Sensitivity of NAEP to the Effects of Reform-Based Teaching and Learning in Middle School Mathematics

Fran Stancavage
American Institutes for Research

Lorrie Shepard
University of Colorado at Boulder

Don McLaughlin
Statistics and Strategies

Deborah Holtzman
American Institutes for Research

Charles Blankenship
American Institutes for Research

Yu Zhang
Federation of State Boards of Physical Therapy

September 2009
Commissioned by the NAEP Validity Studies (NVS) Panel

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics (NCES).

This report is based on work that was jointly supported by NCES (contract ED-01-CO-0026-005) and the National Science Foundation (grant 454755). Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of NCES or the National Science Foundation.

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Albert E. Beaton
Boston College

Peter Behuniak
University of Connecticut

George W. Bohrnstedt
American Institutes for Research

James R. Chromy
Research Triangle Institute

Phil Daro
University of California, Berkeley

Lizanne DeStefano
University of Illinois

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Robert Linn
University of Colorado at Boulder

Donald M. McLaughlin
Statistics and Strategies

Ina V.S. Mullis
Boston College

Jeffrey Nellhaus
Massachusetts State Department of Education

P. David Pearson
University of California, Berkeley

Lorrie Shepard
University of Colorado at Boulder

David Thissen
University of North Carolina, Chapel Hill

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Janis Brown
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
1070 Arastradero Road, Suite 200
Palo Alto, CA 94304-1334
Phone: 650/ 843-8192
Fax: 650/ 858-0958

Acknowledgments

The authors wish to thank the many individuals and organizations who helped to facilitate the study reported here. First and foremost, our thanks go to the schools, teachers, and students who provided the data for the study, as well as to the state CMP coordinators who helped us recruit the schools. Regrettably we cannot acknowledge these collaborators by name because we promised confidentiality to the schools and states that participated in the study.

We thank the many staff members at AIR, Westat Inc., Pearson Educational Measurement, and the Santa Clara Valley Mathematics Project who helped with field work for the study: AIR handled overall coordination and BAM materials management, Westat administered the assessments in the schools, Pearson handled NAEP materials management and scored the NAEP assessments, and the Mathematics Project scored the BAM assessments. We also thank Janis Brown and Janice Earle, our project officers on the NCES contract and NSF grant, for their patient support.

Finally, we thank the many colleagues who helped interpret the assessment items, contributed to the analyses, or provided critical review of the manuscript. Most notably, Hugh Burkhardt and Gloria Dion were invaluable in helping us to interpret the BAM and NAEP items; Phil Daro, Ed Haertel, Kilchan Choi, and Kirk Walters all made significant contributions to the analyses; and John Mazzeo and the entire NVS panel provided the critical feedback that enabled us to bring this report to a successful conclusion.

Contents

List of Tables

List of Figures

Introduction

Large-scale assessments often face the difficulty of needing to be fair to multiple programs and educational jurisdictions—often with quite different philosophies and curricula. Historically, commercial test publishers and many state assessments solved the problem by testing the intersection of various curricula, focusing on lowest-common-denominator, basic skills. In contrast, NAEP was designed to be "comprehensive," that is, to represent the union of curricular goals both across jurisdictions and across time. It should, in principle, be able to capture and report achievement gains due to effective application of either of the major curricular approaches (traditional or reform) current in the United States today, as well as hybrids that combine aspects of both methods.

By themselves, comprehensive assessments do not necessarily solve all of the problems of curriculum fairness or instructional sensitivity. Studies of TIMSS, for example, have shown that the ranking of countries would change substantially if the definition of mathematics had focused on particular content strands, or only on multiple-choice, short-answer open response or extended open-response item types (Schmidt, Jakwerth, & McKnight, 1998). Similarly, when the definition of mathematics was further manipulated by sorting items measuring the same topic into three groups—items measuring basic understandings and knowledge, items requiring use of routine procedures, and items testing problem solving or reasoning—and then emphasizing one or another of these item groups in scaling, changes in country rank ranged from none to 20, with a mean change of 7 ranks. Schmidt et al. (1998) also noted that our ability to evaluate the effects of different curricular weights using existing assessments is limited by the range of what gets included in the test. Important differences may already have been negotiated out of the shared test frameworks.

With respect to the validity of NAEP for detecting the effects of mathematics reform, we know that logistical constraints prevent NAEP from including tasks that would tap the full reach of experiences for students in classes with the most challenging curricula. It would be too disruptive in administrative settings to include tasks that the great majority of students could not do, and it would not be feasible to include mathematical investigations conducted over a period of days or in collaboration with other students. Therefore, there is a plausible concern that NAEP might underreport achievement for students in reform-based classrooms. A competing view might be that NAEP provides a wide range of both procedural and problem-solving tasks, thus enabling more accurate assessment of the skills of low-performing students than would be obtained from assessments that focus only on open-ended performance tasks.

In *Knowing What Students Know* (KWSK) (Pellegrino, Chudowsky, & Glaser, 2001), a National Research Council (NRC) committee composed of experts in learning theory and measurement modeling emphasized even more strongly the importance of aligning assessments and proficiency goals. They argued that large-scale assessments can contribute to teaching and learning, but only if they are designed to "focus on the most critical and central aspects of learning in a domain as identified by curriculum standards and informed by cognitive research and theory" (p. 241). In the judgment of the KWSK committee, the constraints under which current large-scale assessments have been built have limited their ability to measure important cognitive competencies. Pellegrino et al. (2001) provide a succinct summary of the problem motivating the present study:

> Most large-scale test developers opt for having many tasks that can be responded to quickly and that sample broadly. This approach limits the sorts of competencies that

> can be assessed, and such measures tend to cover only superficially the kinds of knowledge and skills students are supposed to be learning. Thus there is a need for testing situations that enable the collection of more extensive evidence of student performance. (p. 243)

Goals and Objectives

This study is a validity study of the National Assessment of Educational Progress (NAEP), intended to test the adequacy of NAEP for detecting and monitoring the effects of mathematics education reform. In addition, it addresses the broader question:

- What are the attributes of assessments that provide the best information on various aspects of student learning?

To provide a context for assessing student learning where we could be reasonably certain of observing substantial learning gains in mathematics over the course of a school year, we selected NSF's Connected Mathematics Project (CMP). CMP is similar to other NSF-funded middle school math programs in that it places a strong emphasis on developing students' understanding of key mathematical concepts through problem solving. To support this end, students are encouraged to reason effectively and use multiple representations and approaches. Students practice skills and procedures, but such practice occurs in conjunction with efforts to promote conceptual understanding of the associated concepts. Furthermore, CMP was identified in an informal survey of mathematics education experts[1] as the most widely implemented middle school mathematics reform curriculum.

The American Association for the Advancement of Science evaluation of middle grade mathematics textbooks (2000) documented the strong similarities in core characteristics among several reform oriented curricula, including CMP. These similarities suggest that any findings from a CMP-based study would likely generalize to students enrolled in other NSF or reform curricula.

The mathematics education experts mentioned above also helped us to identify the NSF-sponsored *Balanced Assessment in Mathematics* (BAM) as a reform-based assessment tool that reflects learning goals laid out by the National Council of Teachers of Mathematics (NCTM). Ridgway, Zawojewski, and Hoover (2000) provide a rationale for the appropriateness of BAM as an outcome measure for CMP. It is important to note, as they did, that the developers of CMP did not develop their own "Connected Mathematics Test" because they wanted students' mathematical proficiencies to transfer to a broader range of tasks.

The current study design was intended to support a comparison of the relative effectiveness of three different types of large-scale assessments—BAM, NAEP, and state assessments—for measuring the learning gains of students participating in a well-implemented reform mathematics curriculum. The following research questions were addressed:

1. For students exposed to a well-implemented reform curriculum, how great are the one-year gains that can be measured on an instrument well aligned with reform-

[1] We are grateful to Alan Schoenfeld, Norman Webb, Iris Weiss, Phil Daro, Patricia Kenney, William Schmidt, Mattye Pollard-Cole, and Lew Romagnano for their advice regarding mathematics reform curricula and assessments.

curriculum objectives (but not so particularistic as to simply comprise an end-of-course test for a specific syllabus)?

2. For students exposed to a well-implemented reform curriculum, how great are the one-year gains captured by NAEP, an instrument that has been designed more broadly to represent the union of curricular goals both across jurisdictions and across time?
3. For students exposed to a well-implemented reform curriculum, how great are the one-year gains captured by a sampling of vertically equated state assessments selected to represent both traditional multiple-choice and standards-based assessments?[2]

Based on the theoretical rationale above, our hypothesis was that NAEP would be more effective than traditional multiple-choice tests, but not as effective as BAM, in capturing student learning gains in CMP classrooms. The plausibility of this hypothesis also had empirical support in the findings reported by Ridgway et al. (2000). In their study comparing gains in CMP classrooms on ITBS and BAM, the effect size gains on ITBS (measured in standard deviation units on the pretest) were only .52 to .55 for sixth, seventh, and eighth grade CMP students, but gains were .71 to .84 for these same students on BAM.

If BAM shows greater gains in CMP classrooms than are found for the same classrooms on NAEP, we want to know if this is because BAM measures more or less than NAEP. Two competing hypotheses can be represented by the following Venn diagrams, A and B.

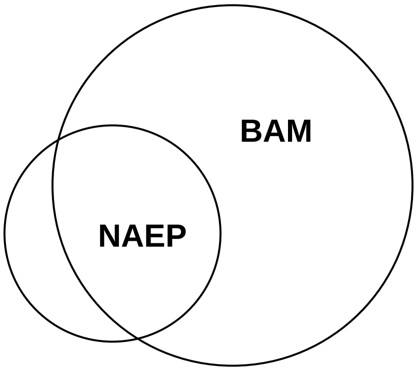


Figure A

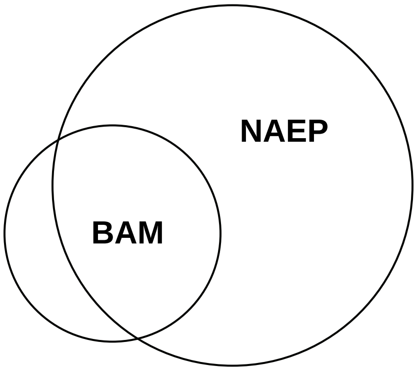


Figure B

The underlying theory of mathematics reform holds that students taught with richer conceptual approaches will develop greater procedural skills as well as conceptual knowledge and problem solving abilities. Figure A shows NAEP nested within the larger BAM domain, illustrating the claim that BAM goes further in representing the cognitive processes involved in applying mathematical knowledge. According to this model, students who are adept with BAM tasks would also score well on NAEP. Figure B represents the rival hypothesis that BAM captures only a subset of the larger NAEP domain. According to this model, students who do well on BAM would not necessarily do well on NAEP. In fact, in the latter case, it should be possible to describe the particular content or skill areas not covered by BAM where students would perform relatively more poorly.

[2] The state assessment analyses, however, were ultimately limited to a single state due to a variety of problems with the state data.

