Psychometric Review of the Maryland School Performance Assessment Program (MSPAP)

Psychometric Review Committee

Ronald K. Hambleton, Chair
University of Massachusetts at Amherst

James Impara
University of Nebraska at Lincoln

William Mehrens
Michigan State University

Barbara S. Plake
University of Nebraska at Lincoln

And Contributions From

Mary J. Pitoniak
April L. Zenisky
University of Massachusetts at Amherst

Lisa F. Smith
Kean University of New Jersey

- December 2000 -

This psychometric report was funded by the Abell Foundation. The Abell Foundation is not responsible, however, for the content and recommendations, and no endorsement by the Abell Foundation should be assumed or inferred. This report benefitted from a careful reading and critique from the Maryland Department of Education (MDOE). However, again, no endorsement from MDOE should be assumed or inferred.

TABLE OF CONTENTS

Psychometric Review of the Maryland School Performance Assessment Program (MSPAP)

Ronald K. Hambleton, Chair
University of Massachusetts at Amherst

James Impara
University of Nebraska at Lincoln

William Mehrens
Michigan State University

Barbara S. Plake
University of Nebraska at Lincoln

Executive Summary

In 1989, the Governor's Commission on School Performance (Sondheim Commission) ublished a report on what Maryland should do to improve its public schools. One of the recommendations was the establishment of a statewide assessment. What followed from the recommendation was an annual statewide assessment, the Maryland School Performance Assessment Program (MSPAP) covering six content areas and administered to students in grades 3, 5, and 8. Today, the MSPAP is a major component of the state's accountability program for schools.

Main Features of the MSPAP

The MSPAP has a number of important features (Features less common in other states or unique to Maryland are noted with an asterisk):

1. *Student assessment is based on tasks that are multi-step, multi-question activities built around a common theme.

2. *Subject-matter is commonly measured through activities that integrate several subject areas. MSPAP tasks typically try to assess more than a single subject. Thus, a MSPAP task may try to measure student learning in, for example, mathematics, reading, and science.

3. *Many of the tasks have pre-assessment activities. These activities are not scored but are intended to acquaint students with the task itself and the sort of work required in the scored portions of the task. Either a large group of students or randomly formed smaller groups of students from each class do these pre-assessment activities.

4. *Students are always asked to construct their answers. This means that multiple-choice and true-false items are not used in the assessments.

5. The MSPAP is criterion-referenced--this means that students are judged not against each other, but against standards of performance that are set in relation to well-defined domains of content measured by the assessments. In theory, every student could be placed in the highest performance category, or every student could be placed in the bottom performance category. Placement is based on test performance, and there are no quotas on the number of students who can or should be placed in each performance category.

6. *Scoring is sometimes based on "patterns of student responses" across related sets of activities as opposed to simply summing up points for correct work on individual questions. This means that a score assigned to a student may be based on his/her responses to a number of questions, rather than scoring each question independently of the others.

7. Individual students in the same subject but in different grades cannot be compared because of the design of the assessments. The significance of this feature is that gain or change scores for students over school years cannot be meaningfully interpreted.

8. At each grade, 20 tasks are used to assess student and school performance across the six content areas. Students are assigned on a random basis to take only six or seven of the tasks. Via "statistical equating," the 20 tasks from one year's assessment can be linked to the 20 tasks in the previous year's assessment so that changes in the level of school, district, and state performance over time can be judged.

9. *Approximately two-thirds of each assessment are tasks that have been administered in previous years.

10. All schools are working against the same performance targets. Neither race nor family educational background of students, for example, is factored into the scoring process.

11. The assessment is principally intended to provide a measure of school-wide performance, not individual or classroom performance. Individual scores are available but not recommended for use by the Maryland Department of Education.

12. MSPAP scores can have both positive and negative consequences for individual schools.

13. MSPAP is intended to provide information for judging school achievement and growth and to provide data to improve classroom instruction.

