF 8	-01	-01	-01
MATH ATT 8	04 *	01	03 *
COURSE/PROGRAM			
ADVANCED TRACK	04	02	02
GENERAL TRACK	-06 *	02	-04 *
GENERAL MATH TAKEN	-29 *	-14 *	-22 *
ALGEBRA I TAKEN	12 *	-07 *	06 *
ALGEBRA II TAKEN	17 *	11 *	12 *
GEOMETRY TAKEN	10 *	06 *	08 *
ALGEBRA IN 8TH	08 *	07 *	08 *
INSTRUCTIONAL			
TRADITIONAL INSTRUCT	04 *	02	03 *
INDIVIDUALIZATION	-05 *	-04 *	-05 *
EMPH. MATH APPLICATIONS	-03 *	-02	-03 *
HIGHER-ORDER THINKING	08 *	05 *	06 *
KNOWLEDGE/COMPUTATION	-02	-01	-01
UNDERSTANDING	04 *	01	03 *
TIME ON HOMEWORK	03 *	01	02
OUTSIDE			
SCIENCE MUSEUMS	02	-01	00
COMPUTER IN HOME	06 *	09 *	07 *
HELP WITH HOMEWORK	-03 *	-04 *	-03 *
R-SQUARED	68	67	75

Note
Decimal points omitted
*p<.01


8th GRADE
10th GRADE
12th GRADE
Teacher Background
Teacher Background
Teacher Context
Teacher Background
Teacher Context
Instructional Practices
Instructional Practices
Instructional Practices
Course Program Choices
Course Program Choices
Course Program Choices
Student Achievement Factors
Student Achievement Factors
Student Achievement Factors
Student Personal Factors
Student Personal Factors
Student Personal Factors
Student Reading Ability
Student Outside Learning
Student Reading Ability
Student Outside Learning
Student Reading Ability
Student Outside Learning
OTHER STUDENT, PARENTAL, AND HOME BACKGROUND VARIABLES