Note that both of these models are consistent with finding greater gains on BAM (or on NAEP). To test these competing hypotheses, one would examine the types of items on each assessment on which students showed the greatest gains. Evidence from these analyses would help to rule out the hypothesis that CMP students do relatively worse on one of the assessments because they are failing to master important content or cognitive skills that are included on that assessment but not on the other.

Ceiling and floor effects on one or both assessments could also be mistaken for differences in instructional sensitivity. A separate issue has to do with whether the classrooms selected reflect a sufficiently strong implementation of the CMP curriculum to provide a valid test of the hypotheses. These competing explanations for effects are addressed by the following questions:

1. Are differences between assessments the same for high achieving eighth students who have been tracked into algebra? Are they the same for seventh grade CMP students?
2. Within each of these class conditions, are differences between assessments the same for classrooms with the greatest and least gains?
3. Within each of these class conditions, are differences between assessments the same for subsamples of students who begin the year with higher or lower levels of achievement?

Methods

Study Design

The core study design called for pretesting and posttesting (September and May) of grade 8 students enrolled in CMP classrooms using both NAEP and BAM. All gains were measured in standard deviation units of the respective pretests. If both assessments were equally sensitive to learning gains in these reform-based classrooms, then gains in standard deviation units should be roughly equivalent. To ensure that differences in measured gains for each assessment were not due to either restricted range of ability for the sampled eighth graders, or to a ceiling effect on one or both of the assessments, additional classrooms were recruited. Specifically, to determine if there were ceiling effects on either of the assessments, we also administered the assessments, fall and spring, to seventh grade CMP classrooms in the same schools. And, because many middle schools implementing CMP track higher-performing eighth graders into algebra courses, we also administered the assessments, fall and spring, to grade 8 algebra classrooms in the same schools. Including only CMP eighth graders would have decreased the study's generalizability and narrowed the opportunity for substantial learning gains by restricting the range of sampled abilities. The majority of the eighth grade algebra students had participated in CMP in grades 6 and 7, since all schools selected had implemented CMP at those grade levels. Furthermore, according to members of our Mathematics Expert Advisory Board, and as noted in previous evaluations of CMP, students continue to show growth on the higher-order skills emphasized in reform-based instruction even after progressing into classrooms using more traditional curricula.

Site Selection

To ease recruitment efforts and to provide a variety of state assessments for comparison, the sample was divided across four states such that 10 schools, each including one classroom per condition, were selected from each state. To decrease data collection costs, the 10 schools were drawn from geographically clustered areas. This did not restrict the spectrum of socioeconomic conditions across schools because each state selected had a range of urban, rural, and suburban schools within close proximity of one another.

A key requirement for the validity of the study design was that the CMP curriculum be fully implemented at the study sites, and particularly in the CMP classrooms selected for assessment. Appropriate schools and teachers were identified with the help of the CMP coordinators in each state. In addition, we collected survey data to assess the extent of CMP implementation in study classrooms.[3]

We selected only schools that used CMP in grades 6 and 7 as well as grade 8. This was done to ensure that the eighth grade students had prior experience with extended constructed-response type problems and therefore wouldn't show large gains due solely to unusually low pretest scores. In the four sampled states, anywhere from 15 to 50 percent of eighth graders were in classes that either blended algebra with CMP content or offered only traditional algebra curricula.

Although our design called for 10 schools per state, with three classrooms per school corresponding to each of the three conditions, some of the schools sampled did not have eighth grade algebra classes. There were also sample losses due to one whole school in State 3, and one seventh grade classroom in a school in State 2, that refused to participate at the last minute. In consequence, the total number of classrooms included in the study was 108 rather than the target 120 (see table 1).

Table 1. Final Counts of Participating Classrooms, by State and Condition

State	CMP 7th	CMP 8th	Algebra 8th
1	10	10	7
2	9	10	5
3	9	9	9
4	10	10	10

Assessment Instruments

National Assessment of Educational Progress **(NAEP).** In order to accurately reflect the breadth of the NAEP assessment, all 10 item blocks that make up the 2005 NAEP eighth grade mathematics assessment were administered at both pretest and posttest. Because we chose to use the national NAEP item parameters in the analysis (to insure that we would derive NAEP results that were as close as possible to operational), we did not have to use the full NAEP balanced-incomplete-block (BIB) spiral, as we would have if we intended to calculate our own item parameters. Rather, only five booklets, each including 2 of the 10 operational NAEP blocks, were used. Each student completed one booklet,

[3] We also collected samples of student work, but the work samples are not analyzed in this paper.

comprising two 25-minute item blocks, in the fall, and another booklet, comprising two different item blocks, in the spring. For each administration, the five booklets were distributed evenly across states and classrooms.

The five NAEP booklets contained 177 individual items. Most items (about 70 percent) were in multiple-choice format; the rest were constructed-response items. About 80 percent of the items (including all of the multiple-choice items) were scored dichotomously; the others were polytomous. Appendix A provides sample NAEP items.

Balanced Assessment of Mathematics **(BAM).** Commercially available BAM assessment booklets contain five tasks, with each task requiring students to demonstrate depth of understanding in a particular content area. However, we were advised that a five-task BAM booklet would be too long to administer comfortably in a single class period. Therefore, in agreement with CTB, who sold us the usage rights and complete scoring guides for the BAM, we constructed eight of our own assessment booklets, each containing four tasks drawn from the pool of seventh, eighth, and ninth grade BAM tasks and balanced across four of the five content strands that are used by NAEP.[4] Each of our BAM booklets started with a seventh grade task and continued with either three eighth grade tasks or two eighth grade tasks and one ninth grade task.

BAM tasks are divided into conceptually related but separately scored items (usually about three or four per task); our eight booklets contained a total of 118 individual BAM items. Nearly all of the BAM items that we used were constructed-response items, and about 60 percent were scored polytomously. Appendix B provides a sample BAM task ("Fish Ponds"), containing six items.

To ensure adequate content coverage, each student took two of the four-task BAM booklets over two class periods at pretest and a different pair of BAM booklets at posttest. For each classroom, all eight BAM booklets were distributed in equal frequencies, and all eight BAM booklets were used in both fall and spring.

Assessment Administration

Data collection took place during the 2005-2006 school year. Each testing session occurred within a single mathematics class period. For both the fall 2005 and spring 2006 administration cycles, all three testing sessions (one NAEP and two BAM) occurred within the same week. To guard against biases resulting from order effects between NAEP and BAM, a counter-balanced design, with classrooms as the sampling unit, was employed. For reasons related to item security and the uniformity of administration conditions, field staff from Westat, the contractor normally responsible for administering NAEP, conducted all administrations. Westat has trained field staff located throughout the country, so we were able to find qualified test administrators who were local to the chosen study sites in most cases.

Unique student identifiers were used to ensure that individual students did not receive the same NAEP or BAM booklets twice. Despite our efforts it was discovered during data cleaning that six students, who were then excluded from analysis, had received one of the booklets more than once due to confusion over student identifiers at the time of administration.

[4] None of the available BAM tasks were judged to have the fifth NAEP content strand—measurement—as a primary focus. Measurement activities, however, were embedded in other BAM tasks.

Scoring Open Response Items

To support valid and generalizable conclusions about the two assessments, it was important that all constructed response items be scored reliably and in a manner consistent with the normal uses of each assessment. Pearson, the company contracted to score NAEP, was enlisted to score our NAEP data using their standard procedures. The study booklets were intermingled with operational NAEP booklets for scoring. In addition, trend papers, which are responses to the same items scored during previous scoring sessions, were included to maintain accurate score calibration. Also as part of the process, 25 percent of the papers were evaluated by a second scorer, allowing a check on inter-rater reliability.

For BAM, scoring was conducted by the Santa Clara Valley Mathematics Project, housed at San Jose State University and directed by David Foster of the Noyce Foundation. The scoring sessions were led by experienced Mathematics Assessment Collaborative trainers who had been leading similar scoring processes on behalf of the Santa Clara Valley Math Project's BAM assessment program for 7 years. Trend scoring is not typically done with BAM since the various booklets are not formally equated. However, scorers were calibrated during training, and scoring consistency was checked throughout the scoring process. The first 20 tasks for each scorer were double-scored and examined before the scorer was allowed to continue. A second scorer rescored 25 percent of the student papers for each scorer. If discrepancies were found, one of the session leaders would intervene to determine the accurate score. If the discrepancies indicated an inconsistency in how the papers were being scored, scoring would cease and the calibration process would start anew. Also, lead scorers rescored 12 percent of the tasks to produce audit scores. Finally, 5 percent of these audited tasks were randomly sampled and evaluated for inter-rater reliability. The results of this process indicated that 98 percent of the original scores were within one mark of the audit scores, and that the median correlation between the original scores and the audit scores was .99, with a median difference between the original score and the audit score of .01 points.

For both assessments, scoring of fall and spring booklets occurred simultaneously so that scorers could not know whether they were scoring a pretest or a posttest.

Constructing the Analysis Sample

Sample attrition. There were 1,874 students for whom we had valid, complete test scores on all four of the BAM administrations. For NAEP there were 2,008 students with complete and valid test scores on both administrations. Because the BAM required four administrations and NAEP only two, the difference between the numbers of students with complete data for each set of assessments was not unexpected. The number of students with complete, valid test scores for all six test administrations was 1,784. All analyses involving gains as measured by NAEP or BAM were based on this group of 1,784 students. Most students who missed an administration were either absent, or not enrolled in the classroom at the time of testing. After taking account of the missing data created by the fact that our sample was a combination of students who were present at two different points in the year (spring and fall), the response rates were comparable to those achieved by NAEP for the same grade level and assessment.[5] The higher percentage of absent students in the

[5] James Chromy, in a 2005 NAGB report titled "Participation Standards for 12th Grade NAEP," indicated that the combined student and school pre-substitution response rate for 2005 grade 8 mathematics was 90 percent, noting that due to NCLB policy school response rates were now nearing 100 percent in most states. The report can be found on the NAGB website at www.NAGB.org.

fall compared to the spring is attributable to the fact that testing occurred very early in the school year and some students were classified as "absent" who were later determined to not be enrolled. No effort, however, was made to reclassify these students in our files based on subsequent information.

Table 2. Testing Session Attendance

	Fall NAEP	Fall BAM 1	Fall BAM 2	Spring NAEP	Spring BAM 1	Spring BAM 2
Test Administered	2,308 (85.9%)	2,337 (87.0%)	2,312 (86.0%)	2,397 (80.3%)	2,407 (80.6%)	2,384 (79.9%)
Absent	219 (8.2%)	192 (7.1%)	217 (8.1%)	130 (4.4%)	115 (3.9%)	137 (4.6%)
Not Enrolled	40 (1.5%)	38 (1.4%)	38 (1.4%)	361 (12.1%)	362 (12.1%)	362 (12.1%)
Excluded	26 (1.0%)	25 (0.9%)	25 (0.9%)	16 (0.5%)	14 (0.5%)	14 (0.5%)
Not In Sample	74 (2.8%)	76 (2.8%)	76 (2.8%)	59 (2.0%)	61 (2.0%)	60 (2.0%)
Refused	20 (0.7%)	19 (0.7%)	19 (0.7%)	22 (0.7%)	26 (0.9%)	28 (0.9%)

NOTE: Accommodations were not provided. Students with disabilities or English language learners who required accommodations in order to participate meaningfully in the assessment were either excluded or tested (for the convenience of the school) but coded as "not in sample." Students who left the target classrooms after fall testing are coded as "not enrolled" in the spring. Students who *entered* the target classrooms after fall testing were generally neither enrolled in the sample nor tested in the spring since their scores could not in any case be used in the analysis. These students do not appear in table 2. In a few cases the schools requested that newly-entered students be tested with the rest of the class in the spring; these students are counted in table 2 as "test administered" in the spring and "not enrolled" in the fall, but necessarily do not contribute to our analysis sample.

