

TECHNICAL REPORT

1999 Maryland School Performance Assessment Program (MSPAP)

Maryland State Department of Education
CTB McGraw-Hill
Measurement Incorporated

May 23, 2000

INTRODUCTION	6
TEST DEVELOPMENT	7
TEST ADMINISTRATION.....	11
SCORING	13
<u>QUALITY CONTROL</u>	15
<u>CONCLUSION</u>	17
SPECIAL ISSUES.....	17
<u>MATHEMATICS OUTCOMES</u>	17
<u>SCALING OF MATHEMATICS</u>	17
<u>ALGORITHMIC SCORING</u>	18
<u>STUDENT PARTICIPATION IN MSPAP</u>	19
SCALING AND EQUATING.....	20
<u>ITEM SET CALIBRATIONS</u>	20
<u>EQUATING STUDIES</u>	22
<u>Equating the Content Area Scores Across Clusters</u>	23
<u>Rater Year Effects Study</u>	24
<u>Equating 1998 and 1999 Scale Scores</u>	26
<u>Comparison of 1998 and 1999 Mean Scores</u>	28
RELIABILITY	29
<u>COEFFICIENT ALPHAS</u>	29
<u>STANDARD ERRORS OF MEASUREMENT FOR PROFICIENCY LEVEL CUT SCORES</u>	30
VALIDITY.....	30
<u>BETWEEN CONTENT AREA CORRELATIONS</u>	30
<u>TEST DIFFICULTY CONCERNS</u>	31
<u>CONTENT VALIDITY EVIDENCE</u>	31

<u>OUTCOMES COVERAGE</u>	32
<u>FACE VALIDITY EVIDENCE</u>	ERROR! BOOKMARK NOT DEFINED.
<u>CONSTRUCT VALIDITY</u>	32
<u>STATISTICAL TEST BIAS</u>	32
<u>CONSEQUENTIAL VALIDITY EVIDENCE</u>	34
<u>CONCLUSION</u>	34
SCORE INTERPRETATION	35
<u>SCALE SCORES</u>	35
<u>PROFICIENCY LEVEL DESCRIPTIONS</u>	35
<u>SCHOOL PERFORMANCE STANDARDS</u>	36
<u>INDIVIDUAL STUDENT SCALE SCORES</u>	38
<u>OUTCOME SCORES</u>	38
MSPAP SCORE REPORTS	39
REFERENCES	41
TABLE 1	45
NUMBERS OF TEAMS, READERS, AND SCORING LEADERS	45
TABLE 2	46
READER ACCURACY SET MEAN SCORES BY TEAM - GRADE 3	46
TABLE 3	47
READER ACCURACY SET MEAN SCORES BY TEAM - GRADE 5	47
TABLE 4	48
READER ACCURACY SET MEAN SCORES BY TEAM - GRADE 8	48
TABLE 5	49
FREQUENCY OF ACCURACY SET MEAN SCORES BY GRADE	49
TABLE 6	50

READER ACCURACY SET MEAN SCORES BY CONTENT AREA - GRADE 8.....	50
TABLE 7	51
FREQUENCY OF ACCURACY SET MEAN SCORES BY CONTENT AREA - GRADE 8.....	51
TABLE 8. SUMMARY FINDINGS FROM CALIBRATIONS.....	53
TABLE 9. CALIBRATION FOR CLUSTERS WITH CHOICE SETS.....	55
TABLE 10. CLUSTER EQUATING RESULTS	56
TABLE 11.....	59
RATER YEAR EFFECTS STUDY PERFORMANCE (98SS ₉₈) OF STATE SAMPLE ON 1998 MSPAP.....	59
TABLE 12.....	60
RATER YEAR EFFECTS STUDY RAW SCORE COMPARISONS	60
TABLE 13.....	61
1992, 1993, 1994, 1995, 1996, 1998, AND 1999 RATER YEAR EFFECTS STUDIES:.....	61
TABLE 14.....	62
RATER YEAR EFFECTS STUDY TRANSFORMATION VALUES.....	62
TABLE 15.....	63
PERFORMANCE OF STATE ON 1998 MSPAP AND 1999 EQUATING SAMPLE ON 1998 MSPAP	63
TABLE 16.....	64
EQUATING STUDY TRANSFORMATION VALUES.....	64
TABLE 17.....	65
COMPARISON OF 1998 AND 1999 MSPAP PERFORMANCE BY GRADE AND SCALE	65
TABLE 18. COEFFICIENT ALPHA FOR 1999 MSPAP CONTENT AREAS	66
TABLE 19. STANDARD ERRORS OF MEASUREMENT - GRADE 3.....	67
TABLE 20. STANDARD ERRORS OF MEASUREMENT - GRADE 5.....	68
TABLE 21. STANDARD ERRORS OF MEASUREMENT - GRADE 8.....	69
TABLE 22. BETWEEN CONTENT AREA CORRELATIONS FOR GRADE 3	70

TABLE 23. BETWEEN CONTENT AREA CORRELATIONS FOR GRADE 5	71
TABLE 24. BETWEEN CONTENT AREA CORRELATIONS FOR GRADE 8	72
TABLE 25. BETWEEN CONTENT AREA SCALE SCORE CORRELATIONS AT SCHOOL LEVEL FOR GRADE 3.....	73
TABLE 26. BETWEEN CONTENT AREA SCALE SCORE CORRELATIONS AT SCHOOL LEVEL FOR GRADE 5.....	74
TABLE 27. BETWEEN CONTENT AREA SCALE SCORE CORRELATIONS AT SCHOOL LEVEL FOR GRADE 8.....	75
TABLE 29. OUTCOME DIFFICULTY INDICATORS	77
APPENDIX A	78
TEST MAPS FOR 1999 MSPAP	78
APPENDIX B	79
NUMBER OF ITEMS COMPRISING EACH OUTCOME FOR 1999 MSPAP	79
APPENDIX C	80
SCALED SCORE RANGES FOR EACH PROFICIENCY LEVEL	80

TECHNICAL REPORT
1999 Maryland School Performance Assessment Program (MSPAP)

Maryland State Department of Education
CTB McGraw-Hill
Measurement Incorporated

May 23, 2000

Introduction

Maryland School Performance Assessment Program (MSPAP) assessments are criterion referenced performance tests designed, developed, and implemented by the Maryland State Department of Education (MSDE) in collaboration with classroom teachers and other Maryland educators. MSPAP is the major strategy for implementing Maryland's reform initiative and provides information relevant to assessing school performance and guiding school improvement plans and activities. The primary focus of the information provided from MSPAP assessments is *schools*, although information about individual student performance is also available.

Each May since 1991, MSPAP has been administered to Maryland students in grades 3, 5, and 8. Each student participates in nine hours of testing (reading, writing, language usage, mathematics, science, and social studies) over a five-day period, approximately one hour and 45 minutes of testing time per day. The assessments are based on the Maryland Learning Outcomes (available on the Maryland State Department of Education's website at <http://mdk12.org>) that were adopted by the State Board of Education in 1990.

MSPAP is comprised of three test forms, or clusters, and one equating form or cluster from the previous year's test per grade (e.g., 3A, 3B, and 3C). Clusters are non-parallel test forms because content areas are spiraled throughout each cluster. For example, in social studies, *Peoples of the Nation and the World*, *Geography*, and *Economics* might be assessed in one cluster; *Political Systems*, *Peoples of the Nations and the World*, and *Economics* in another cluster; and *Political Systems*, *Geography*, and *Peoples of the Nations and the World* in the third cluster. Each test form or cluster assesses a combination of reading, writing, language usage, science, social studies, and mathematics.

Students are randomly assigned to testing groups. Random testing groups help to ensure that groups of students assigned to take each test cluster are heterogeneous in ability. In addition, random testing groups minimize influences on student performance that may occur when students are assessed in intact classroom groups by their regular classroom teachers.

Test clusters are assigned randomly to testing groups within schools and across schools in each school system and the state. Local Accountability Coordinators (LACs) implement a simple procedure (spiraling) to ensure this random assignment. Spiraling also ensures that the numbers of clusters administered within each school system and across the state will be nearly equivalent, and that schools with only three testing groups will always be assigned each of the three clusters. The Maryland State Department of Education's (MSDE's) Assessment Office approves final cluster assignments.

MSPAP is equated across years through random equivalent groups and equating clusters. Equating clusters are assigned to a representative sample of schools that have four or more testing groups in a grade and that were not used in the previous year's equating sample. Each equating cluster is given a test from the previous year's MSPAP administration so that the current year's test can be adjusted for difficulty.

Test Development

MSPAP assesses school performance on the Maryland Learning Outcomes through assessment tasks--collections of inter-related assessment activities or "items" that are organized around a theme (e.g., *Recycling* or *Salinity*). Tasks require students to respond to questions or directions that lead to a solution of a problem, a recommendation or decision, or an explanation or rationale for the responses. Some tasks assess one content area; other tasks assess multiple content areas. Activities comprising the tasks may be group or individual activities; hands-on, observation, or reading activities; and/or activities that require extended written responses, limited written responses, lists, charts, graphs, diagrams, webs, and/or drawings.

Test development consists of five phases: planning, design, development, review and revision, and field testing followed by further revisions.

Planning. MSDE instructional and assessment staff select tasks from previous MSPAP administrations to be reused. Staff then determine the learning outcomes needed to complete test clusters and plan new tasks to assess the outcomes. Up to 50% of the test may consist of reused or rolled over tasks.

Design. MSDE instructional staff members write task outlines comprised of a topic area, the time allotted for the task, and the outcomes to be assessed. They design calendars showing the types of test activities and the balance of content areas for each day of testing.

Development. Approximately 170 Maryland teachers across grades 3, 5, and 8 are recruited, screened, and hired by MSDE to write MSPAP tasks and activities; develop scoring tools; and write test administration directions. Task writers are given specifications for the content areas and outcomes to be assessed; the number

of assessment activities per outcome and task; and the background reading materials to be used in the assessment.

Task writers are trained on the principles of performance assessment, characteristics of MSPAP, bias and sensitivity issues, and Maryland Learning Outcomes. They receive information on scoring, measurement, and administration issues; and guidelines for developing graphics and selecting tools and materials. Task writers also receive concentrated training in the areas for which they are responsible: task writing, scoring, or test administration.

Task writers develop drafts of tasks to which reading and writing cues and prompts are added where appropriate. MSDE specialists and task writers participate in an extended review and revision process that includes raising questions and resolving issues and concerns about the tasks.

One characteristic of MSPAP is the use of authentic texts. Local school media specialists select reading materials in topic areas, and reading content area staff review the materials for bias, sensitivity, and readability. After third and fifth grade “average readers” read the materials with the state reading specialist, an analysis is conducted to determine if the readability is appropriate. Only materials that average readers can read independently and show evidence of construction of meaning are used in MSPAP.

Task writers select materials that can be used in their entirety. Occasionally, the publisher/copyright owner will not grant permission to use a text or material, and the task must be altered to accommodate other materials.

After tasks have been drafted, they are examined to see that all activities provide a measure of the intended outcomes. Draft scoring tools, answer cue information, and sample responses are then developed. MSDE specialists and staff from the scoring contractor for MSPAP (Measurement Incorporated) review draft scoring tools and test booklets (*Answer Books*, *Resource Books*, and *Examiner’s Manuals*) to identify problems. They then make revisions where necessary.

Review and Revision. MSPAP tasks are reviewed for:

- technical soundness,
- feasibility,
- controversial and sensitive topics,
- developmental appropriateness,
- scorability, and
- clarity.

Psychometric specialists conduct technical reviews that include verifying the number of outcome measures in a content area and test cluster and the independent responses in a content area. At least eight independent outcome measures for each content area in each cluster are needed for scaling purposes. Four measures for each outcome measured in a cluster are needed to calculate outcome scores. The test design specifies that an outcome be measured in at least two clusters within a grade.

Local Accountability Coordinators (LACs) and assessment staff conduct feasibility reviews that include examining tasks for:

Timing - Is adequate time allotted to tasks? Are the time blocks listed correctly in test materials?

Ease of Administration - Can tasks be administered by all teachers using the same directions?

Setting - Will all classrooms accommodate the administration of each task?

Clarity and Complexity of Directions - Are directions clear and concise?

Cluster Balance - Are content area tasks evenly distributed throughout the week? Are tasks varied within a day?

Formatting - Is there adequate student response space in the *Answer Book*?

Tools and Materials - Are materials appropriate? Adequately described? Feasible to administer? Cost effective?

Assessment and content staff conduct controversial and sensitive topic reviews in which they examine tasks for controversial language, stereotyping, and treatment of minorities, genders, and persons with disabilities. To ensure that MSPAP is free from controversial and sensitivity topics, task writers use *Guidelines to Avoid Bias and Sensitivity* that were adapted from *Bias Issues in Test Development* published by the National Evaluation System, Inc. (National Evaluation System, 1991). During the 1999 editorial review, the editors of CTB McGraw-Hill reviewed MSPAP for bias and sensitivity following the Macmillan/McGraw-Hill publication guidelines (Macmillan/McGraw-Hill, 1993).

Third and fifth grade teachers, educational psychologists, and early learning university faculty conduct developmental appropriateness reviews, to ascertain that assessment tasks are developmentally appropriate for the grade level in which they are to be administered.

Assessment specialists and experienced MSPAP Scoring Coordinators conduct scorability reviews to verify that tasks are scorable and that they yield meaningful measures of what

students understand and are able to do. Outcome/activity matches, that identify the outcome(s) being assessed by each activity, are verified.

Content specialists conduct clarity reviews to confirm that tasks are clearly written.

After MSPAP tasks have been reviewed, they are organized into an *Answer Book*, a *Resource Book*, and an *Examiner's Manual* for each grade and cluster (3A, 3B, 3C; 5A, 5B, 5C; 8A, 8B, 8C). All test booklets are then reviewed and edited for consistency, accuracy, organization, and comprehension.

Role playing is conducted to ensure that directions and timing are clear and correct. One MSDE specialist is the “teacher” and the other is the “student.” They use the *Answer Book*, *Resource Book*, and *Examiner's Manual* as if they were taking the test. This mock administration allows for cross checking of all materials the students and test administrator will need during the actual test administration.