MSPAP is unique among state assessment programs, and it has been in place for 10 years. Many of the features above, but especially features 1, 2, 3, 4, 6, and 9, are less common in other state assessments around the country. In some states, the focus may be on paper and pencil tests to reduce testing program costs or to deliver scores back to schools quickly. In other states, the focus may be on the assessment of basic skills and multiple-choice test items are quite adequate to assess the curricula. In many states, state law dictates that individual scores be provided. There is no shortage of reasons for differences in assessment programs across states.

The Abell Foundation commissioned two groups--a content panel consisting of subject-matter specialists and a psychometric panel consisting of psychometricians and experts in assessment -- to investigate a number of important issues concerning the MSPAP. The questions and recommendations from the psychometric review panel follow.

Psychometric Review

The psychometric panel agreed to provide commentary on four broad policy issues:

1. The pros and cons of an assessment that reports individual scores versus an assessment that reports school-wide scores, and, how the assessment would need to change if a policy decision was made to report individual scores.

2. An analysis of trends in student performance on the MSPAP compared with other national indicators, the validity of such comparisons, the identification of various factors that contribute to changes in MSPAP scores, and how these factors which are not directly related to increases in student achievement can be minimized.

3. The advantages and disadvantages of the MSPAP's dual role as both an accountability tool and a model for classroom instruction.

4. Whether or not any existing research on assessment would indicate the MSPAP might have a greater negative impact on high poverty schools relative to other schools in Maryland, any more so than any other type of assessment.

In addition, the psychometric panel agreed to address the technical merits of the MSPAP. Using the new American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) Test Standards , the psychometric panel addressed (5) choice of grade levels for MSPAP, (6) test development procedures, (7) scoring, (8) validity evidence, (9) standard-setting, (10) detection of potentially biased test items, (11) reliability evidence, (12) statistical equating

of forms administered within a given year, and forms from one year to the next, and (13) score reporting. In addition, to provide needed background, the psychometric panel compiled two reports-the first reviewed criticisms and concerns directed against the MSPAP, and the second addressed several features of performance assessments and their impact on the reliability and validity of scores.

Findings and Recommendations

A summary of our findings and recommendations follow:

1. On the issue of student versus group reporting, it is clear that this issue is one to be decided by policy-makers. Certainly, if the decision is to report student scores, and there are both positive and negative reasons for doing so that are identified in our report, substantial changes will be needed to the structure of the current educational assessments. The amount of changes, to a large extent, will depend on the intended purposes of reporting individual scores. Test design features that would need to be reconsidered in a new assessment design include the current use of non-parallel forms, lack of content coverage of the individual forms or clusters, a less than acceptable level of score reliability for individual score reporting, and the current non-standardized test conditions. These test design features, however, are not problems with the current focus on school-level score reporting.
2. On the issue of trends of results on national tests versus MSPAP results in Maryland, our panel has reservations about the utility of these comparisons. MSPAP results in reading have shown somewhat greater changes (improvements) over the last ten years than changes on national tests such as the Scholastic Assessment Test (SAT) and the National Assessment of Educational Progress (NAEP). But SAT and NAEP do not measure the same constructs that the MSPAP tries to measure, and instructional initiatives have been focused on the MSPAP curricula. More changes would be expected on MSPAP than either SAT or NAEP.
3. On the issue of MSPAP providing an accountability function as well as modeling classroom instruction, the panel notes that it is difficult for an assessment system to be optimal to serve two different purposes. An optimal assessment system needs to be efficient and standardized across students, and lead to reliable and

valid information. Modeling classroom instruction is a very different purpose and does involve advanced group preparation for tasks, group work, teacher input, etc. Time to completion of tasks may not be a major consideration in modeling classroom instructional practices. It seems impossible for an assessment system to achieve both purposes in an optimal way.