While the booklets were distributed evenly among all sampled students, it was important to verify that the 1,784 students who were included in the analysis received each of the assessment booklets in equal proportions as well. Tables 3 and 4 show the percentages of students included in the analysis receiving each of the booklets for BAM and NAEP respectively. From these data it is apparent that non-response was spread evenly across all booklets, and that the group of students who were included in analyses had received each of the assessment booklets with equal likelihood.

Table 3. Percentage of Students Included in the Analysis who Received Each of the Eight BAM Booklets at Each BAM Administration

Booklet	Fall 1	Fall 2	Spring 1	Spring 2
1	14.0	11.3	10.7	13.2
2	11.7	12.4	12.7	12.6
3	13.5	12.3	11.8	13.6
4	11.6	13.0	13.7	11.7
5	13.7	11.1	11.3	13.2
6	11.9	14.4	12.9	11.8
7	13.0	12.1	12.7	12.2
8	10.8	13.6	14.3	11.7

Table 4. Percentage of Students Included in the Analysis who Received Each of the Five NAEP Booklets at Each NAEP Administration

Booklet	Fall	Spring
1	19.0	20.6
2	19.7	18.8
3	21.0	19.6
4	19.8	20.9
5	20.5	20.0

Sample demographics. Our sample was a convenience sample of nominated schools—all using the same (CMP) curriculum—that agreed to be in the study. It is not, therefore, intended to representative of all U.S. school children. Nevertheless, it is worth asking whether or not the students comprising our sample are similar in demographic characteristics to a representative sample of U.S. students. As can be seen in table 5, there were a number of differences between our analysis sample and the 2005 national NAEP grade 8 mathematics sample (public schools only).

Table 5. Demographic Characteristics of Analysis Sample Compared to 2005 National NAEP Grade 8 Mathematics Sample (Public Schools Only)

	Analysis Sample	NAEP Sample
White	65	60
Black	8	17
Hispanic	14	17
Asian	12	5
SD	5	11
ELL	5	6
Free/reduced lunch eligible[1]	31	39

[1]Students are classified as free/reduced lunch eligible, not eligible, or information not available. Only 3 percent of the NAEP sample was classified as "information not available" in 2005. However, 18 percent of the students in our analysis sample were so classified.

The demographics of our sample, as would be expected, also differed across conditions, with the upper-track algebra classrooms containing more Asian and White students, and fewer low-SES, Hispanic, Black, SD, and ELL students than the lower-track eighth grade CMP classrooms. Demographics for the each of the three conditions are displayed in table 6.

Table 6. Demographic characteristics of analysis sample, by condition

	Condition		
	CMP 7th	CMP 8th	Algebra 8th
White	63	57	73
Black	10	10	3
Hispanic	16	23	4
Asian	10	8	19
SD	8	7	1
ELL	5	9	1
Free/reduced lunch eligible[1]	36	42	16

[1]Students are classified as free/reduced lunch eligible, not eligible, or information not available. Information was not available for 16 percent of the students in the CMP 7th condition, 18 percent of the students in the CMP 8th condition, and 19 percent of the students in the "algebra 8th" condition.

Computing Overall Test Scores

For our analysis, we first used the national NAEP item parameters to calculate NAEP theta scores.[6] We next undertook a joint scaling of the NAEP and BAM items in which the NAEP item parameters were held constant, item parameters were estimated for the BAM items, and

[6] In doing so, however, we ignored the NAEP subscale structure and treated all of the item parameters as though they were based on a single theta scale.

BAM theta scores were computed. Both NAEP and BAM theta scores were then converted to the NAEP reporting metric.

In a test with a block design, students' performance is only partially observed. (That is, no one student answers all of the items in the pool of items that define theta.) To circumvent this problem, maximum likelihood (MML) regression was used to estimate characteristics of the NAEP and BAM score distributions for groups of students in our sample. We were then able to generate distributional draws (plausible or imputed values of theta) for each individual student and use these plausible values in our further analyses. Plausible values can be thought of as a mechanism for accounting for the fact that the true scale scores describing the underlying performance for each student are unknown.

BAM scaling. As noted in the section on assessment instruments, each student completed eight BAM tasks, and, within each task, students responded to a series of discrete but interrelated items, each of which earned separate points based on the BAM scoring rubrics. (Partial credit scoring is used for some items, while others are dichotomously scored.) These separately scored units were treated as "items" for the IRT analysis. In total, our BAM spiral included 118 "items." Possibly as a consequence of sample size limitations, some items did not show much variation in performance, and adjustments were required before the IRT model would converge. Specifically, 10 very easy or very difficult items were dropped, and two or more response categories were combined for 24 of the partial credit items. The latter adjustments had the effect of reducing distinctions at the upper end of the difficulty spectrum, since the collapsed items no longer differentiated as many levels of "correctness."

Scaling BAM together with NAEP had the advantage of allowing us to report results for both assessments in the same metric and also to examine the relationship of individual items across instruments. However, there are two potential problems with this approach. First, in normal usage BAM is not scored using IRT, but by adding up the total number of points earned. (Test forms are not equated.) Second, because of the limitations of sample size, and the fact that we did not administer the full NAEP spiral (in which every block is paired with every other block), it was not possible to do a true joint scaling in which all item parameters were estimated together. Therefore, it could be argued that scoring BAM using our joint scaling disadvantaged BAM by forcing it onto the NAEP theta scale.

In seeking to evaluate this concern, we undertook a confirmatory factor analysis in which separate NAEP and BAM factors were calculated and then correlated with one another. A high degree of correlation would be taken as evidence that the two factors were very similar and a joint scaling was appropriate. (The correlation represents the cosine of the angle between the two factors in a "factor space.") Unfortunately, due to the matrix sampling of items across students, it was not possible to combine all of the NAEP items and all of the BAM items in a single factor analysis. Rather, each analysis had to be restricted to a single NAEP booklet and a single BAM booklet. Nine different booklet pairs were examined. Across these nine cases, the average observed correlation between the two factors was .88, and the true correlation between the NAEP factor and the BAM factor was estimated to be approximately .9. We judged this degree of correlation to be high enough to support the use of the joint scaling. However, in order to feel fully confident of our findings, we also carried out a parallel analysis of student growth on NAEP and BAM using a percent correct metric. The results of this analysis are reported in appendix C.

Test reliabilities. The reliabilities of the two assessments were calculated as

(Observed score variance – error variance)/Observed score variance

We obtained the observed score variance by taking the variance of the first set of plausible values. We obtained the error variance by taking the variance across the five sets of plausible values for each student, and then averaging across students. Based on these calculations, the reliabilities of the two assessments were similar, and averaged around .9 for both assessments. Table 7 shows the reliabilities separately for the fall and spring administrations.

Table 7. Test Reliabilities

	NAEP	BAM
Fall	0.89	0.91
Spring	0.90	0.91

Analysis of State Test Scores

Although we had hoped to be able to compare students' growth on a variety of state tests to their growth on NAEP and BAM, these comparisons were ultimately restricted to a single state due to problems with the state data from the other three states. More specifically, two of the states did not have vertically equated tests and the third state was transitioning to a new test system and could only provide scale scores for the posttest year (2006). For the fourth state we did compute fall-to-fall student gains (fall of sixth grade to fall of seventh grade, or fall of seventh grade to fall of eighth grade, depending upon condition), and compared these to the fall-to-spring gains on NAEP and on BAM.

Results

Mean Scale Scores for NAEP and BAM

Table 8 shows the NAEP and BAM mean scale scores (and standard deviations) of the students in our sample, by condition, in the fall and spring. As can be seen, the eighth-grade algebra students, with average scores of 314.47 in the fall and 327.69 in the spring, performed considerably higher, on both NAEP and BAM, than the other two groups of students. The *national* average for grade 8 NAEP was 279 (s.d. = 36) in 2005. This was somewhat lower than the spring scores earned by students in any of our three conditions. However, our sample was not intended to be nationally representative, and—as was shown in table 5—was in fact skewed toward higher-performing demographic groups compared to the nation. Moreover, it should be noted that the comparison is inexact because national NAEP is administered in February and our spring testing was administered in May, very close to the end of the school year.

Table 8. Mean Scale Scores (Standard Deviations) of the Students in Our Sample[1]

Condition	N	NAEP Fall	NAEP Spring	BAM Fall	BAM Spring
7th Grade CMP	643	270.20 (33.87)	283.12 (38.32)	268.30 (36.42)	282.69 (37.95)
8th Grade CMP	535	271.10 (30.11)	282.26 (32.64)	271.71 (30.58)	282.87 (33.07)
8th Grade Algebra	606	314.47 (29.52)	327.69 (30.55)	315.97 (27.82)	324.25 (28.35)

[1] Mean booklet percent-correct scores for fall and spring are given in appendix C, table C-1.

For each of the three conditions, the average gains on NAEP are in the range of 11-13 NAEP score points. This is reasonably consistent with results on national NAEP, in which the difference between mean scores for grades 4 and 8 was 41 points in 2005.

NAEP vs. BAM effects

Although we initially hypothesized that BAM—being more closely aligned with the reform curriculum—would record larger gains than NAEP, the data suggest that both assessments are equally sensitive to the gains of the CMP students in our study, and NAEP appears better able to detect gains in the algebra classrooms (an effect size of .448 for NAEP compared to .298 for BAM). Table 9 displays the average gains in effect size units for each of the three conditions on both NAEP and BAM. Effect size units were calculated within each condition by dividing the gain by the standard deviation from the fall administration. Effect sizes using the percent-correct metric are shown in appendix C. Results in the percent-correct metric were very similar to the results reported here using IRT scaling.

Table 9. Fall-to-Spring Gains Measured by NAEP and BAM, in Fall Standard Deviation Units, with Standard Errors

Condition	N	NAEP[a]	BAM[b]	Difference
7th Grade CMP	643	0.381 (0.033)	0.395 (0.033)	-0.014 (0.043)
8th Grade CMP	535	0.371 (0.039)	0.365 (0.039)	0.006 (0.052)
8th Grade Algebra	606	0.448 (0.040)	0.298 (0.040)	0.150** (0.054)

NOTE: Significance of difference indicated by: $^{+}$ p<0.10, * p<0.05, ** p<.01, *** p<.001

[a] Across the three conditions, disattenuated effect sizes for NAEP (adjusted for fall test reliability) are .404, .393, and .475 respectively.

[b] Disattenuated effect sizes for BAM are .415, .384, and .313 respectively.