Field Testing. A field test is conducted to collect information on the feasibility of conducting tasks in a classroom setting, clarity of directions to students and examiners, reliability of tools and materials, and timing and scorability of tasks.

In October 1998, schools in the School District of Philadelphia administered the 1999 MSPAP field test. Schools with student populations that closely matched Maryland's population with respect to race/ethnicity and gender were chosen. In addition, in the selected schools, reading/writing instruction, collaborative learning, and hands-on learning were part of daily instruction. All new tasks appearing on the 1999 assessment were administered to two classrooms, each containing 25 to 30 students.

Observers from Maryland monitored the testing process to determine whether timing, directions, questions, or materials needed to be revised. As a result of field test administrative and scoring feedback, some tasks were slightly revised to correct timing, directions, and confusing questions. After the revisions were made, a post field test meeting confirmed that the test was ready for the May 1999 administration. Additional information on the field test may be obtained from MSDE (Westat, 1999).

Field test responses also helped to identify possible anchors (range finding), training, and sample responses for use in scorer training. These sample responses were selected to represent all possible score points and were based on exact agreement after discussion. (Additional sample responses for scorer training were selected from live responses “hijacked” after the MSPAP operational administration in May 1999.)

Development of Scorer Training Materials. Following field test scoring, the scoring contractor reviewed and revised scoring tools, answer cues, and sample responses to create scoring guides for each task. Each activity was presented, followed by the scoring

tool and answer cue information (typical response content, key ideas, etc.). Sample responses were selected to illustrate each score point. In the few instances in which field test scoring had not yielded any samples at a given score point, a teacher-developed sample response was utilized. Responses from the May 1999 administration supplemented these teacher-developed samples. Scoring guides were task-specific, with the exception of language in use. This generic guide was used for anchor responses to a wide array of language usage items.

The scoring contractor's senior staff developed detailed annotations to assist the Maryland-based scoring team coordinators and team leaders to train their teacher teams on scoring MSPAP. In addition, supplementary guides dealing specifically with poetry were developed to assist the expressive writing teams to apply the genre-general rubric to this particular expressive form.

Preparation of Scorer Training Materials. Training materials (training and qualifying sets) were prepared using field test and operational responses. Training sets were used for instruction and practice in task scoring. Qualifying sets were used to test the readers' ability to score accurately and to supplement the training provided by the training sets. These sets included responses from all activities to be scored by the team and were formatted to resemble the portion of the *Answer Book* that the team would score. Work was also begun on the accuracy sets that would be used twice a week during scoring to diagnose and prevent individual and/or room-wide drift away from scoring criteria. These sets closely resembled the qualifying sets described above. Preparation of training materials continued to mid June, when training began.

Pre-Packaging of Manipulatives. Tools and manipulatives for hands-on activities are pre-packaged for each testing group and its examiner by contractors. The materials are delivered to elementary and middle schools in school systems electing to use the service. When possible, materials are pre-cut or pre-measured, such as the amount of detergent or soil, and packaged for each student or teacher.

Test Administration

Each May, the tests are administered in Maryland elementary and middle schools—to third and eighth grade students each morning of the first week; to fifth grade students each morning of the second week.

When tests are delivered to schools, they are signed for, inventoried, and immediately placed in secure storage. Two weeks prior to testing, school test administrators review test materials (*Examiners Manual, Answer Book, and Resource Book*) for only the cluster they will administer.

MSPAP questions are designed to elicit a variety of answers based on various kinds of

information and presented in diverse ways. Responses might involve writing sentences, making lists, writing essays, sketching drawings, or creating tables or graphs.

Students use two booklets in taking the test: a *Resource Book* and an *Answer Book*.

The *Resource Book* contains supplementary or resource materials, such as stories, maps, and charts, or other information a student needs to complete test activities. There are three versions of the *Resource Book* for each grade level, one for each form of the test.

Students also use an *Answer Book* that contains test questions and space for recording responses. For some items, students use information in the *Resource Book* to work in small groups on “pre-assessment” or group activities to help them focus on a test question. Group interaction ends before students begin work in their *Answer Books*, which is always done individually. Pre-assessment activities set the context for a test item, but do not cue or provide an answer.

Teachers use an *Examiner’s Manual* to administer each form of the test. The *Examiner’s Manual* contains specific instructions on how to administer each MSPAP task during the entire five-day testing period. The *Examiner’s Manual* is a script that clearly tells the test examiner exactly what to say and do to move students through the test. It does not allow a test examiner to improvise in providing directions nor to provide examples unless such examples are included in the script. The purpose is to allow all students a fair chance by standardizing the way the test is given in all schools throughout the state.

Test Administration and Coordination Manual. A *Test Administration and Coordination Manual* provides information on test security and on specific test procedures to Local Accountability Coordinators who are responsible for test administration in local school systems. MSDE trains Local Accountability Coordinators in test administration. They, in turn, provide training to school test coordinators who are responsible for test administration in schools. School test coordinators train the teachers who will administer the test.

Eligible school test examiners are state-certified academic, special education, gifted and talented, English as a Second Language (ESL), and Chapter 1 classroom teachers. Test examiners are responsible for the smooth and standardized test administration and the protection of secure test materials. School staffs not eligible to serve as test examiners may provide assistance during test administration as proctors only. Proctors assist the test examiner with the distribution and collection of testing materials and monitor the testing behaviors of students by keeping them on task. Proctors may not have access to secure test materials.

Participation of all grade 3, 5, and 8 students in MSPAP, except those excused or exempted according to MSDE policy, is mandatory. The compulsory school attendance

law and State Board of Education regulations on public school standards are the bases of MSDE's policy of mandatory participation.

MSPAP Observations. In May 1999, MSDE staff conducted MSPAP observation in which they observed the MSPAP administration to see how teachers, school staff, and students responded to tasks and to gather information on the administration. Test examiners submitted comments about the test on a "Concerns or Comments on the Administration of the MSPAP" form. Some examiners made general comments; others commented on specific tasks. Since some tasks will be reused in the next year's administration, comments were reviewed in MSDE roundtable discussions. Based on the comments and concerns of test administration observations and the feedback from teachers, tasks are adjusted as necessary before they are administered again.

After the test has been administered, all test booklets and materials are returned to the test contractor in the same boxes in which they arrived. All scrap materials are destroyed.

Scoring

Four teams of Maryland teachers scored the assessment activities in each test form for each of the three grades using scoring guides developed by Measurement Incorporated (MI) project staff, scoring tools generated by Maryland educators, and selected sample responses chosen by Maryland educators. Each team scored the open-ended student responses and assigned the appropriate score point on a customized scan sheet. During June and July 1999, *Student Answer Books* for approximately 184,750 students were scored.

The four school sites and scoring assignments for 1999 were:

Clusters A and	Grades 3, 8:	Mattawoman Middle School, Charles County, Waldorf
Clusters A and	Grade 5:	Grasonville Elementary School, Queen Anne's County, Grasonville
Cluster B;	grades 3, 5, 8:	Western School of Technology and Environmental Sciences, Baltimore County, Baltimore
Clusters B and C;	grades 3, 5, 8:	Chesapeake High School, Baltimore County, Baltimore

All booklets for a given grade/cluster were scored at the same site due to measurement implications of a multi-site model.

MSDE and MI staff estimated that it would take approximately 25 minutes of reader time to score all scorable units in the answer booklet for each of the 3 clusters at each of the 3 grades, for each of the 9 grade/cluster combinations.

So that work loads were reasonable, the scorable units within each of the 9 grade/cluster combinations were distributed across 4 teams. At the eighth grade, a team for each of the four content areas (mathematics, science, social studies, and reading/writing/language usage) scored within their subject areas to the greatest degree possible. Each team scored assessment activities within one primary content area, although content area integration sometimes required that teams address multiple content areas. When integration occurred, enhanced training ensured accurate score decisions by all team members. Additionally, teams were selected to provide a good “fit” with the content areas being addressed by the task(s) being scored by a team. For example, a reading/science task would be predominately scored by a team of science and English/language arts specialists.

At grades 3 and 5, where most teachers work across subject areas, it was not considered crucial that each scoring team score items in only one content area. It was important to attempt to equalize reader scoring time per team, and to ensure that no one team was responsible for too many items requiring mentally demanding, complex thought processes, which might negatively affect the accuracy of readers and teams due to mental fatigue.

Staffing and Reader Distribution Throughout Scoring Sites. For each grade and cluster, four teams scored a unique set of MSPAP items--a total of 12 teams per grade and 36 teams across three grades. For each team, the data processing contractor provided a customized answer sheet. Each student’s answer booklet had four customized answer sheets included with it when delivered to the scoring site.

Based upon six years of experience, MI project management established a target of 744 readers to score the 1999 MSPAP assessment, with each reader working 18 to 20 days after 2 to 3 days of training and qualifying. The number of readers required for each team varied depending upon the estimate of the relative scoring time per customized answer sheet after the 36 teams had been created. The average number of readers per team was 21. However, team size varied from 13 to 29 readers distributed across sites, grades and clusters as shown in Table 1.

See Table 1

Two leadership positions were assigned to each scoring team: a Scoring Coordinator and a Team Leader. Scoring Coordinators received five days of training by MI Project Leaders to prepare them for training readers (scorers) on their teams, monitoring readers for quality and production during the scoring process, and administering scoring in concert with MI project staff. Team Leaders, who assisted Scoring Coordinators, received three days of training.

Quality Control

Scoring accuracy is maintained by: check sets, accuracy sets, spot checks, and retraining.

Check sets, covering all MSPAP tasks, were administered on Monday mornings to help Scoring Coordinators and Team Leaders determine whether individual readers and the team of readers were continuing to score accurately and consistently, especially on items that were complex and difficult to score. As scoring progresses, readers may “drift” away from score points, especially after a weekend away from scoring. As inconsistencies and inaccuracies were detected, Scoring Coordinators and Team Leaders held discussions with the team and assisted individual readers to improve accuracy.

Accuracy sets were administered on Tuesday and Thursday mornings to determine whether teams of readers maintained appropriate levels of accuracy during the scoring process. Each accuracy set included a student response for each scorable unit, and each reader's average score was recorded so that the mean score for each accuracy set could be calculated. These mean scores were used to construct Tables 2 through 7, which will be used to analyze quality control for this scoring project.

Readers in 35 of the 36 teams were given at least 5 accuracy sets, usually 6 to 7 sets. Readers who scored below 70 percent on any accuracy set received additional training immediately from the Scoring Coordinator or the Team Leader and were released from retraining only after the leaders determined that scoring problems were resolved. If the scoring problems were not resolved, the reader was dismissed from the scoring project.

In *spot checking*, a Scoring Coordinator or Team Leader rescored a booklet to estimate a reader's overall accuracy, to determine specific items with which a reader was having difficulty, or to ascertain specific items that were causing individual readers to perform poorly on check sets or accuracy sets.

In *retraining*, Team Leaders or highly accurate readers used the scoring guide and student papers to assist readers who had experienced problems maintaining appropriate accuracy levels. Small groups of readers who shared a common scoring difficulty were also retrained to improve their scoring accuracy.

Reader accuracy results. In 1999, 206 accuracy sets were administered across all 36 scoring teams. The reader accuracy set mean scores for each scoring team are shown in Tables 2, 3, and 4 for grades 3, 5, and 8 respectively.

 See Tables 2-5

The results are summarized in Table 5 by grade and across all three grades. The results are reasonable and acceptable for scoring open-ended performance assessment items. Fifty-five percent of the sets had mean scores between 80 to 89%, and 28 percent were at or above 90% accuracy. Seventeen percent had mean set scores between 70 to 79%, and only three of the accuracy set mean scores were below 70% accuracy. The results for the 1999 MSPAP were similar to those for the previous three years. The accuracy set mean scores were similar to past years.

The averages across the accuracy sets for each team could be calculated because the sets contained the same number of scorable units. However, it was not possible to calculate the averages across different teams because the numbers of scorable units varied considerably from team to team. When the accuracy set mean scores were studied in terms of content area, the results were reasonably predictable yielding no major surprises.

Bearing in mind that few teams addressed only one content area, it is possible to look at results for predominant content areas in the eighth grade. Results by content area for the eighth grade are displayed in Tables 6 and 7. From past scoring of performance assessments it was reasonably predictable that the scoring of mathematics would yield relatively higher and somewhat more consistent accuracy set scores.

 See Tables 6-7

In grades 3 and 5, the items to be scored within each content area were distributed across teams to such a degree that it was not possible to analyze accuracy set mean scores systematically by content area. Past experience in scoring open-ended performance assessment items indicated that the relationships between content area and accuracy set scores at grades 3 and 5 would be similar to those at grade 8. In addition, MI Project Leaders and the Scoring Coordinators and Team Leaders felt that it was more difficult to train readers to score items consistently in reading/writing/language usage than in other content areas. These responses more often measure higher level skills and objectives; and they more often require holistic scoring decisions rather than more discrete decisions.

Conclusion

The factors that interacted to produce improvements in training and scoring productivity are:

Early field testing to provide an adequate time frame for scoring booklets, selecting training materials, and preparing annotated scoring guides.

An adequate time frame for planning and implementing activities for both CTB (the data processing contractor) and MI.

Increased experience of MI and Maryland project staff. Many readers and leadership staff in Maryland had not only gained another year's experience in scoring MSPAP activities, but had also become increasingly involved in other MSPAP activities, such as task development or rangefinding (field-test scoring).

Special Issues

Mathematics Outcomes

Prior to the 1996 MSPAP, 13 mathematics outcomes were measured, more than twice as many outcomes as were measured in other content areas. The number of measures needed in a cluster made designing the mathematics component difficult and often made individual tasks too long. Therefore, some mathematics outcomes were combined, thereby reducing the number of mathematics outcomes to nine. All mathematics outcomes are still tested, but there are fewer mathematics measures. For example, because geometry and measurement were combined, instead of needing four measures of each outcome for reporting purposes, only four total measures are needed. The supervisors of mathematics in each school system accepted this change.

The 1999 MSPAP included limited problem solving. The problem-solving outcome has been difficult to include in the test because of the scope of true problem solving. Additionally, scoring time and training needed to be slightly modified. However, it was important to include problem-solving activities because of their emphasis at the national and state levels.

Scaling of Mathematics

Mathematics Content and Mathematics Process have been scaled separately because of high local item dependency (CTB, 1992) since 1991. However, Mathematics Process has been a short test (MSDE, 1998) with lower reliability than Mathematics Content.

Furthermore, it is the source of frequent item calibration and test equating problems. MSDE and CTB conducted a series of research study investigating the feasibility of scaling Content and Process items together since 1995. The 1996 and 1997 Content and Process calibration data were used for the study. The study suggests that scaling Content and Process together did not create more fit problems, as compared to separate scaling. Furthermore, the local-item-dependency problems seen in the common scaling were similar in magnitude to those of the separate scaling. Based on these results, the National Psychometric Council approved the plan to combine Content and Process into a common scale starting 1999.

Mathematics Content score and Process score have been averaged to yield the Mathematics Total score since 1991. The main psychometric concern of scaling Content and Process together is the comparability of the Mathematics total score across years. A series of linking studies were conducted to link the newly created combined Content and Process scale to the Mathematics Total scale using data from the 1998 equating study. Linear approximation of the equipercentile equating was used to align the means and standard deviations of the two distributions. The linking was successful. Scale scores on this combined Mathematics test were easily transformed onto the Mathematics Total score scale. Detailed descriptions of the special re-scaling and linking project can be obtained through MSDE.

Algorithmic Scoring

Prior to 1995, students who were absent on one or more days of MSPAP testing could not obtain a content area scale score if they missed any day on which the content area was assessed. Algorithmic scoring is a process for deriving a score for students who were absent, but who had 60% or more of the responses in a content area and a minimum of eight independent measures.

Algorithmic scoring uses a maximum-likelihood estimation, which is a general method of finding good parameter estimates in a model. Since table scoring is based on complete score records, the ability estimates of absent students are inaccurate (underestimated). Therefore, students scored algorithmically can have their ability more accurately estimated using a maximum likelihood estimator, which approximates student ability using the data available. Beginning with the 1995 MSPAP, CTB McGraw-Hill scored all students algorithmically. (Before 1995, CTB used table scoring.)

To be eligible for algorithmic scoring, a student must have attempted at least 60% of the content area and at least eight independent items. Exceptions include the content areas of writing and language usage, as well as any “short” test. Because writing is a three-item test, if a student responds to the extended writing prompt (scored 0-3) and to one of the two limited writing prompts (scored 0-2), then a student should receive a score. (From 1992 to 1994 only one extended and one limited writing process comprised the writing

test. Therefore, MSPAP added another limited writing process to the writing scale in 1995. If students missed one of the limited writing process prompts, they still received a writing score.) Language usage is the content area most vulnerable to absence vulnerability because language usage measures are captured throughout the week. Therefore, language usage is scored for absent students as long as six or more of the responses in the student's language usage vector have either valid scores or score codes. Score codes are assigned when the student response is invalid which may be a blank, an off-task, or an nscorable response.

Algorithmic scoring increased the number of students who received at least one score. In 1999, across all grades and content areas, more than 15,000 more scores were computed using algorithmic scoring. This method of scoring gave a more accurate reflection of student performance in a school or school system.

Student Participation in MSPAP

It is the policy of Maryland to include all students to the fullest extent possible in all state assessment programs. Testing accommodations that meet state guidelines are provided to help students with disabilities and English as a Second Language (ESL) students participate more fully in assessments and better demonstrate their knowledge and skills.

MSPAP permits five categories of accommodations (scheduling, setting, equipment, presentation, and response) with 31 accommodations under the five categories for students with Individualized Education Programs (IEPs) and ESL students. Most accommodations do not invalidate student scores; however, in some cases, the student will not receive a score if the validity of the work that has been accommodated has been compromised. For example, if an examiner must read sections of the test to a student, the reading construct has been comprised. The student is not reading but listening; therefore, the student will not receive a reading score for the test. The student will, however, receive scores in all other content areas.

Students with disabilities may be exempted from MSPAP if they are not pursuing the Maryland Learning Outcomes, but instead, are pursuing alternative or life skill outcomes. ESL students may be exempted if they do not have the minimum language proficiency required for participation in MSPAP. ESL exemptions are limited to one test administration, i.e., a student exempted in grade 3 cannot be exempted again in grade 5.

Students may be excused from testing for a variety of reasons, such as demonstrating inordinate frustration, distress, or disruption of others and/or require accommodations that the school is unable to provide.

Students who are exempted do not take the test and are not included in the calculation of MSPAP scores for a school. Students who are excused do not take the test, but are included in the calculation of MSPAP scores. In other words, the school is not held responsible for students who are exempted from the test; it is held responsible for students who are excused from the test.

Scaling and Equating

Scaling and equating the MSPAP consists of two major phases. In Phase I, item calibrations are conducted to obtain the item parameters for each cluster. Misfitting items are identified and removed from the scale. Cluster equating is conducted to adjust the differences in difficulty among test forms. In Phase II, the results of two studies were used to link students' performance on the 1999 scale to the 1998 score scale. The first, *Rater Year Effects Study*, was designed to determine differences between raters who scored the 1998 MSPAP and raters who scored the 1999 MSPAP. The second, *Year to Year Equating Study*, was designed to equate the scores of two samples of students who were administered the 1998 and 1999 MSPAP in 1999.

The results of the two studies were combined to produce values that could be used to transform students' 1999 MSPAP scale scores to the 1998 score scale. This transformation permits comparisons to be made between the performance of students administered the MSPAP in 1998 and 1999. Since the current year MSPAP score scale is always linked to the score scale of the previous year, comparisons can be made between the performance of students across years.

Item Set Calibrations

As in previous years, 1999 MSPAP items were calibrated separately by cluster. The calibrations for each cluster was based on stratified random samples drawn from the pool of students in the state who were administered the cluster. The strata consisted of the 24 Maryland LEAs. Within each grade, students were sampled such that their proportional representation in the calibration sample corresponded to their LEA's proportional representation in the state. Separate samples were drawn for each set of items to be calibrated.

Table 8 shows that item calibrations, or item scalings, were carried out for reading, writing, language usage, mathematics, science, and social studies. No items were deleted due to special issues or at the request of MSDE prior to the initial scaling.

The Two-Parameter Partial Credit model (CTB McGraw-Hill, 1992, p. 4-4), as implemented by the PC based program PARDUX (Burket, 1992), was used for scaling the responses to the 1999 MSPAP items. Trait estimates, as well as standard errors of measurement for these estimates, were developed using the same procedures that were

used in previous test editions. For two items assessing writing content, PARDUX could not provide parameter estimates. These items typically had difficulties that were extreme and different from the other items in the scale. For each of these items, plots of students' observed performance were used to fit tracelines "by hand." That is, the graphical display capability of PARDUX was used to examine observed item tracelines. Item parameters that produced tracelines that most accurately represented the observed data then were identified interactively.

The same two types of model fit analyses used to evaluate MSPAP items in the past were used again in 1999. The two types of analyses used an analogue to Yen's Q_1 (Yen, 1981) fit statistic and an analogue of Yen's Q_3 dependency statistic (Yen, 1984). The Q_1 statistic was used to compare observed and expected tracelines statistically. Also, graphical representations of these lines were examined. The Q_3 statistic was used to examine local dependence. Even though local dependence is still examined, it is important to remember that there have been no testlets of dependent items constructed since 1992.

Items with differences between students' observed and expected performance that exceeded criterion values were flagged for further study. These criterion values are described in detail in the Technical Report for the 1991 MSPAP. The items that exceeded the criterion values used for the 1999 MSPAP are given in Table 8.

There are limitations to the usefulness of fit statistics such as Q_1 . First, chi-square measures such as Q_1 are greatly influenced by the deviation of observations from very small expectations; this influence results in high chi-square values for deviations of no practical significance. Another limitation is that performance on an item is implicitly included in the model via the trait estimate. With shorter tests, such as writing, there is substantial part-whole contamination in comparing item observed performance with predictions that implicitly include that item via that trait estimate. Lastly, the Q_1 statistic criterion is very conservative; it often flags items that in fact fit really well. Due to these limitations, the Q_1 statistic was used as a flag for potential misfit. The fit of each flagged item was then further evaluated using detailed fit information and both graphically within PARDUX.

If very large differences between students' observed and expected performance occurred on an item, the item was judged to have poor fit and was deleted. Table 8 shows that in 1999 no items were deleted due to poor fit.

When reading for literacy experience is measured, students in cluster 3A, 5B, and 8C were allowed to select one from three or four passages. When writing for personal expression was measured, students in 3A, 5B, and 8C were allowed to choose the topic they wanted to write about and the form of writing they wanted to use. Table 9 details the calibration information for the reading and writing choice clusters. The writing choices of poem and play were not widely selected by students. The fit of each flagged item was then further

evaluated using detailed fit information within PARDUX. Table 9 shows that no items were deleted due to poor fit.

 See Table 8 and 9

Equating Studies

To adjust for differences in difficulty among test forms, MSPAP is equated horizontally. Equivalent scores are established on test forms in a grade (e.g., Cluster 3A, 3B, 3C), but not across grades (e.g., grades 3 and 5). Therefore, MSPAP scores can be compared within a grade, but not between grades.

To equate horizontally, equivalent group design (administering tests to be equated to groups of examinees equivalent in terms of the skill measured by the tests) is used. In MSPAP, equivalent design is implemented by randomly assigning students to test groups by Local Education Agencies (LEAs). Each test group of randomly assigned students for a given grade in a school is administered one of three test clusters.

Rater-year effect equating is conducted to determine and adjust for rater or scorer variance from one year to the next. In 1999, raters rescored approximately 1,500 *Student Answer Books* per grade from the 1998 MSPAP administration.

To adjust for differences in difficulty from year to year, a test form from the previous year's edition is administered. For the 1999 annual equating, 2,500 students per grade were selected to take a 1998 cluster. In each school system, one or more schools were randomly selected; in each school, a test group of randomly assigned students was selected. In each school system, the number of schools chosen for equating was proportional to the system's representation in the state as a whole. Because a minimum of three test groups in each grade take MSPAP, only schools with more than three test groups in a grade were selected for equating.

The next step in the equating study was to identify a group of students in each grade who took the 1999 MSPAP and who were equivalent to the 1999 group of students administered the 1998 MSPAP cluster. Following MSPAP administration, CTB counted the number of valid students from each LEA who took the 1998 MSPAP for the equating study and randomly sampled from the equating schools in the LEA the same number of students who took the 1999 MSPAP. This procedure ensured that the numbers of students from each LEA were identical in the two groups used for the equating.

The critical assumption that must be met to use the equivalent group design is that the groups taking the tests must be equivalent, not representative. CTB proportionally

samples from all LEAs to construct equating groups to avoid the appearance that one LEA or another exerts any undue influence on the equating results.

Analysis procedure

The equating process involves constructing an equation that permits the translation of scores obtained on one test to correspond to scores on a second test. It was the responsibility of CTB to express the 1999 obtained MSPAP scores on the 1992 score scale so that performance in the test years are comparable.

The method used derives a linear equation that can be used to adjust the scores on one test so that they correspond to the scores given for comparable performance on the target test. In the case of cluster equating, this target test was the 1999 cluster that had the most regular cumulative score distribution. In the case of the 1998-1999 equating, this target was the 1998 clusters administered in 1999 for the equating study.

When tests are scaled using item response theory, it is necessary that linear equating be done. Traditionally, linear equating based on equivalent groups has involved merely equating means and standard deviations. However, considering only means and standard deviations can produce unsatisfactory equating for tests such as MSPAP that have few items or unusual score distributions. Therefore, for equating MSPAP a procedure was used that was more detailed and robust than equating means and standard deviations. This procedure, the linear equipercentile procedure, determined the linear transformation that most closely aligned the greatest number of score points possible.

The linear equipercentile procedure had several steps. First, pairs of scores on the two tests that had the same percentile rank were identified. Then, the linear function that most accurately described this equipercentile result was determined. For the vast majority of tests, the score pairs fell on a straight line; therefore, the linear function ran through all the pairs.

As in previous years, the operating principle for equating was "the greatest accuracy for the greatest number." In other words, the equating line was located so that it passed through as many scores as possible. It was also located with attention on the Proficiency Level 3/4 cut score.

Equating the Content Area Scores Across Clusters

The procedures used to equate content area scores are comparable to those used to equate content area scores of previous MSPAP forms. Specifically, cumulative scale score distributions for the calibration samples were obtained for each grade and content area. In each grade, the content area scores of one cluster were designated as the target distribution. FLUX was used to carry out an equipercentile equating procedure to align

distributions of content area scores from each of the two other clusters so that they matched the target distribution as closely as possible. A linear transformation that produced the closest alignment between the target and a non-target score distribution was identified and used to adjust the non-target scores to the score scale.

Table 10 specifies the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS) for each content area and cluster. Note that the LOSSes and HOSSes are the same for the three clusters used to assess a given content area in a grade.

 See Table 10

Table 10 also indicates the percentage of students in the calibration samples at the LOSS and the HOSS, which is a useful measure of floor and ceiling effects. The table shows that there are substantial floor effects in writing and language usage. These tests are uniformly difficult and short, and many students in the calibration samples received scale scores at the LOSS.

Rater Year Effects Study

For this study, the responses of approximately 1,500 randomly selected students who had taken the 1998 MSPAP (Clusters 3C, 5A, or 8A) were re-scored by raters who scored the 1999 MSPAP. The 1999 raters were trained, using Scoring Guides developed for the 1998 MSPAP, by Measurement Incorporated (MI), the hand-scoring contractor for the MSPAP.

Analyses of the rater effects were conducted separately by scale within Grades 3, 5, and 8. To determine the magnitude of the rater effect for each scale, the 1998 item parameters were used to generate 1998 scale scores for the students in the study. The first set of scale scores ($98SS_{98}$) was based upon the ratings that the students received when they were tested in 1998. The second set of scale scores ($98SS_{99}$) was based on the ratings that these students received when they were re-scored by the 1999 raters. Both sets of scale scores were expressed on the 1998 score scale.