One critical element of the study design is that the sampled students must be benefiting from well-implemented reform-based mathematics instruction and thus exhibiting the types of learning gains that are intended in such a curriculum. Correspondingly, one obvious argument against the meaningfulness of our findings is the possibility that the sample failed to capture such students. In other words, the concern that NAEP is not sensitive to the types of gains fostered by reform-based mathematics curriculum would not be adequately addressed by our study if the sample was dominated by classrooms in which such gains were not present. Because of the way the sample frame was developed (working with each state's CMP coordinator, who theoretically would be motivated to recruit the most successful CMP schools), one might expect the CMP classrooms in our study to be among the best in each state. Moreover, all of the schools selected had fully implemented CMP curriculum in grades 6 through 8, and, based on a short survey which we administered at the time of testing, most students in our sample had had at least one prior year of middle school CMP instruction. As shown in table 10, 83 percent of the CMP students and 94 percent of the algebra students had had at least one prior year of CMP in middle school.

Table 10. Percentage of Students with at Least One or Two Prior Years of CMP Instruction in Middle School, for Each of the Three Conditions in Our Study

	Condition		
	CMP 7th	CMP 8th	Algebra 8th
At least one prior year of CMP	83	83	94
Two prior years of CMP	na	64	78

Teachers of the seventh and eighth grade CMP classes also provided information relevant to judging the extent (if not the quality) of CMP implementation. Seventy percent of our CMP teachers reported teaching CMP for at least 3 years, and the mean number of years teaching CMP was 4.5.

With regard to the amount of CMP materials used in the target classes, teachers were given a list of all grade 6 through 8 CMP "units" (eight units are available at each grade level) and asked to mark the extent to which they used each unit. Ninety-seven percent of the CMP teachers reported using at least some material from at least four CMP units, and 80 percent reported using at least some material from at least six CMP units (table 11).

Table 11. Percentage of CMP Teachers Reporting the Use of Materials From at Least Four or at Least Six CMP Units in Their Target Class

	At Least 4 CMP Units	At Least 6 CMP Units
Used at least some material from…	97	80
Used at least half the investigations[1] from…	85	47

[1] "Investigations" are the primary components of CMP units.

Although these facts are reassuring, given the difficulties of implementing reform curriculum by teachers who have primarily been exposed to traditional teaching practices, it was worth considering what the results would be if we restricted the analysis to the highest gaining classrooms in our sample. For this analysis, we divided the classrooms in half within each condition, based on an average of NAEP gains and BAM gains for individual students, aggregated to the classroom level.

The results are displayed in table 12. As in the main analyses, the data show that both assessments are roughly equivalent in being able to detect learning gains in high-gaining CMP classrooms, thus providing further evidence that the results of the study are not likely due to poor implementation of the CMP curriculum. Also as in the main analyses, effect sizes in the high gaining-algebra classes favor NAEP (an effect size of .743 for NAEP, compared to .510 for BAM).

Table 12. Fall-to-Spring Gains Measured by NAEP and BAM, in Fall Standard Deviation Units (with Standard Errors), for Top and Bottom Halves of the Sample, as Based on Averaged NAEP and BAM Gains at the Classroom Level

		N	NAEP	BAM	Difference
7th Grade CMP	Students in the Top-Gaining Half of Classrooms	332	0.537 (0.047)	0.480 (0.044)	0.057 (0.059)
	Students in the Bottom-Gaining Half of Classrooms	311	0.222 (0.050)	0.312 (0.052)	-0.090 (0.070)
8th Grade CMP	Students in the Top-Gaining Half of Classrooms	239	0.518 (0.055)	0.513 (0.054)	0.005 (0.073)
	Students in the Bottom-Gaining Half of Classrooms	296	0.234 (0.056)	0.227 (0.057)	0.008 (0.075)
8th Grade Algebra	Students in the Top-Gaining Half of Classrooms	285	0.743 (0.065)	0.510 (0.064)	0.233** (0.087)
	Students in the Bottom-Gaining Half of Classrooms	321	0.242 (0.052)	0.147 (0.054)	0.096 (0.073)

NOTE: Significance of difference indicated by: $^+$ p<0.10, * p<0.05, ** p<.01, *** p<.001

Finally, as shown in table 13, we looked at how well the two assessments were able to detect gains for students who started the year with higher or lower levels of achievement (based on average NAEP and BAM fall test scores). Across conditions (but particularly in algebra), the somewhat larger effect sizes for students who started the year in the bottom half of the distribution suggest that there may be some ceiling effects in both assessments. Nevertheless, the findings still follow the same pattern as seen in the main analyses and in the analyses of high-gaining classrooms. That is, there were no significant differences in the ability of NAEP and BAM to detect gains for either grade 7 or grade 8 CMP students, but NAEP registered greater gains for algebra students, irrespective of whether they started in the top or bottom half of the distribution. (However, the difference in sensitivity among algebra students who started in the top half of the distribution is weaker, and it is only significant at the 0.10 level.)

Table 13. Fall-to-Spring Gains Measured by NAEP and BAM, in Fall Standard Deviation Units (with standard errors), for Top and Bottom Halves of the Sample, as Based on Averaged NAEP and BAM Fall Scores at the Student Level

		N	NAEP	BAM	Difference
7th Grade CMP	Top Half of Students, Fall Achievement	321	0.515 (0.068)	0.475 (0.061)	0.040 (0.085)
	Bottom Half of Students, Fall Achievement	322	0.590 (0.069)	0.699 (0.076)	-0.109 (0.099)
8th Grade CMP	Top Half of Students, Fall Achievement	267	0.413 (0.082)	0.438 (0.078)	-0.024 (0.107)
	Bottom Half of Students, Fall Achievement	268	0.632 (0.077)	0.606 (0.082)	0.026 (0.108)
8th Grade Algebra	Top Half of Students, Fall Achievement	303	0.376 (0.082)	0.150 (0.086)	0.226^{+} (0.118)
	Bottom Half of Students, Fall Achievement	303	1.007 (0.087)	0.657 (0.073)	0.350** (0.109)

NOTE: Significance of difference indicated by: $^{+}$ p<0.10, * p<0.05, ** p<.01, *** p<.001

NAEP and BAM Effects by Content Area and Cognitive Complexity

Another interesting question involves looking at the ability of the two assessments to detect gains within particular mathematics topic areas and by level of cognitive demand. Items were classified into the five content areas and three levels of "mathematical complexity" used by NAEP. The 2005 NAEP Mathematics Framework describes mathematical complexity as follows:

> Mathematical complexity of an item answers the question, "What does the item ask of the students?" Each level of complexity includes aspects of knowing and doing mathematics, such as reasoning, performing procedures, understanding concepts, or solving problems. The levels are ordered, so that items at a low level would demand that students perform simple procedures, understand elementary concepts, or solve simple problems. Items at the high end would ask students to reason or communicate about sophisticated concepts, perform complex procedures, or solve nonroutine problems. (NAGB, 2004, Ch. 2, pp. 2-3)

Classification of the BAM items was done by an expert panel that included individuals who were experienced developers of both assessments. The same expert panel reviewed and validated the operational classifications of the NAEP items.

Table 14 displays the results of the gain analysis by topic area. Here we see that the greater sensitivity of NAEP for gains made by the eighth grade algebra students was particularly concentrated in items from the algebra (p=.0141), geometry (p=.0135), and measurement (p=.0263) strands. In eighth grade CMP classes, none of the content areas exhibited differential gains by test instrument, but in seventh grade CMP, the two assessments were each differentially better at detecting gains in particular content areas. Specifically, NAEP was more sensitive to gains in geometry at seventh grade (p=.0273), while BAM was more sensitive to gains in the content area of number properties and operations (p=.0427). One subtopic within number properties and operations that may contribute to this latter difference (favoring BAM) is proportional reasoning. The seventh grade CMP curriculum has two units devoted to deepening students' understanding of ratio and proportion concepts, which are deemed critically important content to be introduced and mastered at the middle school level. The relative weakness of NAEP in this area was noted in a recent validity study by an expert panel of mathematicians, who found the set of eighth grade NAEP items classified as number properties and operations to be seriously lacking in coverage of ratios and proportional reasoning (Daro, Stancavage, Ortega, DeStefano & Linn, 2007).[7]

Table 14. Fall-to-Spring Gains Measured by NAEP and BAM, in Fall Standard Deviation Units (with Standard Errors), for Items Grouped by Mathematics Topic Area

	Number Properties & Operations	Measurement	Geometry	Data Analysis & Probability	Algebra
NAEP	47 items	28 items	37 items	24 items	41 items
7th Grade CMP	0.210 (0.043)	0.198 (0.048)	0.302 (0.046)	0.161 (0.049)	0.387 (0.050)
8th Grade CMP	0.180 (0.047)	0.172 (0.056)	0.177 (0.057)	0.184 (0.054)	0.256 (0.052)
8th Grade Algebra	0.254 (0.049)	0.242 (0.052)	0.244 (0.051)	0.115 (0.053)	0.374 (0.050)
BAM	33 items	3 items	14 items	15 items	43 items
7th Grade CMP	0.332 (0.043)	0.157 (0.053)	0.151 (0.052)	0.179 (0.051)	0.320 (0.041)
8th Grade CMP	0.209 (0.052)	0.050 (0.059)	0.127 (0.056)	0.104 (0.057)	0.279 (0.050)
8th Grade Algebra	0.194 (0.047)	0.070 (0.057)	0.056 (0.057)	0.028 (0.053)	0.202 (0.049)

Significant NAEP vs. BAM differences:

7th Grade CMP: Number Properties & Operations (BAM>NAEP p=.0427); Geometry (NAEP>BAM p=.0273)
8th Grade CMP: no significant differences
8th Grade Algebra: Measurement (NAEP>BAM p=.0263); Geometry (NAEP>BAM p=.0135); Algebra (NAEP>BAM p=.0141)

Decomposing the two tests by content area also illustrates the effect of shortening the total test on the ability to measure gains. As would be expected, shorter tests are less reliable and therefore less able to detect gains from fall to spring.

The results of the gain analysis by complexity level are shown in Table 15. Since very few items on either instrument were classified as high complexity, high- and medium-complexity items were analyzed together. Here we see that the NAEP/BAM differences in detecting gains among algebra students were concentrated in the low-complexity items. That is, within the algebra condition, the effect size differences between the two assessments are

[7] This panel was examining the 2007 NAEP item pool, not the 2005 item pool, but there was substantial overlap among the two years' item pools and it is likely that the 2007 judgments also apply to 2005.

highly significant for low-complexity items (p=.0021), but not significant for high- and medium-complexity items.