Linear equipercentile equating procedures, as implemented in the computer software program FLUX (Burket, 1992), were used to align the $98SS_{98}$ s with the $98SS_{99}$ s. The linear transformation that best expressed the adjustment to the $98SS_{98}$ s was used to define the magnitude of the rater effect for each scale assessed in each of the three grades.

Results

Table 11 shows the mean 1998 scale scores ($98SS_{98}$) for the samples used in the Rater Effects Study and the mean scale scores for the state reported in the 1998 Forms Effects Study for Clusters 3C, 5A, and 8A. The table shows that for all three grades, the samples tended to have slightly higher scale scores than did the population of students who were administered this cluster. Overall, the differences were typically less than one tenth of a standard deviation.

The average raw scores obtained in 1998 and the values obtained when they were re-scored in 1999 are given and compared in Table 12. Positive values, given in the last column of the table, indicate that the 1999 raters graded the students more leniently than did the 1998 raters; that is, they gave the students higher scores on the average. Negative values, in this column, indicate that the 1999 raters graded the students more severely than did the 1998 raters; that is, they gave the students lower scores on the average.

A comparison between the mean differences reported in the current study and those reported for 1992 through 1999 MSPAPs are given in Table 13 in terms of standardized mean differences. Positive differences indicate that the raters who scored in the year that the study was done were more lenient than the raters who scored in the previous test year. Negative differences mean that the raters who scored in the year that the study was done were more severe than the raters who scored in the previous test year.

Table 13 shows that in terms of raw scores the rater effects generally were quite small in 1999. The 1999 results indicate small differences between the 1998 and 1999 rater groups. The 1999 results also indicate that the 1998 and 1999 raters were not consistently more lenient or severe relative to previous study years.

Table 13 shows that, in terms of raw scores, the rater effects generally were quite small in 1999, ranging from zero- to one-tenths of a standardized mean difference in either direction for all content areas in the three grades. The exceptions are Writing in Grade 5 and Language Usage and Social Studies in Grade 8. The 1999 results indicate small differences between the 1998 and 1999 rater groups. The 1999 results also indicate that the 1998 and 1999 raters were not consistently more lenient or severe relative to previous study years. The values of the multiplicative (R_1) and additive (R_2) components of the transformations that best aligned the $98SS_{99S}$ with the $98SS_{98S}$ are given in the first two columns of Table 14. When applied to the 1998 parameters, these values adjust the 1998 parameter values for the 1999 rater effects. To illustrate the magnitude of the adjustment, the transformation values were applied to a scale score of 500. The value of 500 was chosen because the average 1998 scale score was near 500. Since the values given in Table 14 are expressed in terms of the scale score metric, they will resemble but not mirror the raw score results given in Table 12, since raw scores and scale scores have a non-linear relationship.

See Tables 11-14

Equating 1998 and 1999 Scale Scores

Method

For this equating study, equivalent groups of students who were administered the 1998 and 1999 MSPAP were required, since no anchor items were available to link the tests administered in the two years. Accordingly, in 1999 approximately 2,500 third grade, fifth grade, and eighth grade students were selected to take 1998 MSPAP test books in May, 1999, while their counterparts were administered the 1999 MSPAP. The third grade students took Cluster 3C from the 1998 MSPAP; the fifth grade students took Cluster 5A; and the eighth grade students took Cluster 8A. These are the same books as those that were used for the Rater Effects Study just described.

The test groups in each grade were selected using stratified random selection procedures. Following a priori decisions to involve in the study no more than one test group per school and to use only Maryland schools with more than three classrooms, schools within each LEA were randomly selected to provide test groups for the Equating Study. Schools were selected separately for Grades 3, 5, and 8. The number of schools selected within each LEA was proportional to the representation of the LEA in the state. Within each school selected to contribute a test group in a given grade, the test group was randomly selected. Since all eligible students in a grade were randomly assigned to test groups, this test group was representative of the students in the school in the grade of interest.

Students' responses to the 1998 test books were scored by the same 1999 raters who were trained to score the 1998 books for the Rater Year Effects Study. For each scale, the students were screened to ensure that they had ratings for all the items used to assess that scale in the cluster of interest. Only those students meeting the screening criteria were used in the analyses for a given scale.

To develop equivalent groups for the administration of the 1999 test, it was decided a priori to select students who had been administered the clusters used as targets in the 1999 cluster equating. The target clusters typically had the most items, therefore the most reliable measurement. The target clusters also typically had smooth score distributions and items with good fit. The target clusters for the cluster equating can be found in Table 10.

The equivalent groups administered the 1999 target clusters in each grade were developed separately for each scale within the grade. To do this, the number of 1999 students

selected from each LEA for the analyses was the same as the number of students from that LEA who took the 1998 test books for the Equating Study and had valid scores on the scale. For example, if in the Equating Study 24 students from LEA #1 took 1998 Cluster 3C and had valid reading scores. To develop an equivalent group to use for the equating of the 1998 and 1999 Reading scales, 24 students from the same LEA who had valid scores on the 1999 target cluster would be randomly selected.

See Table 15

Analyses

The students in the Equating Study who took the 1998 test books were scored using the 1998 item parameters estimated for the items in these books. The use of these parameters ensured that these students' scale scores would be expressed in terms of 1998 scale scores; since these students' responses were scored by 1998 raters, it is useful to designate these scale scores as 98SS₉₉. The students who took the 1999 test books were scored using the 1999 item parameters estimated for the items in these books, so that these students' scores were expressed in terms of 1999 scale scores. Since these students' responses were scored by the 1999 raters, their scale scores can be designated 99SS₉₉. In the equating analyses, the lowest and highest obtainable scale scores from the 1998 MSPAP were used. This was done so that the scale scores for all students would not have scores that fell beyond the range of scale scores obtainable in 1998.

Equating procedures implemented by FLUX (Burket, 1992) were used to align the 99SS_{99S} with the 98SS_{99S}. The linear transformation that best aligned the 99SS_{99S} with the 98SS_{99S} was used to express the 99SS_{99S} on the 1998 scale.

Results

It is important to emphasize that the equivalence of the two samples used in the equating is critical for the soundness of the equating. The only data available to measure the equivalence of these samples were the distributions of students across LEAs, which indicated that the equating groups matched exactly in terms of the number of students taken from each LEA.

In the paragraphs that follow, comparisons are made between the test performance of the equating samples administered the 1998 books and the state as a whole in 1998. These comparisons are useful for the purposes of documentation and general information.

Table 15 describes the sample of students' 98SS_{99S} and compares these scores to state means estimated for 1998. In examining this table, it is important to keep in mind that the 98SS₉₉ reflect performance on 1998 items evaluated by 1999 raters, adjusted for the differences between the 1998 and 1999 raters. In other words, these statistics reflect the scores that would have been obtained had 1998 raters been used. The table shows that the scale scores are relatively similar across the grades when the State and the sample results are compared. Inspection of the case counts by LEA in each grade reveal that the proportions of students from each LEA were quite similar to the proportion of students that the LEA represents in the state.

The values of the multiplicative (T_1) and additive (T_2) components of the transformations that best aligned the 99SS_{99S} with the 98SS_{99S} are given in the first two columns of Table 16. In addition, the result of applying these transformation values to a scale score of 500 is shown in the third and fourth columns of the table to provide a sense of the size and direction of the test effect. Positive values in the fourth column of the table indicate that a scale score of 500 obtained on the 1999 MSPAP was transformed to a score greater than 500 on the 1998 scale. Negative values indicate that a scale score of 500 obtained on the 1999 MSPAP was transformed to a score less than 500 on the 1998 scale.

 See Tables 15-16

Comparison of 1998 and 1999 Mean Scores

Table 17 provides data permitting comparisons between the MSPAP performance of the students in 1998 and 1999 on the average. Both the 1998 and 1999 results reflect the average scale scores obtained by the student populations in three grades.

Caution must be exercised when interpreting the differences observed in Table 17. This is especially true for the writing results since they were short tests and had large standard errors. All differences observed in the last column of Table 17 are too small to allow an interpretation of the trend of the performance of the Maryland students by themselves. However, consistently higher scores for the 1999 students suggest some degree of growth occurred in each grade for several content areas.

When considering these results, it is important to remember that many different statistics can be used to describe student performance. Average scores are a convenient statistic, but when distributions are as skewed as many are for the MSPAP, the median may be a better indicator of typical test performance. The reports produced by the state of Maryland summarize performance in terms of Proficiency Standards; these bands

constitute another set of statistics by which performance can be described. The statistic used will affect the results one obtains and the conclusions one draws about growth or declines in performance over years. The average scores reported in Table 17 may not provide the same picture of student performance as that obtained when other statistics are used to describe this performance.

See Table 17

Review and Decision Points for the 1999 Equating. As an equating assurance check, review and decision points were examined for all clusters. MSDE, the National Psychometric Council, and CTB McGraw-Hill reviewed the cluster scaling and equating, rater year effect equating, annual equating, and performance results before each subsequent step of the process was undertaken. Through this process, the test characteristic curves and percentile rank correspondences were found to be acceptable for the 1999 MSPAP equating.

Reliability

Coefficient Alphas

Coefficient alpha is a reliability measure suitable when items have a variety of score levels (Allen & Yen, 1979). The coefficient alphas based on the calibration sample are reported in Table 18 by grade and cluster. Refer to Table 8 and 9 for the sample sizes and the number of items comprising each scale. The alpha coefficients for each grade and content area are generally high except for writing choice clusters. The coefficient alphas for each MSPAP test within each cluster are consistent with other constructed response tests (e.g., see KIRIS Accountability Cycle Technical Manual, 1998).

The writing test is comprised of three items spanning at least two different writing purposes, unlike mathematics, which usually has more than 30 items per cluster. The coefficient alphas obtained in the MSPAP writing assessment are typical of short tests. The MSPAP writing results are similar to the coefficient alphas obtained on the Maryland Writing Test (MWT), a performance assessment comprised of two items. The coefficient alphas for the MWT range from 0.50 to 0.55. Therefore, the reliabilities for the writing portion of the MSPAP are considered acceptable as well.

See Table 18

Standard Errors of Measurement for Proficiency Level Cut Scores

The standard error of measurement (SEM) is displayed in Tables 19 to 21. These SEMs are for individual scores in each content area. No test provides an exact point estimate. Instead, all scores have some degree of error. The SEM, produced through the Two-Parameter Partial Credit model, is influenced by the amount of information provided by each item and the number of items contributing to a content area. In this way, it is similar to the coefficient alpha. As can be noted from the tables, SEMs are usually smaller in the middle of the scale distribution (i.e., Proficiency Level 3/4 cut) and larger at the ends (i.e., HOSSes and LOSSes). Because the SEM is a function of item and test information, higher standard errors of measurement are not surprising in writing and language usage, which are all short tests of three to nine items.

See Tables 19 to 21

Validity

Validity evidence refers to the accuracy with which the test appears to measure what it is supposed to measure. MSPAP validity evidence is collected to support and validate intended interpretations and uses of scores from the assessment. Additionally, it is important that MSPAP assess the skills and knowledge that are documented in the Maryland Learning Outcomes document. The validity evidence described below is organized around these goals.

Between Content Area Correlations

Correlations were calculated to examine the relationships between the content area scale scores at each grade level. The relationships can be described as moderate to strong (see Tables 22 through 24). These findings are similar to the moderate to strong correlations found among MSPAP content area scale scores, CTBS/4, and teacher ratings calculated in a special study of the 1991 MSPAP test edition (see CTB McGraw Hill, 1992, Tables 9-8 through 9-10).

Correlations were also calculated to examine the relationships between the content area scale scores at each school (Tables 25 through 27). The relationships can be described as strong.

See Tables 22 to 27

Test Difficulty Concerns

MSPAP was developed with standards for the year 2000. The test was built around what students should be learning. Two impacts of test difficulty are (1) the test information function does not overlap well with student scores, and (2) higher standard errors are found at the lower and upper regions of the distribution. Since 1992, the fit between the test and student achievement has been improving.

Content Validity Evidence

Content validity evidence refers to the degree to which an assessment reflects the content it was designed to assess. The Maryland Learning Outcomes, the basis for learning, instruction, and MSPAP assessment activities, are based on national curriculum standards and learning theories. For example, the reading outcomes are similar to the NAEP reading assessment objectives and based on the reader response theory. Similarly, the writing outcomes are based on long-recognized modes of discourse, and the mathematics outcomes are based on the National Council of Teachers for Mathematics (NCTM) standards for curriculum and evaluation. The science outcomes are based on Project 2061 by the American Association for the Advancement of Science (AAAS). The social studies outcomes are underpinned by the work of groups including the Association of American Geographers, the Commission on History in the Schools, and the Joint Council on Economic Education. Moreover, the assessment tasks are developed by content area and grade specialists, specifically teachers. Each task development team is given specifications on which outcomes to assess in their task. After tasks are completed, they are reviewed.

In conclusion, the MSPAP has evidence of substantive content validity. It is a performance-based assessment that uses authentic and real-life situations as assessment tasks. In addition, reading selections are full-length published works rather than excerpts contrived for use in a test. Furthermore, the test is administered to random groups of students who work in small groups that reflect authentic situations. MSDE content chairs assign tasks to be written for a group of outcomes.

A high degree of match between assessment activities and the outcomes they assess is ensured through multiple reviews during the development of tasks, scoring tools, and scoring guides. All MSPAP tasks have been reviewed by the writers, hand-scoring teams, test administration teams, and are field-tested. These reviews allow for the opportunity to confirm that the specified outcomes, as defined by the Maryland Learning Outcomes document, are being assessed.

Outcomes Coverage

Coverage of outcomes by assessment activities is proportionally balanced according to the relative importance of the outcomes at different grade levels. A high degree of match between assessment activities and the outcomes they assess is ensured through multiple reviews during task development and development of scoring tools and guides. All of these reviews allow for the opportunity to confirm that the specified outcomes are indeed being measured as defined by the Learning Outcomes document. Appendix B presents the Maryland Learning Outcomes and the number of items measuring each outcome by grade and cluster for 1999 MSPAP.

See Appendix B

Construct Validity

Construct validity is considered to be the unifying concept for all views and types of evidence of test score validity (see, for example, Messick, 1989, p. 13). One way to assess the construct validity of MSPAP is to compare its results with similar tests. Since MSPAP reflects the NCTM standards and the reader-response model of reading, MSPAP results can be compared to Maryland's National Assessment of Educational Progress (NAEP) results.

Maryland's fourth grade NAEP reading performance showed 29% of the students achieving at or above the "proficient" level on the 1998 NAEP Reading State Assessment. On the 1998 MSPAP, 41.6% of the state's third graders and 40.4% of the fifth graders scored at the satisfactory level or above in reading. (On the 1999 MSPAP, 41.2% of the state's third graders and 41.4% of the state's fifth graders scored at the satisfactory level or above in reading.)

In mathematics, 22% of Maryland's fourth graders performed at or above the "proficient" level in the 1996 NAEP assessment. On the 1996 MSPAP, 38.7% of the state's third graders and 47.8% of the state's fifth graders scored at the satisfactory level or above in mathematics. (On the 1999 MSPAP, 38.9% of the state's third graders and 46.2% of the fifth graders scored at the satisfactory level or above in mathematics.)

Statistical Test Bias

As a technical term, 'test bias' is not easily defined. A reasonable conceptual approach is to consider a test biased if students of the same degree of proficiency receive reliably

different scores on the test (Camilli & Shepard, 1994). A test that fits this definition would then be biased in favor of those who receive the higher scores and against those who receive the lower scores. The difficulty is that, in practice, there is no method available to determine whether or not two different students have the same degree of attainment.

In order to overcome the lack of a 'pure' measure of attainment, overall scores on the test are commonly used as the best available measure to evaluate 'bias' at the item level. This approach relies on the assumption that bias, if it exists, is presented in some, as opposed to all, the items on the test. Therefore, to the degree that items are identified as biased, it may be true that the test is biased. However, if no items are identified as biased, then it is a reasonable conclusion that test bias is not a threat to test validity.

Differential item functioning (DIF) procedures examine the possibility that non-essential item characteristics may result in misleading poor performance for minority, female, or other defined groups of students. Although the terms item bias and DIF are used interchangeably, DIF does not necessarily imply unfairness. Evidence of DIF is usually considered as a signal to test developers to examine an item more closely to consider whether or not it is defective before using it again.

Items that are biased against groups of students who take the MSPAP or items that function differently for different student groups diminish construct validity. A measure of DIF generalized from the Linn-Harnisch procedure (1981) is used to flag differentially functioning items. MSDE has studied items flagged for DIF to inform subsequent assessment task development. MSDE examines performance of African-Americans, Asians, and Hispanics in comparison to Caucasians, and examines the performance of females in comparison with males.

In the generalized Linn-Harnisch procedure, the parameters for each item and the student scale score estimates are estimated using the two-parameter partial credit model (Burke, 1991). Students from ethnic and gender groups are divided into ten scale score categories. Within each category, the expected proportion of students getting the item correct (based on the IRT model using all students) is compared to the observed proportion of students from that ethnic or gender group who got the item correct. Items are flagged when the discrepancy between the expected and observed proportions is large and occurred at multiple score levels. The Linn-Harnish procedure is designed to flag both uniform and non-uniform DIF. Items with uniform DIF showed DIF at all score categories where items with non-uniform DIF showed DIF at some but not all score categories. The computational details for Linn-Harnish procedures are summarized below.

During item calibration, the item parameters estimated for the items assessing a given subject area are used to score **all** of the examinees in the calibration sample. The examinees for each target group (e.g., African American) are then sorted into ten equally

numerous score categories (deciles). For each item, using the mean attainment estimate for the examinees of the target group in each decile, the predicted and observed examinee success rates are calculated and compared separately in each decile. A positive difference between the observed and predicted values indicates that the target group members in that decile did better than expected. The positive differences are summed to obtain a positive difference value, D+. Similarly, a negative difference indicates that the target group members in that decile did less well than was expected. The negative differences are also summed to obtain a negative difference value, D-. These two sums of differences are summed to obtain an overall difference, D.

DIF was defined in terms of overall differences in performance. Items for which absolute value of D was greater or equal to 0.10 were flagged as exhibiting DIF or biased. That is, D smaller than -0.10 indicates that the item is against the target subgroup and D greater than 0.10 indicates that the item is in favor of the target subgroup. Table 28 presents the number of items for MSPAP 1999 being flagged as exhibiting DIF using the criterion described above. It can be seen that very few items were flagged for bias either in favor or against African American target groups. While present, the small numbers of flagged items in the Asian, Hispanic, and female groups may be the result of statistical imprecision due to the relative small sizes of these groups in Maryland.

 See Table 28

Consequential Validity Evidence

MSDE, in conjunction with the University of Pittsburgh, is conducting a study to examine the impact of MSPAP on curriculum, instructional and assessment practices, student performance, staff development, and school-based decision-making. It will also examine how the impact varies by content area (reading, writing, language usage, mathematics, science, and social studies), school characteristics (percent minority students, percent free or reduced lunch, MSPAP performance), and grade level (3, 5, 8 and off-grades 2, 4, 7).

Evidence is being collected at system, school, and classroom levels via questionnaires, interviews, and reviews of curriculum, assessment, and professional development materials.

Conclusion

MSPAP scores, in combination with other performance measure, are used to determine school performance consequences such as state mandated intervention in schools failing to demonstrate progress and rewards for schools consistently making significant improvement.

Validity evidence and other technical information provide reasonably strong assurance that MSPAP scores can be appropriately used for evaluating school performance and guiding school improvement.

Score Interpretation

Two types of scores are available and relevant to school performance and for use in school improvement planning: scale scores and outcome scores. These two types of MSPAP scores are discussed below. For more detailed discussions about score interpretation of MSPAP, consult “Score Interpretation Guide” (MSDE, 1998).

Scale Scores

MSPAP was designed to produce scale scores for the content areas of reading, writing, language usage, mathematics, science, and social studies. MSPAP scale scores indicate a school's level of performance in each content area. MSPAP scale scores range, in general, between 350 and 700 with a mean of approximately 500 and a standard deviation of approximately 50. Scale scores from the same grade level and content area have the same meaning and are directly comparable from year to year. Scale scores are not comparable across grade levels or content areas because of differences in test content and difficulty.

MSPAP scale scores, like other test scale scores, have little intrinsic meaning other than higher scale scores represent higher performance in a content area. Interpretation of the scale scores is aided by proficiency level descriptions. Proficiency level descriptions were developed to help bring meaning to scale scores and to guide interpretation for school performance and improvement.

Proficiency Level Descriptions

Proficiency levels. Proficiency levels and descriptions are intended to inform and guide interpretation of MSPAP scale scores. They describe what students at a particular level generally know and can do in relation to the Maryland Learning Outcomes. The descriptions generally apply to all students at each level rather than to specific students within a level. Individual students whose scale score locates them at a particular proficiency level may or may not be able to demonstrate all of the knowledge, skills, and processes contained in that proficiency level description.

Listed in Appendix C are the scale score ranges for each proficiency level in each content area and grade. Detailed proficiency descriptions for each content area and grade appear in Appendix B of the Score Interpretation Guide (MSDE, 1998).

As Appendix C indicates, each proficiency level represents a range of performances and of scale scores. For example, grade 3 reading scale scores lower than 490 indicate Level 5 proficiency, those between 490 and 529 indicate Level 4 proficiency, those between 530 and 579 indicate Level 3 proficiency, and so forth.

MSPAP emphasizes high standards of performance. Since MSPAP scale scores can range as low as 350, there is a wide range of scores in Level 5. Generally speaking, students at Level 5 do not consistently demonstrate Level 4 proficiency. However, they may have provided some responses to assessment activities that, with increased consistency, would have placed them at Level 4.

Proficiency level descriptions and proficiency cut scores were established by committees, which consist of teachers, principals, content area supervisors, and assistant superintendents. The committees matched MSPAP items to proficiency level descriptions between Levels 1-5 and used the resulting item classifications to establish the location of the cut scores between proficiency levels.

Developing the Descriptions. The committee that established the proficiency level cut scores also developed descriptions for each level. For both the establishment and refinement of the descriptions, the committee examined each assessment activity at a proficiency level, the accompanying scoring criteria for each activity, and student responses to each activity. They used professional judgment to determine and list the knowledge, skills, and processes each activity required of students and to synthesize the lists of required knowledge, skills, and processes into descriptions, in Maryland learning outcomes terms, of what students at each proficiency level know and can do.

Interpretation and use of the proficiency levels and proficiency level descriptions. Proficiency level descriptions apply generally to any group of students, based on performances by all students and schools in Maryland. The descriptions are not customized specifically for individual students, single schools, or other groups.

School Performance Standards

A cornerstone of the Maryland School Performance Program (MSPP) is the process of setting standards of satisfactory and excellent performance levels for schools to meet by 2000.

Development of the standards for MSPAP followed the same procedures used in establishing the school performance standards for all areas reported in the annual *Maryland School Performance Report*. A state Standards Committee researched

information on standard setting, identified criteria for standards, and defined the terms *satisfactory* and *excellent*.

Satisfactory performance denotes a level of performance that is realistic and rigorous for schools, school systems, and the state. It is an acceptable level of performance on a given variable, indicating proficiency in meeting the needs of students.

Excellent performance denotes a level of performance that is highly challenging and clearly exemplary for schools, school systems, and the state. It is a distinguished level of performance on a given variable, indicating outstanding accomplishment in meeting the needs of students (Thorn, Moody, McTighe, Kelly, & Peiffer, 1990, page 7).

Two groups participated in the standards setting process:

A 20 member Standards Committee of teachers, administrators, content area and assessment specialists, parents, students, university professors, and

A 17 member Standards Council of representatives of local boards of education, teacher's unions, businesses, students, and the Maryland General Assembly.

The process of setting standards included several steps. Initially, the Standards Committee recommended a proficiency level to describe satisfactory and excellent performance and the percentage range of students who should score at these levels (i.e., 60% to 80% at the satisfactory level). These recommendations were reviewed by the Standards Council, which refined this work to describe satisfactory and excellent performance by proficiency level and set a percent of students who should be in each category. These two steps depended on a group decision reached through a convergence process.

The recommendations from the Standards Council were reviewed by the State Board of Education and comments were given through public meetings. Following the public meetings, the standards were formally adopted by the State Board of Education.

The Standards Committee recommended level 3 as the proficiency level that describes satisfactory performance and levels 1 and 2 as the proficiency levels that describe excellent performance. Once the ranges for satisfactory and excellent school performance were established, the recommendations were forwarded to the Standards Council. They were asked to choose a single percentage for each standard for school performance. The

Council concurred with the definitions for satisfactory and excellent performance. In addition, the Council recommended 70% for satisfactory and 25% for excellent. For a given school to achieve satisfactory performance in a particular area/grade level, 70% of students must achieve satisfactory performance (level 3 and above). To achieve excellent performance, a school must meet the satisfactory requirement and 25% of these students must achieve excellent performance (level 2 and above). The State goal is that all schools will reach the satisfactory standards by the year 2000.

Interpretation and use of school performance standards for school improvement planning. The score reports produced by MSDE for each school system and school contain numbers and percentages of students at each proficiency level and at satisfactory and excellent standards. School and system staff use these percentages, along with the proficiency level descriptions, to evaluate their school's performance in relation to the Maryland Learning Outcomes. They also use this information to assess their school's progress in reaching standards.

Only those students tested are considered when determining a school's proficiency level, because of the focus on the strengths and weaknesses of the students in the school. Since the school performance standards focus on how well a school is performing on the outcomes, any student who should have been tested is included in the calculation¹. This includes students who were excused from the MSPAP test administration and students who were absent during the test administration. Therefore, proficiency level percentages may be higher than standard percentages, because the proficiency level percentages are usually based on a smaller number of students.

Individual Student Scale Scores

Scale scores and outcome scores for individual students are not interpretable because each student takes only one-third of the total test. Since the primary focus of MSPAP is school performance rather than individual performance, individual student scores are not to be used for decisions for individual student's performance.

Outcome Scores

Within each of the six content areas assessed on MSPAP, i.e., reading, there are more specific outcomes, i.e., reading to be informed. Outcome scores are based on subsets of items that comprise a content area scale. These scores are the scores that would be expected on an outcome if a student had taken all of the items which measure that outcome. For an outcome score to be reported, at least four measures of the outcome must be present in the test form that the student took. There are two types of outcome scores: Outcome Scores and Outcome Scale Scores.

Outcome Scores. MSPAP outcome scores range from 0 to 100% and are reported for each outcome assessed in each MSPAP content area. They are conceptually analogous to Maryland Functional Testing Program domain scores and can be interpreted like these scores². Outcome scores indicate the proportion of mastery of the knowledge, skills, processes and other requirements that comprise an outcome area. In other words, the MSPAP school outcome score is the average percentage of all score points available on that outcome that a school achieved across all test clusters administered in the school.

Outcome scores are not directly comparable across grades and content areas within a grade, nor are they directly comparable across years because of differences in content and test difficulty. However, they can be compared using information on the relative difficulty of each outcome. Moreover, outcome scores cannot be directly linked to MSPAP proficiency levels.

Interpretation and Use of the Outcome Scores. School improvement teams use profiles of a school's Outcome Scores in a content area along with other information about a school, to determine a school's instructional program's relative strengths and weaknesses in each MSPAP content area.

Content area relative difficulty values are reported on Table 29. Relative difficulty refers to the average proportion of the maximum possible score for an outcome across clusters. The relative outcome difficulty index ranges from 0 to 100%. Lower percentages indicate harder outcomes, and conversely, higher percentages indicate easier outcomes. This information is used in conjunction with outcome score averages. An index of relative difficulty was developed because of the desire to compare outcome score averages within each content area to one another.

 See Table 29

Outcome Scale Scores. Outcome scale scores are directly comparable across outcomes in the same content area, across years, and to the MSPAP proficiency levels. These scores are expressed on the MSPAP scale score scale and range, as are the content area scale scores, from 350 to 700. Therefore, they can be interpreted in relationship to the underlying score scale and proficiency levels.