Table 15. Fall-to-Spring Gains Measured by NAEP and BAM, in Fall Standard Deviation Units (with Standard Errors), for Items Grouped by Complexity

	Low Complexity	Medium/High Complexity
NAEP	112 items	65 items
7th Grade CMP	0.401 (0.038)	0.317 (0.041)
8th Grade CMP	0.272 (0.042)	0.242 (0.047)
8th Grade Algebra	0.469 (0.045)	0.286 (0.045)
BAM	39 items	69 items
7th Grade CMP	0.350 (0.043)	0.383 (0.036)
8th Grade CMP	0.211 (0.049)	0.358 (0.044)
8th Grade Algebra	0.280 (0.047)	0.249 (0.041)

Significant NAEP vs. BAM differences:
7th Grade CMP: no significant differences
8th Grade CMP: Medium/High Complexity (BAM>NAEP p=.0659)
8th Grade Algebra: Low Complexity (NAEP>BAM p=.0021)

In grade 8 CMP, by contrast, the difference in test sensitivity, while not as great, was concentrated in high- and medium-complexity items, and favored BAM over NAEP (P=.0659). No significant differences were observed in grade 7.

Offsetting or Compensatory Differences Between Tests

Findings from the content and item complexity analyses contradict somewhat the overall study findings of “no difference” between NAEP and BAM in capturing learning gains in seventh and eighth grade CMP classrooms and a general superiority of NAEP in capturing gains in eighth grade algebra classes. Instead, there appear to be some compensatory or offsetting strengths of the two tests worthy of further exploration.

To further examine the hypothesis that gains on NAEP reflect predominantly low-complexity gains while gains on BAM represent disproportionately gains on medium- and high-complexity items, regression analyses were performed predicting total test score gains from subtests constituted from the two item types. Regression coefficients and standard errors are reported in Table 16. While there appeared to be a pattern in Table 15 in the relative magnitude of effect sizes, some of the differences between NAEP and BAM were not statistically significant. Here, however, the pattern is more consistent across all of the class conditions, and strikingly different between the two tests. Total score gains on BAM are accounted for predominantly by the subtest comprising medium- and high-complexity items, which has regression coefficients twice those of the low-complexity subtest. On NAEP, the reverse is true, with total test score gains being predicted much more strongly by the low-complexity subtest. These differences are likely the result of differences in both the relative *number* of each type of item on the two tests and the relative *measurement quality* of the two types of items.

Table 16. Regression coefficients and standard errors when predicting total test score gain from gains on subtests of low- and medium/high-complexity items.

	BAM			NAEP		
	7th Grade CMP	8th Grade CMP	8th Grade Algebra	7th Grade CMP	8th Grade CMP	8th Grade Algebra
	b/se/beta	b/se/beta	b/se/beta	b/se/beta	b/se/beta	b/se/beta
BAM Low Complexity Items = 39	0.222*** (0.02) 0.308	0.182*** (0.02) 0.259	0.187*** (0.02) 0.253			
BAM Medium/High Complexity Items = 69	0.436*** (0.03) 0.489	0.424*** (0.03) 0.498	0.469*** (0.03) 0.517			
NAEP Low Complexity Items = 112				0.399*** (0.03) 0.463	0.322*** (0.03) 0.387	0.347*** (0.03) 0.415
NAEP Medium/High Complexity Items = 65				0.189*** (0.02) 0.254	0.198*** (0.03) 0.271	0.241*** (0.03) 0.323
r2	0.383	0.338	0.369	0.300	0.220	0.286
r2_adjusted	0.381	0.335	0.367	0.298	0.217	0.284
N	643	535	606	643	535	606

NOTE: Significance of difference indicated by: $^{+}$ p<0.10, * p<0.05, ** p<.01, *** p<.001

The comparisons in Table 15 between NAEP and BAM gains based on subtests of medium/high-complexity items are relatively straightforward because the two subtests have very nearly the same number of items: 65 items and 69 items, respectively. However, the NAEP low-complexity subtest has almost three times as many items as the BAM low complexity subtest: 112 items versus 39 items. Do the greater gains registered by the NAEP low-complexity subtest indicate that NAEP low-complexity items are more instructionally sensitive that BAM low-complexity items? Or are there just more of them? To allow a more direct comparison without the confounding influence of test length, the NAEP low-complexity items were randomly subdivided into three shorter subtests—37 or 38 items in length. Fall-spring effect sizes for the three shorter subtests are reported in Table 17; they can be compared with the effects for the full 112-item subtest, which are reproduced from the top left-hand quadrant of Table 15.

Table 17. Fall-to-spring gains measured by BAM and NAEP total low-complexity items and three random NAEP subtests, in fall standard deviation units

	Low Complexity			
NAEP	Total items (112 items)	Subtest A (37 items)	Subtest B (37 items)	Subtest C (38 items)
7th Grade CMP	0.401	.420	.385	.344
8th Grade CMP	0.272	.422	.251	.259
8th Grade Algebra	0.469	.515	.281	.276
BAM	Total items (39 items)			
7th Grade CMP	0.350			
8th Grade CMP	0.211			
8th Grade Algebra	0.280			

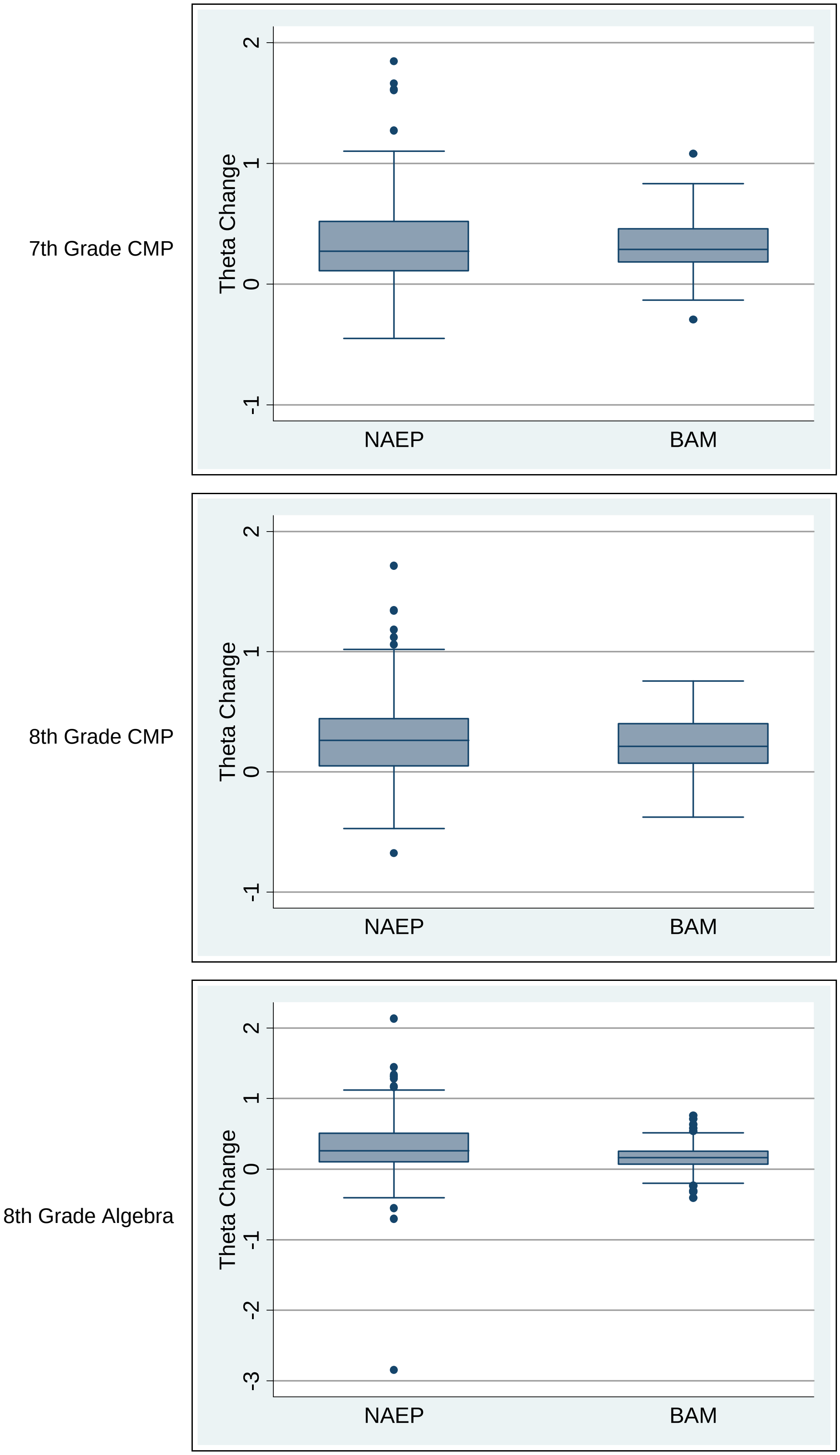
If number of items is a major determinant, shortening the NAEP subtests could reduce their reliability and therefore make it more difficult to detect fall-spring gains. Indeed, the effect sizes for subtests B and C are smaller than those for the full set of 112 NAEP low-complexity items and (with the exception of eighth grade CMP) are no longer greater than the gains for BAM low-complexity items. The results for subtest A were anomalous. Clearly this random draw of items was appreciably different. Even when the five highest gaining items were eliminated from subtest A, the effect sizes continued to be almost as great as for the full 112 items. These findings, as well as the desire to examine the substantive nature of fall-spring gains, led to a more fine-grained analysis of item gains and the distribution of item gains.

Quantitative Analysis of Item Gains

To delve further into the effects observed on each assessment and differences between them, we calculated theta gains, within condition, on each individual item, and we then examined the distributions of these theta gains. Items with fall p-values above .95 or below the c (“guessing”) parameter for a particular condition were excluded from the analysis of that condition because of the high error associated with estimates of theta for items at the extremes of the distribution.[8] Figure 1 shows the distribution of gains among all items, for each assessment, by condition.

[8] The excluded items exhibited no particular pattern by content area or complexity level. As might be expected, the very easy items were concentrated in the algebra condition, while the very difficult items were concentrated in the CMP conditions, particularly seventh grade CMP.

Figure 1. Distribution of Item Gains, NAEP and BAM



NOTE: Parts of box plots are as follows. Horizontal line in shaded box: median. Top and bottom edges of shaded box: 75th and 25th percentiles (Q3 and Q1), respectively.

The difference in the shape of the item gain distributions is most evident in the algebra condition, which is, of course, the condition in which NAEP demonstrated a significantly greater effect size than BAM. In the algebra condition, the distribution of item gains for BAM is both more compressed and centered lower than the NAEP distribution. The box-and-whisker plots in figure 1 also make it easy to observe the fact that NAEP has a few items for which there were very large theta changes from fall to spring. These items, which tended to repeat across class conditions, are discussed substantively in the next section.

Content Analysis of Item Gains

The hypothesis that NAEP and BAM appeared to produce "the same" results for seventh and eighth grade CMP classes because of offsetting, differential gains on low-complexity and medium/high-complexity items prompted a further investigation into the substantive nature of the items contributing the most to the gains on the two assessments. Table 18 lists the 10 top gaining items on NAEP and BAM in each of the three class conditions, ordered from the highest gaining to the tenth-highest gaining. These items were examined by content experts who, in addition to reviewing the content and complexity classifications, looked for other patterns, such as patterns in item format or skill requirements.