MSPAP Score Reports

The four main types of MSPAP score reports are: Standards Reports, Proficiency Level and Participation Reports, Outcome Score Reports, and Outcome Scale Score Reports. MSDE provides these reports at the state, school system, and school levels.

MSPAP Standards Reports. These reports provide information on the percentages of students at satisfactory and excellent levels of performance and indicate whether the standards for satisfactory and excellent school performance have been met. Information on the numbers and percentages of students by grade, content area, race, and gender is available in the MSPAP Disaggregated Standards Report.

MSPAP Proficiency Level and Participation Reports. These reports provide the numbers and percentages of test takers at each of the five MSPAP proficiency levels. They also report numbers and percentages of students who completed assessment activities in each MSPAP content area and received a scale score. Also, numbers and percentages of students who were absent, excused, or exempted from the MSPAP test administration are reported.

MSPAP Outcome Score Reports. Outcome Score Reports contain the average outcome score, or percentage of mastery of an outcome, for a school, school system, or the state. The Outcome Score Reports also include percentages of students in four outcome score ranges: 0-25, 26-50, 51-75, and 76-100. This information is intended to provide a general idea of the percentage of students who have displayed little or no mastery of the knowledge, skills, and processes required in an outcome (i.e., those in the outcome score range 0-25) and the percentage who have displayed near complete mastery of the outcome (i.e., those in the range 76-100).

MSPAP Outcome Scale Score Reports. The Outcome Scale Score Reports contain the median outcome scale score for each learning outcome. The median (50th percentile), the interquartile range (25th to 75th percentiles) and the 5th to 95th. Outcome Scale Score Reports can be used to compare outcome performance within a content area. Unlike outcome scores, outcome scale scores can be compared in a content area because the outcome scale scores have been adjusted for difficulty.

It is important not to over-interpret the relationship between outcome scale scores and proficiency levels. Outcome scale scores represent performance on activities that measure only that outcome. In contrast, proficiency levels are established based on all the outcomes in a content area.

References

- Allen, M., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) New York: American Council on Education.
- Binkley M., Atash, M. N., & Bourque, M. (in press). Standard setting and reporting. In T. Husen and N. Postlethwaite (Eds.), *The International Encyclopedia of Education*, 2nd ed.
- Burket, G. R. (1991). *PARDUX, Version 1.4*. Monterey CA: CTB Macmillan/McGraw Hill.
- Burket, G. R. (1991). *FLUX Version 1.0*. Monterey, CA: CTB Macmillan/McGraw-Hill.
- CTB Macmillan/ McGraw Hill. (1992). *Final technical report: Maryland School Performance Assessment Program, 1991*. Available from Maryland State Department of Education, Division of Planning, Results and Information Management.
- Camilli, G. & Shepard, L. A., (1994). *Methods for identifying biased test items, Measurement Methods for the Social Sciences, Vol 4*, Sage Publications, CA: Thousand Oaks.
- Ebel, R. L. (1979). *Essentials of educational measurement*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall.
- Kenney, P, & Sliver, E. (1999). Content Analysis Project – State and NAEP Mathematics Assessment. *Report of Results from the Maryland MAEP Study*. Learning Research and Development Center, University of Pittsburg.
- Kentucky Department of Education. (1998). *KIRIS Accountability Cycle I Technical Manual*: Lexington: Author.
- Linn, R. L. & Harnisch, D. (1981). Interactions between item content and group membership in achievement test items. *Journal of Educational Measurement*, 18, 109-118.

- Mamillan/McGraw-Hill School Publishing Company (1993). *Reflecting Diversity-Multicultural guidelines for educational publishing professionals*. New York: Author.
- Maryland State Department of Education (1993a). *Scoring MSPAP: A Teacher's Guide*. Baltimore: Author.
- Maryland State Department of Education. (1993b). *Technical report: 1992 Maryland School Performance Assessment Program*. Baltimore: Author.
- Maryland State Department of Education. (1994). *Technical report: 1993 Maryland School Performance Assessment Program*. Baltimore: Author.
- Maryland State Department of Education. (1995). *Technical report: 1994 Maryland School Performance Assessment Program*. Baltimore: Author.
- Maryland State Department of Education. (1996). *Technical report: 1995 Maryland School Performance Assessment Program*. Baltimore: Author.
- Maryland State Department of Education. (1997). *Technical report: 1996 Maryland School Performance Assessment Program*. Baltimore: Author.
- Maryland State Department of Education. (1998). *Technical report: 1997 Maryland School Performance Assessment Program*. Baltimore: Author.
- Maryland State Department of Education. (1999). *Technical report: 1998 Maryland School Performance Assessment Program*. Baltimore: Author.
- Maryland State Department of Education. (1999). *Test administration and coordination manual, 1998*. Baltimore: Author.
- Maryland State Department of Education. (1996). *Score Interpretation Guide, Maryland School Performance Assessment Program - 1996 MSPAP and Beyond, 1996*. Baltimore: Author.
- Maryland State Department of Education. (1999). *Maryland School Performance Report, 1999*. Baltimore: Author.
- Measurement Incorporated. (1999). *1999 Maryland School Performance Assessment Program scoring report*. (Available from the Maryland State Department of Education, Baltimore, MD)
- National Evaluation System Inc. (1991). *Bias Issues in Test Development*.

- Amerst, MA: Author.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.)
New York: American Council on Education/ Macmillan.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14, 3-19.
- Thorn, P., Moody, M., McTighe, J., Kelly, N., & Peiffer, R. (1990, April). *Establishing standards for Maryland's School Systems: A systemic approach*. Available from Maryland State Department of Education, Division of Planning, Results and Information Management.
- Westat, Inc. (1999). *1999 MSPAP Field Test Report*. Available from Maryland State Department of Education, Division of Planning, Results and Information Management.
- Westat, Inc. (1994). *Establishing proficiency levels and descriptions for the 1994 MSPAP assessment program*. Available from Maryland State Department of Education, Division of Planning, Results and Information Management.
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245-262.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.

TABLE 1
NUMBERS OF TEAMS, READERS, AND SCORING LEADERS

Site	Grade/ Cluster	Number Of Teams	Target Number of Readers	Number Of Coordinators	Number Of Leaders
Western Tech	3B	4	68	4	4
	5B	4	83	4	4
	8B	4	83	4	4
Total		12	234	12	12
Chesapeake	3C	4	78	4	4
	5C	4	82	4	4
	8C	4	82	4	4
Total		12	242	12	12
Waldorf	3A	4	91	4	4
	8A	4	95	4	4
Total		8	186	8	8
Grasonville	5A	4	82	4	4
GRAND TOTAL		36	744	36	36

TABLE 2
READER ACCURACY SET MEAN SCORES BY TEAM - GRADE 3

<u>TEAM</u>	<u>SET 1</u>	<u>SET 2</u>	<u>SET 3</u>	<u>SET 4</u>	<u>SET 5</u>	<u>SET 6</u>	<u>AVERAGE</u>
A1	77	83	89	80	90	87	84
A2	85	94	88	84	94	86	89
A3	81	83	90	86	85	75	83
A4	77	73	84	87	73	74	78
B1	88	79	90	88	93	90	88
B2	95	87	90	87	90	90	90
B3	87	91	88	90	85	85	88
B4	82	78	90	69	86	84	82
C1	88	88	82	82	86	84	85
C2	83	85	83	88	90	85	87
C3	92	96	100	99	100	100	99
C4	83	91	84	88	95	99	90

TABLE 3
READER ACCURACY SET MEAN SCORES BY TEAM - GRADE 5

<u>TEAM</u>	<u>SET 1</u>	<u>SET 2</u>	<u>SET 3</u>	<u>SET 4</u>	<u>SET 5</u>	<u>SET 6</u>	<u>AVERAGE</u>
A1	83	77	86	91	83	88	85
A2	77	81	85	78	84	79	81
A3	73	78	84	80	77	--	78
A4	79	93	78	88	84	85	85
B1	82	82	79	83	76	83	81
B2	71	80	87	82	81	82	81
B3	78	87	88	83	84	83	84
B4	72	86	88	79	81	83	82
C1	86	84	93	90	86	88	88
C2	78	80	84	84	82	--	82
C3	91	90	93	94	92	--	92
C4	81	92	89	92	86	92	89

TABLE 4
READER ACCURACY SET MEAN SCORES BY TEAM - GRADE 8

<u>TEAM</u>	<u>SET 1</u>	<u>SET 2</u>	<u>SET 3</u>	<u>SET 4</u>	<u>SET 5</u>	<u>SET 6</u>	<u>AVERAGE</u>
A1	93	87	89	86	--	--	89
A2	90	89	84	83	85	78	85
A3	78	87	75	80	77	--	79
A4	83	83	88	93	88	88	87
B1	96	90	97	97	91	95	94
B2	80	81	91	84	82	79	83
B3	77	81	73	82	83	81	80
B4	84	79	87	91	92	90	87
C1	93	89	94	93	92	96	93
C2	83	84	86	91	81	85	85
C3	84	75	81	77	86	80	81
C4	93	76	91	88	91	79	86

TABLE 5
FREQUENCY OF ACCURACY SET MEAN SCORES BY GRADE

<u>Grade</u>	<u>Less than 70 Percent</u>	<u>70-79 Percent</u>	<u>80-89 Percent</u>	<u>90-100 Percent</u>
3	1 (1%)	8 (11%)	39 (54%)	24 (33%)
5	0 (0%)	16 (23%)	41 (59%)	12 (17%)
8	0 (0%)	12 (17%)	35 (51%)	22 (32%)
ALL GRADES	1 (0%)	36 (17%)	115 (55%)	58 (28%)

TABLE 6
READER ACCURACY SET MEAN SCORES BY CONTENT AREA - GRADE 8

<u>TEAM</u>	<u>SET 1</u>	<u>SET 2</u>	<u>SET 3</u>	<u>SET 4</u>	<u>SET 5</u>	<u>SET 6</u>	<u>AVERAGE</u>
Social Studies							
A1	93	87	89	86	--	--	89
B1	96	90	97	97	91	95	94
C1	93	89	94	93	92	96	93
Social Studies							
A2	90	89	84	83	85	78	85
B2	80	81	91	84	82	79	83
C2	83	84	86	91	81	85	85
Science							
A3	78	87	75	80	77	--	79
B3	77	81	73	82	83	81	80
C3	84	75	81	77	86	80	81
Writing							
A4	83	83	88	93	88	88	87
B4	84	79	87	91	92	90	87
C4	93	76	91	88	91	79	86

*Note: Content areas are somewhat integrated.

TABLE 7
FREQUENCY OF ACCURACY SET MEAN SCORES BY CONTENT AREA - GRADE 8

<u>Content Area</u>	<u>Less than 70 Percent</u>	<u>70 - 79 Percent</u>	<u>80 - 89 Percent</u>	<u>90 – 100 Percent</u>
Mathematics	0 (0%)	0 (0%)	4 (25%)	12 (75%)
Social Studies	0 (0%)	2 (11%)	13 (72%)	3 (17%)
Science	0 (0%)	7 (41%)	10 (59%)	0 (0%)
Writing	0 (0%)	3 (17%)	8 (44%)	7 (39%)
All Content Areas	0 (0%)	12 (17%)	35 (51%)	22 (32%)

TABLE 8. SUMMARY FINDINGS FROM CALIBRATIONS

Content/ Cluster	Sample Size	No. of Items ¹	No. Items Deleted ²		No. Items with Hand-Estimated Parameters	No. of Items with FIT>Criterion ³	No. of Students at Min./Max.
			Special Issues	Fit			
<u>Reading</u>							
3A*	7,500	24	0	0	0	5	100
3B	7,500	11	0	0	0	1	224
3C	7,500	9	0	0	0	1	180
5A	7,500	11	0	0	0	0	142
5B*	7,500	30	0	0	0	1	57
5C	7,500	12	0	0	0	4	126
8A	7,498	14	0	0	0	1	172
8B	7,498	12	0	0	0	1	269
8C*	7,498	30	0	0	0	4	69
<u>Writing/Language Usage</u>							
3A*	7,500	17	0	0	0	1	509
3B	7,503	11	0	0	0	3	902
3C	7,500	11	0	0	0	1	801
5A	7,500	11	0	0	0	1	540
5B*	7,500	18	0	0	0	5	265
5C	7,500	11	0	0	0	1	568
8A	7,498	11	0	0	0	3	591
8B	7,498	11	0	0	0	7	551
8C*	7,498	20	0	0	0	2	171
<u>Mathematics</u>							
3A	7,500	19	0	0	0	0	164
3B	7,500	17	0	0	0	1	97
3C	7,500	21	0	0	0	0	112
5A	7,500	24	0	0	0	2	68
5B	7,500	20	0	0	0	1	98
5C	7,500	31	0	0	0	1	34
8A	7,498	18	0	0	1	3	366
8B	7,498	21	0	0	0	0	401
8C	7,498	22	0	0	0	2	343

(table 8 continue)

Content/ Cluster	Sample Size	No. of Items ¹	No. Items Deleted ²		No. Items with Hand-Estimated Parameters	No. of Items with FIT>Criterion ³	No. of Students at Min./Max.
			Special Issues	Fit			
<u>Science</u>							
3A	7,500	17	0	0	0	0	221
3B	7,500	24	0	0	0	1	127
3C	7,500	18	0	0	0	0	85
5A	7,500	16	0	0	0	0	111
5B	7,500	23	0	0	0	1	95
5C	7,500	19	0	0	0	0	107
8A	7,498	25	0	0	0	3	157
8B	7,498	23	0	0	0	0	231
8C	7,498	19	0	0	0	0	319
<u>Social Studies</u>							
3A	7,500	16	0	0	0	2	203
3B	7,500	17	0	0	0	0	136
3C	7,500	19	0	0	0	0	177
5A	7,500	18	0	0	0	1	69
5B	7,500	15	0	0	0	0	84
5C	7,500	19	0	0	0	0	74
8A	7,498	20	0	0	0	1	231
8B	7,498	18	0	0	0	1	361
8C	7,498	20	0	0	0	0	227

¹ No. of items refers to the number of items defined as assessing each content area prior to scaling and before items were deleted for the reasons specified in the next column. For the Reading and Writing/Language Usage items in 3A, 5B, and 8C, the No. of items is the total number of items in all choice sets; students administered these clusters actually responded to fewer items than the total given.

² The reasons for the item deletion are designated as GA signifying group-administration; MSDE signifying a deletion requested by MSDE; and Fit signifying poor fit.

³ The cut-off Z values used for various N counts are as follows:

<u>N</u>	<u>Z ></u>	<u>N</u>	<u>Z ></u>	<u>N</u>	<u>Z ></u>
1,500	4	2,000	5	3,000	8
4,000	11	5,000	13	6,000	16
7,000	19				

* This is a choice cluster. Sample size, the numbers of items, and the number of misfitting items for this cluster varied over the choice sets.

TABLE 9. CALIBRATION FOR CLUSTERS WITH CHOICE SETS

Content	Cluster	Choice	Sample Size	Number of Items	Number of Items with Fit Exceeding Criterion ¹
Reading	3A	Non-choice	7,500	6	0
		Choice A	1,541	6	0
		Choice B	4,110	6	3
		Choice C	1,849	6	2
	5B	Non-choice	7,500	6	1
		Choice A	1,770	6	0
		Choice B	2,439	6	0
		Choice C	1,192	6	0
		Choice D	2,099	6	0
	8C	Non-choice	7,498	6	0
		Choice A	1,616	6	1
		Choice B	1,670	6	1
		Choice C	2,229	6	1
		Choice D	1,983	6	1
	Writing	3A	Non-choice	7,500	2
Story			5,609	1	0
Poem			1,530	1	0
Play			361	1	0
5B		Non-choice	7500	2	1
		Story	4649	1	1
		Poem	2425	1	1
		Play	426	1	0
8C		Non-choice	7498	2	0
		Story	3686	1	0
		Poem	2665	1	0
		Play	294	1	0
		Other	853	1	0

¹ See footnote of Table 8 for the fitting criterion

TABLE 10. CLUSTER EQUATING RESULTS

Content Area/ Cluster	LOSS	HOSS	% at LOSS	% at HOSS
<u>Reading</u>				
3A	400	650	5	1
3B*	400	650	7	0
3C	400	650	6	1
5A	375	675	3	0
5B	375	675	3	0
5C*	375	675	3	0
8A*	375	650	3	0
8B	375	650	3	1
8C	375	650	2	1
<u>Writing</u>				
3A	455	635	24	1
3B	455	635	23	1
3C*	455	635	34	2
5A	440	595	18	5
5B	440	595	14	8
5C*	440	595	24	8
8A*	425	625	11	6
8B	425	625	12	5
8C	425	625	12	7

(Table 10 Continue)

* : Target Cluster

Content Area/ Cluster	LOSS	HOSS	% at LOSS	% at HOSS
<u>Language Usage</u>				
3A	450	625	12	1
3B	450	625	13	0
3C*	450	625	13	1
5A*	425	625	14	3
5B	425	625	12	4
5C	425	625	14	2
8A*	425	625	12	3
8B	425	625	10	3
8C	425	625	9	2
<u>Mathematics</u>				
3A	375	650	5	0
3B	375	650	3	1
3C*	375	650	3	0
5A	400	650	6	0
5B	400	650	7	0
5C*	400	650	4	0
8A*	400	650	3	0
8B	400	650	4	0
8C	400	650	4	0

(Table 10 Continue)

* : Target Cluster

Content Area/ Cluster	LOSS	HOSS	% at LOSS	% at HOSS
<u>Social Studies</u>				
3A	400	625	7	0
3B	400	625	6	0
3C*	400	625	8	0
5A	400	625	4	0
5B	400	625	5	1
5C*	400	625	6	0
8A*	375	650	4	0
8B	375	650	5	1
8C	375	650	4	0
<u>Science</u>				
3A	375	650	5	0
3B	375	650	4	0
3C*	375	650	3	0
5A	375	650	3	0
5B	375	650	4	0
5C*	375	650	4	0
8A	375	650	8	0
8B	375	650	7	0
8C*	375	650	8	0

* : Target Cluster

TABLE 11
RATER YEAR EFFECTS STUDY PERFORMANCE (98SS₉₈) OF STATE SAMPLE ON 1998 MSPAP

Grade	Scale	State ¹			Sample		
		Mean	SD	N	Mean	SD	N
3	Reading	520.9	43.8	20,214	522.5	43.4	1,400
	Writing	524.4	48.0	20,568	526.2	46.9	1,400
	Language Usage	525.2	59.5	20,636	525.9	58.4	1,400
	Math Content	517.3	57.1	20,587	---	---	---
	Math Process	513.8	52.1	20,587	---	---	---
	MA ²	515.8	52.1	20,587	519.5	48.3	1,400
	Social Studies	509.7	49.2	20,718	511.6	48.2	1,400
	Science	510.2	56.4	20,181	511.3	56.1	1,400
	5	Reading	518.9	49.9	19,945	524.3	48.2
Writing		507.1	54.2	20,136	513.3	52.7	1,393
Language Usage		530.5	59.3	20,291	537.2	56.6	1,393
Math Content		519.1	57.9	20,321	---	---	---
Math Process		511.3	56.0	20,030	---	---	---
MA ²		515.6	55.1	20,030	519.4	54.8	1,393
Social Studies		517.7	56.7	20,282	523.1	53.9	1,393
Science		520.1	55.4	20,332	525.3	53.8	1,393
8		Reading	507.6	37.5	18,033	510.5	35.6
	Writing	503.3	56.8	18,469	509.6	55.8	1,406
	Language Usage	508.5	59.6	18,701	513.6	57.2	1,406
	Math Content	522.5	51.0	18,303	---	---	---
	Math Process	514.8	61.6	18,303	---	---	---
	MA ²	518.9	53.2	18,303	524.5	50.4	1,406
	Social Studies	516.7	53.7	18,552	522.9	49.7	1,406
	Science	527.2	52.2	17,882	531.0	49.3	1,406

¹ State performance results were drawn from the Forms Effect Study carried out for the 1998 MSPAP. The values reported refer to performance on Clusters 3C, 5A, and 8A.

² The State performance results on MA were from the Math Total; The Sample results were from the unified MA.

TABLE 12
RATER YEAR EFFECTS STUDY RAW SCORE COMPARISONS

Grade	Scale	N	Raters Used				Mean Diff. (99 - 98)
			1998		1999		
			Mean	SD	Mean	SD	
3	Reading	1400	13.84	5.44	13.55	5.40	-0.29
	Writing	1400	2.95	1.98	2.60	1.80	-0.35
	Language Usage	1400	7.22	5.59	6.25	5.34	-0.97
	MA	1400	15.78	7.40	15.46	7.23	-0.32
	Social Studies	1400	10.93	5.11	10.21	4.95	-0.72
	Science	1400	15.05	7.06	14.75	6.74	-0.30
5	Reading	1393	13.78	5.64	14.04	5.42	0.26
	Writing	1393	3.39	1.79	2.98	1.86	-0.41
	Language Usage	1393	7.10	5.04	7.45	4.52	0.35
	MA	1393	14.56	6.05	13.65	5.96	-0.91
	Social Studies	1393	13.77	6.61	14.03	6.32	0.26
	Science	1393	9.96	6.04	9.79	5.84	-0.17
8	Reading	1406	14.63	6.14	15.79	6.38	1.16
	Writing	1406	3.60	2.01	4.02	1.96	0.42
	Language Usage	1406	9.20	6.14	10.91	5.97	1.71
	MA	1406	10.22	6.67	10.75	6.78	0.53
	Social Studies	1406	14.86	6.59	17.13	7.72	2.27
	Science	1406	16.74	8.26	17.68	8.42	0.94

TABLE 13
1992, 1993, 1994, 1995, 1996, 1998, AND 1999 RATER YEAR EFFECTS STUDIES:
COMPARISON OF RESULTS IN TERMS OF STANDARDIZED RAW SCORE MEAN DIFFERENCES¹

Grade	Scale	Rater Effects Study						
		1992	1993	1994	1995	1996	1998	1999
3	Reading	0.0	-0.2	0.0	0.0	0.0	-0.1	0.0
	Writing	-0.2	-0.2	0.0	0.2	0.0	0.0	-0.1
	Language Usage	-0.2	-0.4	0.0	0.2	0.0	-0.1	-0.1
	Math Content	0.1	0.0	-0.1	0.1	0.0	0.0	0.0
	Math Process	0.2	0.0	0.0	0.1	0.0	-0.1	0.0
	Social Studies ²	---	-0.6	-0.1	0.1	0.1	0.0	0.0
	Science ²	---	-0.2	0.1	0.0	0.0	0.0	0.0
5	Reading	0.3	-0.2	0.1	0.1	-0.1	0.3	0.1
	Writing	0.4	-0.1	0.0	0.1	0.0	-0.1	0.0
	Language Usage	0.3	-0.2	0.0	-0.1	0.0	0.0	0.1
	Math Content	0.1	-0.1	0.1	0.0	0.0	0.0	0.1
	Math Process	0.2	0.0	0.1	0.0	0.0	0.0	0.1
	Social Studies ²	---	0.1	0.2	0.0	0.1	0.1	0.2
	Science ²	---	-0.1	0.1	0.1	0.0	0.2	-0.1
8	Reading	0.0	0.1	-0.1	-0.2	-0.1	0.1	0.0
	Writing	0.0	0.0	0.2	-0.1	0.1	0.1	0.0
	Language Usage	-0.2	0.1	-0.1	0.0	0.1	0.3	0.1
	Math Content	0.1	-0.1	0.0	-0.1	0.0	0.0	0.0
	Math Process	0.1	-0.1	-0.1	-0.1	-0.1	0.0	0.0
	Social Studies ²	---	-0.1	-0.1	0.0	-0.1	0.0	0.0
	Science ²	---	-0.2	0.0	-0.2	0.0	0.1	0.0

¹ These differences were obtained by dividing the difference between the current and prior year mean ratings by the square root of the pooled variances of these ratings.

² This subject was not assessed in this grade in 1991, so comparisons involving 1991 ratings are not available.

TABLE 14
RATER YEAR EFFECTS STUDY TRANSFORMATION VALUES

Grade	Scale	Multiplier R ₁	Addend R ₂	(A) (R ₁ *500)+R ₂	(A) - 500 ¹
3	Reading	1.015	- 6.510	500.990	1
	Writing	1.150	-69.745	505.255	5
	Language Usage	1.031	- 3.415	512.085	12
	MA	1.028	-13.355	500.645	1
	Social Studies	1.010	1.765	506.765	7
	Science	1.055	-26.574	500.926	1
5	Reading	1.049	- 28.763	495.737	- 4
	Writing	1.000	14.000	514.000	14
	Language Usage	1.274	-154.943	482.057	-18
	MA	1.016	0.099	508.099	8
	Social Studies	1.053	-30.773	495.727	- 4
	Science	1.036	-17.772	500.228	0
8	Reading	0.968	10.571	494.621	- 5
	Writing	1.000	-11.000	489.000	-11
	Language Usage	1.037	-35.782	482.718	-17
	MA	0.993	1.404	497.904	- 2
	Social Studies	0.808	87.705	491.705	- 8
	Science	1.026	-19.497	493.503	- 6

¹ Numbers in this column were purposely rounded to improve their comprehensibility

TABLE 15
PERFORMANCE OF STATE ON 1998 MSPAP AND 1999 EQUATING SAMPLE ON 1998 MSPAP

Grade	Scale	State ¹ (98SS ₉₈)			Sample (98SS ₉₉)		
		Mean	SD	N	Mean	SD	N
3	Reading	520.9	43.8	20,214	520.2	43.4	2,362
	Writing	524.4	48.0	20,568	525.8	46.9	2,362
	Language Usage	525.2	59.5	20,636	524.8	58.4	2,362
	Math Content	517.3	57.1	20,587	---	---	---
	Math Process	513.8	52.1	20,587	---	---	---
	MA ²	515.8	52.1	20,587	516.5	48.4	2,362
	Social Studies	509.7	49.2	20,718	511.0	48.8	2,362
	Science	510.2	56.4	20,181	509.5	56.4	2,362
	5	Reading	518.9	49.9	19,945	523.2	50.2
Writing		507.1	54.2	20,136	510.9	55.5	2,412
Language Usage		530.5	59.3	20,291	534.9	57.8	2,412
Math Content		519.1	57.9	20,321	---	---	---
Math Process		511.3	56.0	20,030	---	---	---
MA ²		515.6	55.1	20,030	518.6	56.5	2,410
Social Studies		517.7	56.7	20,282	523.1	55.5	2,410
Science		520.1	55.4	20,332	525.8	56.0	2,410
8		Reading	507.6	37.5	18,033	508.7	38.8
	Writing	503.3	56.8	18,469	509.7	56.4	2,409
	Language Usage	508.5	59.6	18,701	512.0	59.3	2,409
	Math Content	522.5	51.0	18,303	---	---	---
	Math Process	514.8	61.6	18,303	---	---	---
	MA ²	518.9	53.2	18,303	522.3	56.0	2,409
	Social Studies	516.7	53.7	18,552	524.3	54.4	2,409
	Science	527.2	52.2	17,882	529.7	52.2	2,409

¹ State performance results were drawn from the Forms Effect Study carried out for the 1998 MSPAP. The values reported refer to performance on Clusters 3C, 5A, and 8A.

² The State performance results on MA were from the Math Total; The Sample results on MA were from the unified MA.

TABLE 16
EQUATING STUDY TRANSFORMATION VALUES

Grade	Scale	Multiplier T ₁	Addend T ₂	(A) (T ₁ *500)+T ₂	(A) - 500 ¹
3	Reading	0.802	119.204	520.204	20
	Writing	0.856	97.035	525.035	25
	Language Usage	1.418	-190.756	518.244	18
	MA	0.910	63.192	518.192	18
	Social Studies	0.953	35.947	512.447	12
	Science	1.074	-29.443	507.557	8
5	Reading	0.896	74.275	522.275	22
	Writing	1.156	-64.879	513.121	13
	Language Usage	1.047	14.548	538.048	38
	MA	1.133	-50.390	516.110	16
	Social Studies	1.170	-65.080	519.920	20
	Science	1.005	25.529	528.029	28
8	Reading	0.687	164.341	507.841	8
	Writing	0.991	10.919	506.419	6
	Language Usage	1.235	-112.246	505.254	5
	MA	1.134	-39.247	527.753	28
	Social Studies	0.948	48.876	522.876	23
	Science	1.130	-35.691	529.309	29

¹ Numbers in this column were purposely rounded to improve their comprehensibility.