Table 18. Analysis of the Top 10 Gaining Items on NAEP and BAM by Content, Complexity, Item Format, and Type of Skill Required

NAEP					BAM				
	Content	Format	Comp.	Skill		Content	Format	Comp.	Skill
7th Grade CMP									
Identify/plot points on Cartesian plane	ALG	MC	L	ES	Solve problem using spatial visualization	GEO	CR	MH	MR
Evaluate algebraic expression	ALG	MC	L	ES	Use sampling strategy to estimate large numbers	DAT	CR	L	MR
Apply basic statistical terminology	DAT	CR	MH	ES	Apply fraction and area concepts	NUM	CR	MH	MR
Identify/plot points on Cartesian plane	ALG	MC	L	ES	Apply area concepts	GEO	CR	MH	ES
Identify particular types of number systems	ALG	MC	L	ES/IS	Describe algebraic formula in words	ALG	CR	MH	MR
Apply area concepts	GEO	CR	MH	ES	Understand properties of geometric shapes (congruence)	GEO	CR	L	ES
Identify/plot points on Cartesian plane	GEO	MC	L	ES	Use sampling strategy to estimate large numbers	DAT	CR	L	MR
Apply Pythagorean theorem on Cartesian plane	GEO	CR	MH	ES	Write formula from real world data	ALG	CR	MH	MR
Identify points on quadratic function	ALG	MC	L	ES	Understand properties of geometric shapes (congruence)	GEO	CR	L	ES
Interpret data from linear graph	ALG	CR	MH	MR	Identify and extend pattern in picture and table	ALG	CR	MH	MR
8th Grade CMP									
Apply Pythagorean theorem to find side of triangle	GEO	MC	L	ES	Write formula from real world, spatial data	ALG	CR	MH	MR
Express number using exponents	NUM	MC	L	ES/IS	Write formula from table	ALG	CR	MH	MR
Identify geometric terminology	GEO	MC	L	ES	Apply number operations to real world data	NUM	CR	MH	ES
Apply basic statistical terminology on number line	NUM	MC	L	ES	Interpret a graph involving time and distance	ALG	CR	L	MR
Transform geometric shape, know terminology	GEO	MC	MH	ES	Analyze data from scatter plot	DAT	CR	L	MR
Match linear equation to graph	ALG	MC	L	ES	Write formula from real world, spatial data	ALG	CR	MH	MR

Table 18. Analysis of the Top 10 Gaining Items on NAEP and BAM by Content, Complexity, Item Format, and Type of Skill Required (Continued)

NAEP					BAM				
	Content	Format	Comp.	Skill		Content	Format	Comp.	Skill
Use proportional reasoning to find missing side of similar triangle	GEO	CR	MH	MR	Combine geometric shapes to make new shape	GEO	CR	MH	MR
Interpret bar graph to answer real world question	DAT	CR	MH	MR	Use number operations in context	ALG	CR	MH	MR
Analyze data in stem and leaf plot	DAT	MC	L	ES	Solve problem using spatial visualization	GEO	CR	MH	MR
Apply Pythagorean theorem on Cartesian plane	GEO	CR	MH	ES	Apply number pattern to real world, spatial context	ALG	CR	MH	ES
8th Grade Algebra									
Use proportional reasoning to find missing side of similar triangle	GEO	CR	MH	ES	Analyze data from scatter plot	DAT	CR	L	MR
Interpret graph	ALG	MC	L	ES	Transform geometric shape (spatial visualization)	GEO	CR	MH	MR
Identify/plot points on Cartesian plane	ALG	MC	L	ES	Demonstrate understanding of mean	DAT	CR	L	MR
Apply basic statistical terminology	DAT	CR	MH	ES	Make linear graph from table	ALG	CR	L	ES
Interpret bar graph to answer real world question	DAT	CR	MH	MR	Plot points on Cartesian plane	GEO	CR	MH	MR
Express number using exponents	NUM	MC	L	ES/IS	Solve problem using spatial visualization	GEO	CR	MH	MR
Apply Pythagorean theorem on Cartesian plane	GEO	CR	MH	ES	Use equation to solve problem	ALG	CR	L	MR
Understand concept of fraction	NUM	MC	L	ES	Test conjectures about properties of numbers	NUM	CR	MH	MR
Express number using exponents	NUM	MC	L	ES/IS	Identify pattern in table to solve problem	ALG	CR	MH	MR
Order angles from smallest to largest	MEA	MC	L	ES	Identify points on Cartesian plane	GEO	CR	MH	MR

NOTE: Content areas are: Number properties and operations (NUM), measurement (MEA), geometry (GEO), data analysis and probability (DAT), and algebra (ALG). Complexity levels are: low (L) and medium/high (MH). Types of skill are: enabling skill (ES), isolated content (IS), and mathematical reasoning (MR).

While the subtest analyses presented in tables 15-17suggested that NAEP gains were more concentrated in the low-complexity items, the detailed review of the top-gaining NAEP items showed an overrepresentation of constructed response (CR) and medium/high-complexity items in this set. For example, these sensitive NAEP items were 40 percent constructed response (CR) for each condition, which is twice the proportion of constructed response items on NAEP as a whole. The NAEP high-gaining items were also 40 percent medium/high complexity as opposed to only 26 percent medium/high complexity for the test as a whole. There was, however, considerable overlap between these two item attributes within the set of NAEP high-gaining items: all of the constructed response items in this set were judged to be medium or high complexity in terms of cognitive demand, while only one multiple choice item (in the eighth-grade CMP condition) was classified this way.

As was suggested by the quantitative analysis of item gains displayed in figure 1 (in which the range of gains across BAM items was seen to be smaller than the range of gains across NAEP items), the BAM high-gaining items are relatively more homogeneous than the NAEP high-gaining items. Obviously BAM items are all constructed response items, and they are also more similar to one another in that they all were developed with the explicit intention of eliciting mathematical reasoning—particularly modeling in quantitative and geometric situations. This explicit attention to mathematical reason in test design was, in fact, the reason that we selected BAM to be the reform-oriented mathematics assessment in our study. Such reasoning proficiencies are expected to develop gradually over time if they are featured in instruction, and CMP explicitly incorporates pedagogies and content intended to develop these proficiencies. Table 18 shows the high proportion of the high-gaining BAM items that were judged by our content experts to require mathematical reasoning for their solution (MR).

NAEP is much more heterogeneous than BAM, both with respect to the range of item gains and the nature of the skills tapped by the high-gaining items. As noted above, NAEP constructed response items are more BAM-like in terms of cognitive complexity. In contrast, examination of the high-gaining, low-complexity NAEP items suggests that these items often tap proficiencies that are less related to mathematical reasoning and more focused on the language and presentation of mathematics problems. In Table 18, NAEP items are said to assess "enabling skills" (ES) if they required reading complicated instructions, applying basic terminology, or demonstrating basic skills in an unfamiliar context. In a few cases, NAEP high-gaining items also tapped "isolated" (IS) content, such as scientific notation, that is taught briefly and then disappears from the curriculum. In other cases the specific skills tapped by high-gaining NAEP items, such as locating points on the Cartesian plane, would have been taught in prior grade levels, yet appeared to be "instructionally sensitive" in the grade level we tested. Presumably this occurred because students were able to use and practice these skills sufficiently over the course of the year to dramatically improve proficiency by the time of spring testing.

State Test Effects

As explained previously, the comparison of sensitivity to gain was extended to the state assessment in one of our four states. For this analysis we had 308 students who not only had all of the necessary administrations of both NAEP and BAM, but also state test scores on a vertically aligned state test for both spring 2005 and spring 2006. The state test was one that would be considered moderately reform oriented.

Table 19 shows the effect sizes on all three of the assessments for this group of 308 students. The results suggest that for all three class conditions, the state test is less sensitive than NAEP to the type of instruction experienced by these students, and the degree of difference increases as one moves from seventh grade CMP, through eighth grade CMP, to eighth grade algebra. BAM and the state test are not significantly different in any of the conditions.

While these results are interesting, they should be viewed with caution because of the small sample size and the fact that the analysis was restricted to a single state. Moreover, it is important to note that NAEP and BAM are measuring fall-to-spring gains, while the state test is measuring spring-to-spring gains. To the extent that students "backslide" over the summer, one would expect higher student performance in the spring than the following fall and therefore less room for growth prior to the subsequent spring. The state test may, therefore, reflect a situation in which the actual student gains are smaller than those measured by NAEP and BAM.

Table 19. Gains from Time 1 to Time 2 Measured by NAEP, BAM, and the State Test, in Time 1 Standard Deviation Units, for a Common Sample of 308 Students in One State.[1]

Condition	N	NAEP	BAM	State Test
7th Grade CMP	99	0.450 (0.093)	0.418 (0.092)	0.264 (0.049)
8th Grade CMP	105	0.563 (0.124)	0.279 (0.107)	0.258 (0.053)
8th Grade Algebra	104	0.766 (0.118)	0.361 (0.117)	0.397 (0.077)

[1] Time 1 is fall 2005 for NAEP and BAM, but spring 2005 for the state test. Time 2 is spring 2006 for all three tests.

Significant differences:

NAEP vs. BAM: CM08 p=.0702; AL08 p=.0101

NAEP vs. State Test: CM07 p=.0620; CM08 p=.0227; AL08 p=.0068

BAM vs. State Test: no significant differences

Conclusions

The primary question addressed in this study was as follows: for students exposed to a well-implemented mathematics reform curriculum, how great, and how similar, are the one-year gains that can be measured on (a) BAM, an instrument well aligned with reform-curriculum objectives, and (b) NAEP, an instrument that has been designed more broadly to represent the union of curricular goals both across jurisdictions and across time?

Although we had initially hypothesized that BAM, being more closely aligned with the reform curriculum, would reveal larger gains than NAEP, we found that both assessments were equally sensitive to the gains of our sample of students in CMP classrooms, and NAEP appeared better able to detect gains in the algebra classrooms. This was true even though the BAM test required twice as much time to administer as the NAEP test. Moreover, the finding of equal or larger gains for

NAEP held up even when we looked only at high-gaining classrooms, as a check against the concern that the CMP curriculum may not have been well implemented in all of the sampled classrooms.

In an analysis that separated students based on their fall performance level, we found that both assessments registered greater effects for students who started in the bottom half of the distribution, suggesting some ceiling effect for both instruments. However, the general pattern of equal sensitivity for BAM and NAEP in CMP classrooms, and greater sensitivity for NAEP in algebra classrooms, still persisted.

When gains were examined separately by mathematical content area and level of complexity, further detail emerged. In particular, the greater sensitivity of NAEP for measuring gains in algebra classrooms seemed concentrated in the algebra, geometry, and measurement strands and in lower-complexity items. At the same time, the variability of individual item gains was greater for NAEP than for BAM in this class condition.

The concentration of NAEP gains in low-complexity items for all class conditions was particularly striking and suggested the hypothesis that NAEP and BAM appeared to produce "the same" results for seventh and eighth grade CMP classes because of offsetting, differential gains on low-complexity and medium/high-complexity items, respectively. This hypothesis was given additional support by a detailed analysis of the top-gaining items in each class condition. Our content expert reviewers judged that the highest gaining BAM items were more likely to tap mathematical reasoning skills than were the highest gaining NAEP items. (Although the highest gaining NAEP items were more heterogeneous and also included several BAM-like constructed response items.)

While the detailed analysis of item gains is useful for attempting to tease out the attributes of assessments that provide the best information on various aspects of student learning, they should not obscure the primary finding of this study—namely, the comparable (and sometimes greater) sensitivity of NAEP as compared with BAM. In light of this central finding, the concern that NAEP is underreporting achievement for students in reform-based classrooms seems clearly unfounded. Students in reform-based classrooms—and schools—appear to do at least as well on NAEP as BAM. Moreover, in the one state where a comparison to gains on the state test was possible, NAEP appeared to outperform the state test in all three conditions. This latter tentative conclusion must be tempered, however, by the small sample size available for the state test analysis and also by the fact that the state test was measuring spring-to-spring gains, while NAEP was measuring fall-to-spring gains. The differences in time of administration may mean that there was more room for actual growth on the NAEP and BAM assessments than on the state test due to summer learning loss.

References

American Association for the Advancement of Science (2000). *Middle grades mathematics textbooks: A benchmarks-based evaluation.* Available online at http://www.project2061.org/publications/textbook/mgmth/report/default.htm.

Chromy, J.R. (2005). *Participation standards for 12th grade NAEP.* Available online at http://www.nagb.org/pubs/reports-papers.htm.

Daro, P., Stancavage, F., Ortega, M., DeStefano, L., & Linn, R. (2007). *Validity study of the NAEP mathematics assessment: Grades 4 and 8.* A publication of the NAEP Validity Studies Panel, Palo Alto, CA: American Institutes for Research.

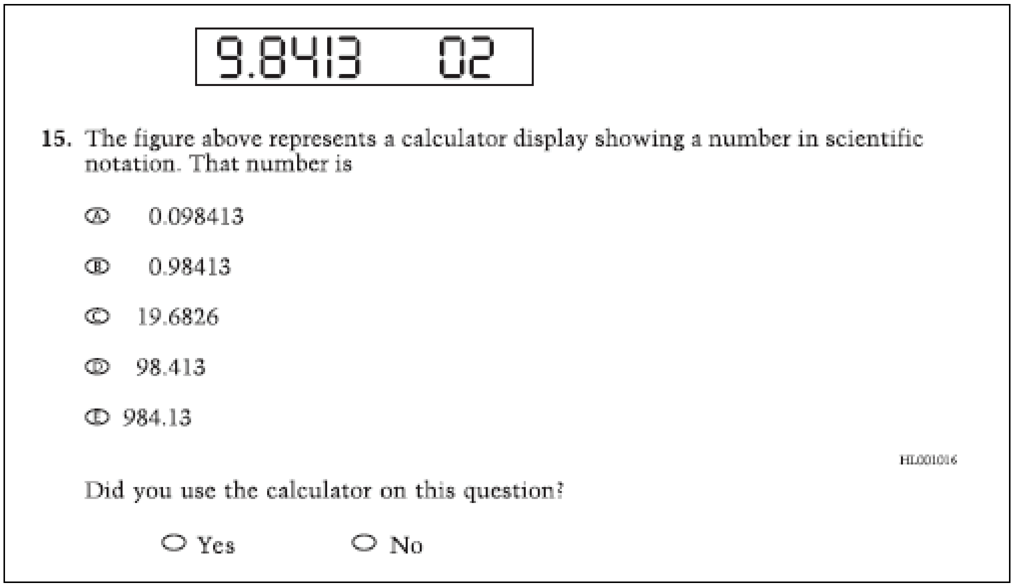
National Assessment Governing Board, U.S. Department of Education. (2004, September). *Mathematics framework for the 2005 National Assessment of Educational Progress.* Available online at http://www.nagb.org/pubs/m_framework_05/toc.html.

Pellegrino, W. J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academies Press.

Ridgway, J., Zawojewski, J. S., & Hoover, M. N. (2000). Problematising evidence-based policy and practice, *Evaluation and Research in Education, 14,* 181-192.

Schmidt, W. H., Jakwerth, P. M., McKnight, C. C. (1998). Curriculum sensitive assessment: Content *does* make a difference, *International Journal of Educational Research*, *29*, 503-527.

Appendix A – Sample NAEP Items



9.8413 02

15. The figure above represents a calculator display showing a number in scientific notation. That number is

Ⓐ 0.098413

Ⓑ 0.98413

Ⓒ 19.6826

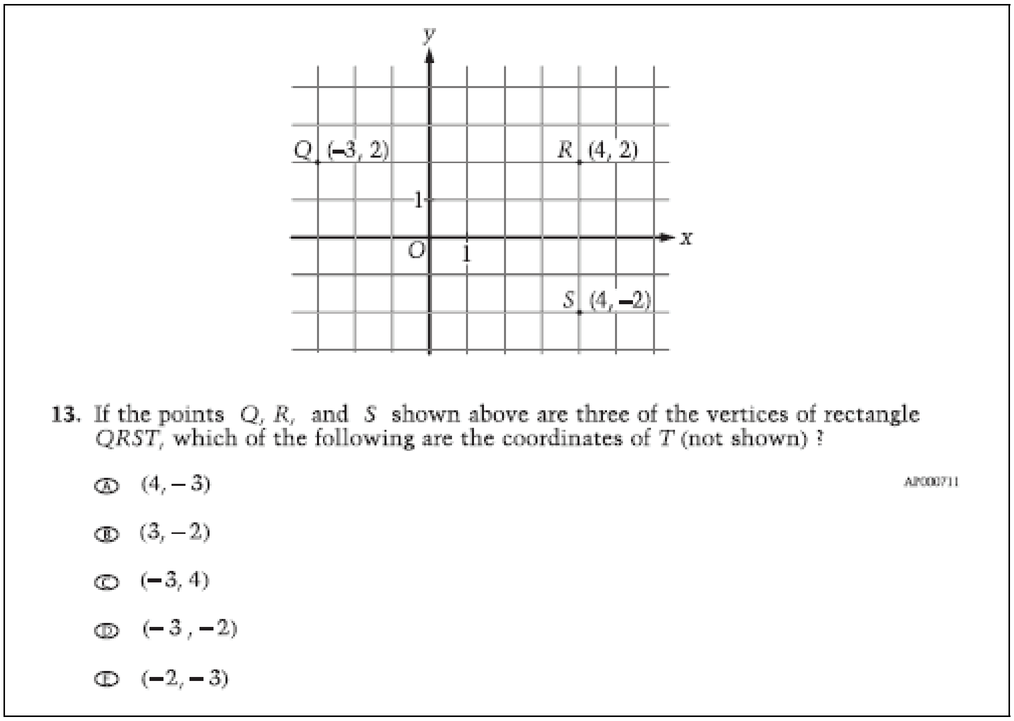
Ⓓ 98.413

Ⓔ 984.13

HL001016

Did you use the calculator on this question?

○ Yes ○ No



y
Q (–3, 2)
R (4, 2)
1
O
1
x
S (4, –2)

13. If the points *Q, R,* and *S* shown above are three of the vertices of rectangle *QRST,* which of the following are the coordinates of *T* (not shown) ?

AP000711

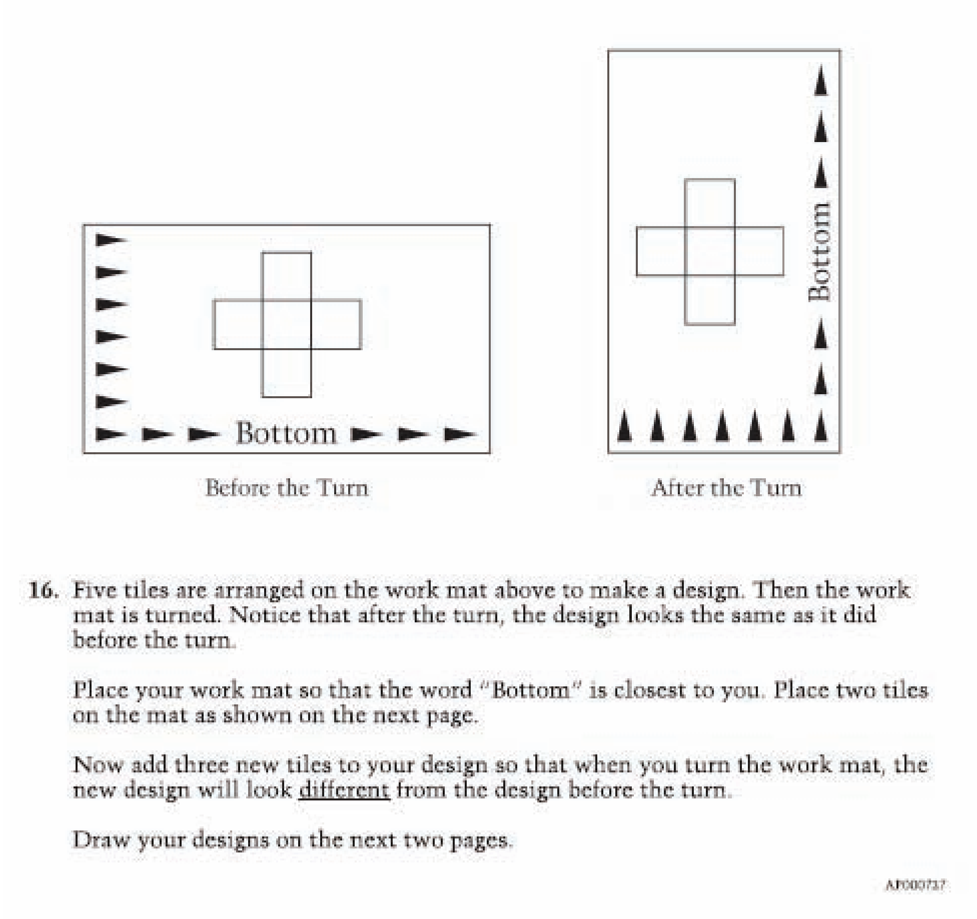
Ⓐ (4, – 3)

Ⓑ (3, – 2)

Ⓒ (–3, 4)

Ⓓ (– 3 , –2)

Ⓔ (–2, – 3)



Before the Turn

After the Turn

16. Five tiles are arranged on the work mat above to make a design. Then the work mat is turned. Notice that after the turn, the design looks the same as it did before the turn.

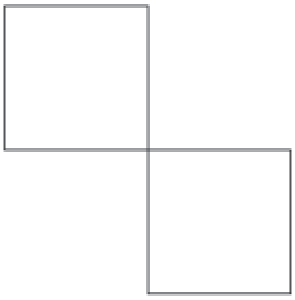
Place your work mat so that the word "Bottom" is closest to you. Place two tiles on the mat as shown on the next page.

Now add three new tiles to your design so that when you turn the work mat, the new design will look <u>different</u> from the design before the turn.

Draw your designs on the next two pages.