TABLE 17
COMPARISON OF 1998 AND 1999 MSPAP PERFORMANCE BY GRADE AND SCALE

Grade	Scale	1998 State Means	1999 State Means	99 - 98 Difference
3	Reading	519.7	517.6	-2.1
	Writing	523.5	525.0	1.5
	Language Usage	524.0	522.7	-1.3
	MA ¹	515.9	514.8	-1.1
	Social Studies	509.0	508.9	-0.1
	Science	509.4	508.7	-0.7
5	Reading	516.0	516.7	0.7
	Writing	508.0	508.2	0.2
	Language Usage	529.7	532.3	2.6
	MA ¹	516.2	509.5	-6.7
	Social Studies	516.5	517.8	1.3
	Science	521.3	523.4	2.1
8	Reading	508.1	508.1	0.0
	Writing	503.9	507.7	3.8
	Language Usage	510.3	507.2	-3.1
	MA ¹	519.1	521.8	2.7
	Social Studies	518.0	521.0	3.0
	Science	528.8	530.9	2.1

¹ The 1998 and 1999 means on MA were from the Math Total, and the unified Math respectively.

TABLE 18. COEFFICIENT ALPHA FOR 1999 MSPAP CONTENT AREAS**Grade 3**

	<u>A</u>	<u>Cluster</u> <u>B</u>	<u>C</u>
Reading	.80	.83	.79
Writing	.59	.56	.72
Language Usage	.90	.92	.91
Mathematics	.84	.80	.85
Science	.85	.87	.86
Social Studies	.85	.82	.82

Grade 5

	<u>A</u>	<u>Cluster</u> <u>B</u>	<u>C</u>
Reading	.79	.82	.85
Writing	.74	.67	.70
Language Usage	.92	.90	.91
Mathematics	.86	.87	.89
Science	.84	.84	.83
Social Studies	.84	.85	.85

Grade 8

	<u>A</u>	<u>Cluster</u> <u>B</u>	<u>C</u>
Reading	.87	.87	.86
Writing	.77	.76	.67
Language Usage	.91	.92	.89
Mathematics	.88	.90	.89
Science	.90	.89	.87
Social Studies	.89	.91	.90

Note: Clusters 3A, 5B, and 8C are choice clusters.

The reported alpha for the choice cluster are the average alpha across all choices.

TABLE 19. STANDARD ERRORS OF MEASUREMENT - GRADE 3

Reading	Scale Score	Cluster		
		3A	3B	3C
SE at HOSS	650	37	32	56
SE at Level 1/2	620	26	22	38
SE at Level 2/3	580	21	18	27
SE at Level 3/4	530	18	15	18
SE at Level 4/5	490	18	17	18
SE at LOSS	400	39	45	38
Writing				
SE at HOSS	635	37	37	36
SE at Level 1/2	614	31	34	32
SE at Level 2/3	577	29	30	25
SE at Level 3/4	528	33	31	26
SE at LOSS	455	58	57	44
Language Usage				
SE at HOSS	625	25	23	22
SE at Level 1/2	620	24	23	21
SE at Level 2/3	576	21	18	19
SE at Level 3/4	521	21	19	19
SE at LOSS	450	26	32	35
Mathematics				
SE at HOSS	650	24	47	30
SE at Level 1/2	626	19	32	22
SE at Level 2/3	583	15	27	18
SE at Level 3/4	531	15	20	15
SE at Level 4/5	489	19	18	16
SE at LOSS	375	58	45	50
Science				
SE at HOSS	650	24	26	33
SE at Level 1/2	619	21	21	27
SE at Level 2/3	580	19	19	22
SE at Level 3/4	527	18	17	18
SE at Level 4/5	488	20	18	17
SE at LOSS	375	43	37	35
Social Studies				
SE at HOSS	625	23	27	21
SE at Level 1/2	622	23	27	20
SE at Level 2/3	580	18	21	18
SE at Level 3/4	525	17	19	18
SE at Level 4/5	495	18	20	20
SE at LOSS	400	40	32	39

Note: HOSS is the highest obtainable scale score, LOSS is the lowest obtainable scale score.

TABLE 20. STANDARD ERRORS OF MEASUREMENT - GRADE 5

<u>Reading</u>	<u>Scale Score</u>	<u>Cluster</u>		
		<u>5A</u>	<u>5B</u>	<u>5C</u>
SE at HOSS	675	40	42	44
SE at Level 1/2	620	24	23	24
SE at Level 2/3	580	20	19	20
SE at Level 3/4	530	20	18	17
SE at Level 4/5	490	20	17	17
SE at LOSS	375	51	40	40
<u>Writing</u>				
SE at HOSS	595	35	52	37
SE at Level 2/3	567	33	44	35
SE at Level 3/4	522	34	36	35
SE at Level 4/5	488	36	35	36
SE at LOSS	440	46	42	45
<u>Language Usage</u>				
SE at HOSS	625	31	31	22
SE at Level 1/2	597	16	21	16
SE at Level 2/3	567	15	16	15
SE at Level 3/4	533	15	20	15
SE at LOSS	425	49	29	54
<u>Mathematics</u>				
SE at HOSS	650	34	34	27
SE at Level 1/2	617	27	25	22
SE at Level 2/3	575	22	18	17
SE at Level 3/4	520	19	16	11
SE at Level 4/5	473	19	23	18
SE at LOSS	400	29	38	28
<u>Science</u>				
SE at HOSS	650	29	24	24
SE at Level 1/2	625	23	21	21
SE at Level 2/3	580	19	17	18
SE at Level 3/4	525	18	19	18
SE at Level 4/5	484	21	22	22
SE at LOSS	375	48	45	49
<u>Social Studies</u>				
SE at HOSS	625	24	30	24
SE at Level 1/2	619	23	28	24
SE at Level 2/3	580	21	23	21
SE at Level 3/4	529	20	28	20
SE at LOSS	400	30	34	32

Note: HOSS is the highest obtainable scale score, LOSS is the lowest obtainable scale score.

TABLE 21. STANDARD ERRORS OF MEASUREMENT - GRADE 8

<u>Reading</u>	<u>Scale Score</u>	<u>Cluster</u>		
		<u>8A</u>	<u>8B</u>	<u>8C</u>
SE at HOSS	650	55	79	64
SE at Level 1/2	650	55	79	64
SE at Level 2/3	580	17	27	19
SE at Level 3/4	530	11	12	11
SE at Level 4/5	490	11	11	11
SE at LOSS	375	44	57	43
<u>Writing</u>				
SE at HOSS	625	53	52	75
SE at Level 2/3	551	26	29	35
SE at Level 3/4	505	26	27	29
SE at LOSS	425	36	32	31
<u>Language Usage</u>				
SE at HOSS	625	30	31	41
SE at Level 2/3	565	19	19	23
SE at Level 3/4	509	17	19	19
SE at Level 4/5	474	19	18	20
SE at LOSS	425	29	21	22
<u>Mathematics</u>				
SE at HOSS	650	26	30	22
SE at Level 1/2	618	19	21	15
SE at Level 2/3	579	14	14	13
SE at Level 3/4	525	17	14	15
SE at Level 4/5	481	20	19	20
SE at LOSS	400	49	50	46
<u>Science</u>				
SE at HOSS	650	26	24	26
SE at Level 1/2	619	20	18	19
SE at Level 2/3	576	16	13	14
SE at Level 3/4	532	15	13	15
SE at Level 4/5	482	15	17	20
SE at LOSS	375	29	48	70
<u>Social Studies</u>				
SE at HOSS	650	34	40	32
SE at Level 1/2	620	25	26	24
SE at Level 2/3	582	17	17	16
SE at Level 3/4	530	14	12	13
SE at Level 4/5	495	15	13	16
SE at LOSS	375	46	48	48

Note: HOSS is the highest obtainable scale score, LOSS is the lowest obtainable scale score.

TABLE 22. BETWEEN CONTENT AREA CORRELATIONS FOR GRADE 3

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.56	1.00				
Lang. Usage	.59	.73	1.00			
Mathematics	.67	.54	.58	1.00		
Science	.76	.59	.62	.78	1.00	
Social Studies	.70	.58	.61	.70	.76	1.00

Note: N ranges from 59,876 to 64,043.

TABLE 23. BETWEEN CONTENT AREA CORRELATIONS FOR GRADE 5

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.57	1.00				
Lang. Usage	.60	.77	1.00			
Mathematics	.63	.57	.62	1.00		
Science	.67	.59	.64	.76	1.00	
Social Studies	.73	.60	.64	.70	.70	1.00

Note: N ranges from 56,982 to 62,922.

TABLE 24. BETWEEN CONTENT AREA CORRELATIONS FOR GRADE 8

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.68	1.00				
Lang. Usage	.70	.81	1.00			
Mathematics	.61	.59	.64	1.00		
Science	.75	.64	.67	.76	1.00	
Social Studies	.66	.61	.64	.69	.76	1.00

Note: N ranges from 55,031 to 58,837.

TABLE 25. BETWEEN CONTENT AREA SCALE SCORE CORRELATIONS AT SCHOOL LEVEL FOR GRADE 3

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.91	1.00				
Lang. Usage	.90	.93	1.00			
Mathematics	.93	.90	.86	1.00		
Science	.96	.92	.89	.97	1.00	
Social Studies	.94	.92	.88	.95	.96	1.00

Note: N=816

TABLE 26. BETWEEN CONTENT AREA SCALE SCORE CORRELATIONS AT SCHOOL LEVEL FOR GRADE 5

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.92	1.00				
Lang. Usage	.91	.94	1.00			
Mathematics	.91	.90	.90	1.00		
Science	.94	.93	.92	.96	1.00	
Social Studies	.95	.94	.93	.94	.97	1.00

Note: N=810

TABLE 27. BETWEEN CONTENT AREA SCALE SCORE CORRELATIONS AT SCHOOL LEVEL FOR GRADE 8

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.96	1.00				
Lang. Usage	.96	.98	1.00			
Mathematics	.91	.92	.93	1.00		
Science	.97	.96	.96	.96	1.00	
Social Studies	.95	.95	.95	.95	.98	1.00

Note: N=262.

TABLE 28. NUMBER OF ITEMS FLAGGED AS DIFFERENTIAL ITEM FUNCTIONING

Grade 3												
	Reading (44 items)		Writing (11 items)		Language Usage (29 items)		Mathematics (57 items)		Social Studies (52 items)		Science (59 items)	
	+ ¹	- ²	+	-	+	-	+	-	+	-	+	-
Black	0	0	0	0	0	0	0	0	0	0	0	0
Asian	1	1	0	0	0	0	2	0	1	3	0	0
Hispanic	0	1	1	0	0	0	0	0	3	1	0	1
Female	0	0	0	0	0	0	0	0	0	0	0	0
Grade 5												
	Reading (53 items)		Writing (11 items)		Language Usage (29 items)		Mathematics (75 items)		Social Studies (52 items)		Science (58 items)	
	+ ¹	- ²	+	-	+	-	+	-	+	-	+	-
Black	0	0	0	0	0	0	0	0	0	0	0	1
Asian	1	1	0	1	0	0	1	1	0	0	1	1
Hispanic	1	3	0	1	0	2	0	1	1	1	1	1
Female	0	0	0	0	0	0	0	0	0	0	0	0
Grade 8												
	Reading (56 Items)		Writing (12 items)		Language Usage (30 items)		Mathematics (61 items)		Social Studies (58 items)		Science (67 items)	
	+ ¹	- ²	+	-	+	-	+	-	+	-	+	-
Black	0	1	0	0	0	0	0	0	1	0	2	4
Asian	0	0	0	0	0	0	0	0	1	0	0	1
Hispanic	0	2	1	1	0	0	0	0	0	0	2	2
Female	0	0	0	0	0	0	0	0	0	0	1	0

Note 1: The minority group members did better than was expected

Note 2: The minority group members did less well than was expected

TABLE 29. OUTCOME DIFFICULTY INDICATORS

Outcome Number	Outcome	Grade3	Grade5	Grade8
<u>Reading</u>				
2.	Reading for Literary Experience	50	58	53
3.	Reading to be Informed	52	45	52
4.	Reading to Perform a Task	46	43	57
<u>Writing</u>				
1.	Writing to Inform	20	45	54
2.	Writing to Persuade	29	47	59
3.	Writing to Express Personal Ideas	40	50	57
<u>Language Usage</u>				
1.	Language In Usage	37	48	51
<u>Mathematics</u>				
1.	Problem Solving	N/A	N/A	N/A
2.	Communication	31	42	32
3.	Reasoning	28	46	34
4.	Connections	29	33	38
5.	Concepts/Relationships	47	46	44
6.	Measurement/Geometry	56	48	38
7.	Statistics	49	45	38
8.	Probability	39	48	40
9.	Patterns/Relationships	47	N/A	N/A
9.	Patterns/Algebra	N/A	54	37
<u>Science</u>				
1.	Concepts of Science	50	42	50
2.	Nature of Science	37	47	44
3.	Habits of Mind	41	41	44
5.	Processes of Science	48	39	49
6.	Applications of Science	37	31	49
<u>Social Studies</u>				
1.	Political Systems	30	38	49
2.	People/Nation & World	43	51	50
3.	Geography	55	48	52
4.	Economics	32	45	46
5.	Skills and Processes	45	46	50
6.	Valuing Self and Others	36	46	57
7.	Understand/Attitudes	34	51	54

Note: N/A means the outcome is not measured at that grade.

Note: The numbers are percentages of the maximum possible scores.

APPENDIX A
TEST MAPS FOR 1999 MSPAP

APPENDIX B

**NUMBER OF ITEMS COMPRISING EACH OUTCOME FOR 1999
MSPAP**

APPENDIX C

Scaled Score Ranges for Each Proficiency Level