AP000737

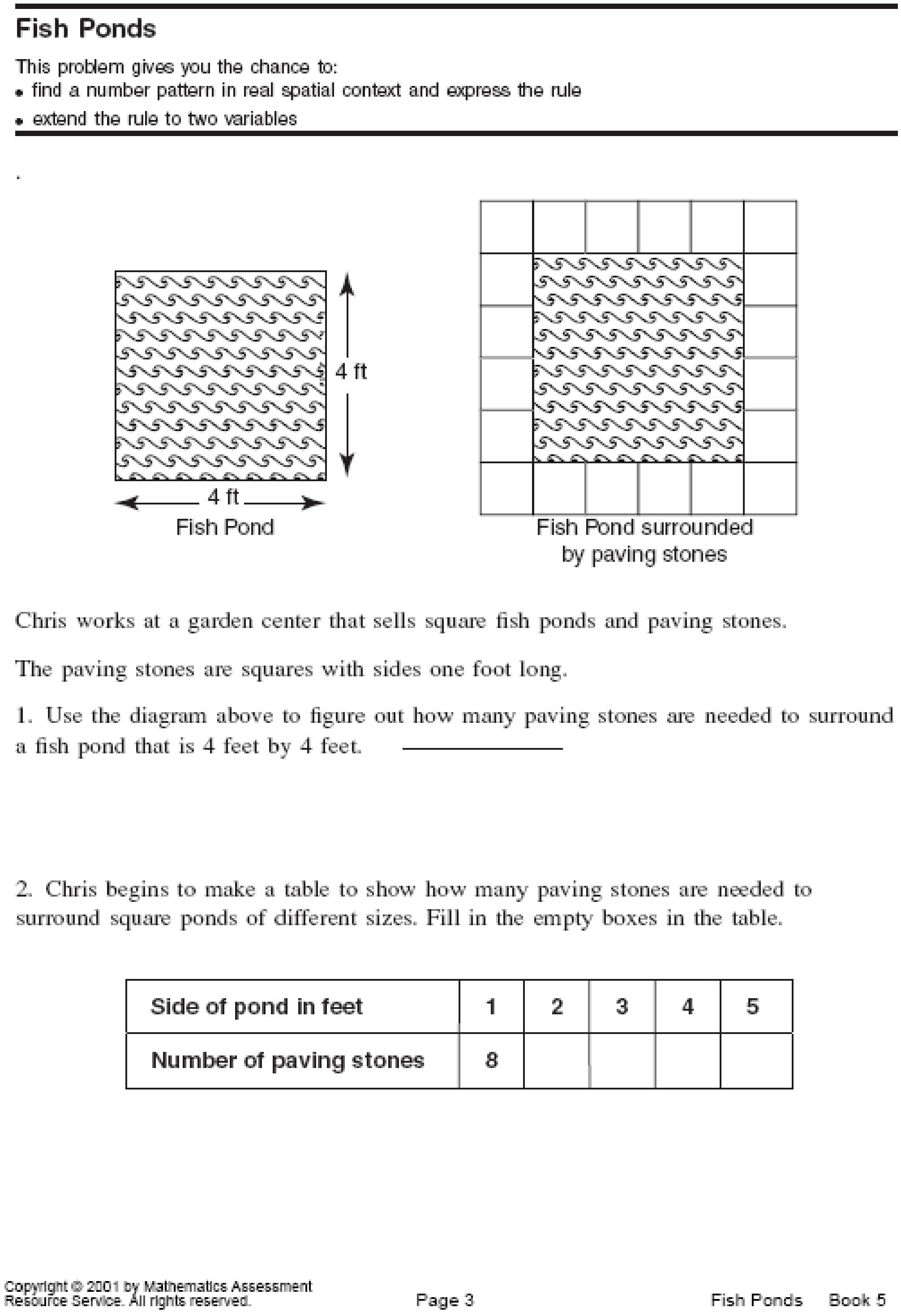
Draw your before design in the space below. Then draw your after design on the next page.



► ► ► Bottom ► ► ►

Draw your after design in the space below.

Appendix B – Sample BAM Task



Fish Ponds

This problem gives you the chance to:

- find a number pattern in real spatial context and express the rule
- extend the rule to two variables

Fish Pond

Fish Pond surrounded
by paving stones

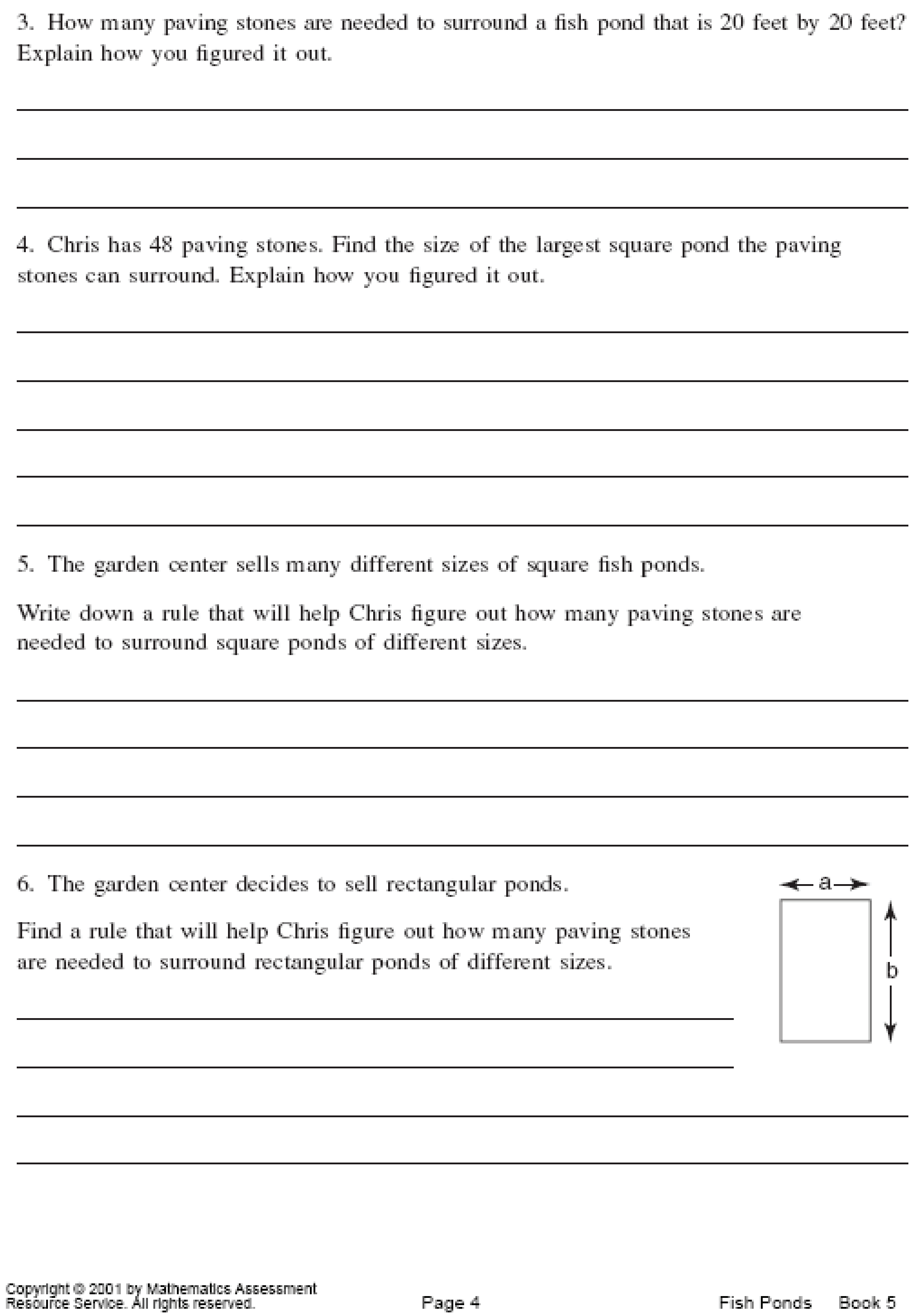
Chris works at a garden center that sells square fish ponds and paving stones.

The paving stones are squares with sides one foot long.

1. Use the diagram above to figure out how many paving stones are needed to surround a fish pond that is 4 feet by 4 feet. ____________

2. Chris begins to make a table to show how many paving stones are needed to surround square ponds of different sizes. Fill in the empty boxes in the table.

Side of pond in feet	1	2	3	4	5
Number of paving stones	8				



3. How many paving stones are needed to surround a fish pond that is 20 feet by 20 feet? Explain how you figured it out.

4. Chris has 48 paving stones. Find the size of the largest square pond the paving stones can surround. Explain how you figured it out.

5. The garden center sells many different sizes of square fish ponds.

Write down a rule that will help Chris figure out how many paving stones are needed to surround square ponds of different sizes.

6. The garden center decides to sell rectangular ponds.

Find a rule that will help Chris figure out how many paving stones are needed to surround rectangular ponds of different sizes.

Page 4

Fish Ponds Book 5

Appendix C – Analyses Using Booklet Percent Correct Metric

In addition to the main IRT-based analyses presented in the body of the paper, we also carried out a parallel set of analyses using a booklet percent correct metric. For BAM, this meant calculating the booklet percent correct, adjusted for booklet difficulty and spread as measured during the fall. Since students took two BAM booklets at each administration, booklet scores from the paired administrations were averaged to obtain a single fal, and a single spring percent correct score for each student. For NAEP, the percent correct was calculated by solving for each student's theta using the existing NAEP item parameters. The thetas were then used to estimate the pooled percent correct for all of the NAEP items. Table C-1 shows the means and standard deviations of the percent correct scores for fall and spring on each assessment and in each of the three conditions.

Table C-1. Mean Booklet Percent-Correct Scores (Standard Deviations) of the Students in Our Sample

Condition	**N**	**NAEP** **Fall**	**NAEP** **Spring**	**BAM** **Fall**	**BAM** **Spring**
7th Grade CMP	643	44.63 (17.28)	51.99 (19.82)	30.89 (18.72)	38.65 (20.59)
8th Grade CMP	535	45.63 (15.49)	51.57 (16.69)	32.38 (15.29)	37.86 (17.68)
8th Grade Algebra	606	69.86 (14.31)	75.98 (13.39)	58.97 (16.24)	63.45 (16.38)

NOTE: Because scores have been adjusted to account for differences in booklet difficulty, the potential range of percent correct scores is -10% to 110%. X's denote means, dark horizontal bars denote medians.

Table C-2 presents the fall-to-spring gains for each assessment, in each of the three conditions. As can be seen, the results are generally similar to the results obtained using the IRT scaling. That is, NAEP and BAM performed essentially the same for measuring gains in seventh and eighth grade CMP classrooms. In eighth grade algebra classrooms, NAEP detected greater gains than BAM.

Table C-2. Fall-to-Spring Gains in Booklet Percent-Correct, as Measured by NAEP and BAM, in Fall Standard Deviation Units, by Condition

Condition	N	NAEP	BAM	Difference
7th Grade CMP	643	0.426 (0.028)	0.414 (0.029)	0.011 (0.035)
8th Grade CMP	535	0.384 (0.034)	0.359 (0.035)	0.025 (0.044)
8th Grade Algebra	606	0.428 (0.032)	0.275 (0.035)	0.153*** (0.044)

NOTE: Significance of difference indicated by: $^{+}$ p<0.10, * p<0.05, ** p<.01, *** p<.001