4. On the issue of MSPAP's impact on high poverty schools, the panel was not able to compile any direct evidence from Maryland–time and resources did not permit this activity, but a literature review on the topic turned up some interesting findings. It would appear that teacher mobility is higher in lower SES schools and this is likely to impact negatively on the MSPAP results in these schools. New teachers in these schools tend to be inexperienced teachers and might be expected to be less familiar with performance assessments like MSPAP. Both of these factors work against high poverty schools performing well on MSPAP. There is also some evidence that students from low poverty schools may take MSPAP and other assessments less seriously. Again, school results will be negatively affected.

Our findings regarding selected psychometric aspects of MSPAP are generally positive. The state and its test contractor have done a good to excellent job of delivering, scoring, and reporting complex assessment materials. Many aspects of the process are state-of-the-art in the assessment field. A summary of our main findings and recommendations follows:

5. While testing students at all grades has merit, the panel recognizes the logistical and cost issues involved. The choice of testing at grades 3, 5, and 8 is reasonable from the perspective of providing data on school accountability. For providing information

that may be useful for improving instruction, the choice of grades is reasonable, but it may not be optimal.

6. Test development appears to be one of the main strengths of the MSPAP. Clear steps are available. One shortcoming is that some of the steps may not always be carried out and if that is the case, a change in practice is needed. The review process seems to be exemplary. Our recommendations in this area for improvement are minor.

7. Scoring seems to be handled in a very responsible way. Procedures appear excellent, but evidence of compliance was less apparent to us. Our recommendations are centered on the insertion of more quality control measures into the scoring process. It is unlikely that scoring can be handled any faster without sacrificing accuracy and reliability.

8. In the area of test score validity, many excellent examples of validity evidence were observed–the work of review committees, for example. At the same time, it is in this area that the panel's biggest concerns are centered–the timing of the administration of the assessment is out of step with the delivery of the subject matter content in the classroom, there was no direct evidence that we saw that demonstrates that the tasks are measuring higher level thinking skills, there appears to be a confounding of the assessment of higher level thinking skills with writing skills, non-standardized testing conditions due to the pre-assessment activities are used, questions remain about the developmental level of the tasks in the assessment, there is a lack of evidence of divergent validity among the subject matter scores, inexperienced teachers who are not familiar with performance based assessments like the MSPAP may be impacting negatively on scores, and there appears to be a lack of documentation of what the

positive and negative consequences have been for individual schools because of the MSPAP scores. The panel is not suggesting that the current assessments lack validity, but we are concerned that more validity evidence is not available at the present time to support the uses of MSPAP. Our panel strongly recommends that more evidence needs to be compiled and the evidence used to support, or modify MSPAP itself. Other changes the panel would recommend be made immediately include: (a) drop the use of group-based, pre-assessment activities, (b) reconsider the use of manipulatives in the assessment, (c) train/retrain teachers in the lower SES schools to handle performance assessments and preparing students to take the MSPAP, and (d) discontinue making individual scores available to parents and students unless a major redesign of the assessment takes place.

8. The setting of performance standards (the scores needed for students to be assigned to the five levels of performance on MSPAP) many years ago was commendable and the documentation of this process involving many Maryland educators was excellent. At the same time, with many shifts in the curricula and assessments, it may be time to reconsider setting the performance standards again. Many measurement specialists recommend that the setting of standards be considered every five years or so. If standards are reset, then validity evidence should be compiled as well.

9. The efforts to carry out item bias analyses (the detection of assessment material that may be inappropriate for subgroups of students such as Blacks, Hispanics, Asians, and females) appear laudatory. The panel would like to see more documentation of this work included in the technical manual.

10. The evidence to support the reliability of group scores is strong and certainly strong enough to support school level reporting. The panel would like to see inclusion of information about the reliability of classification decisions for students and information about the confidence (i.e., standard errors) of group statistics that are being reported.

11. Our findings are that the linking of forms within a year, and linking of assessments from one year to the next is being well done. The panel saw no problems. We are concerned about the reuse of tasks over time (i.e., the security problem), but felt that steps were being taken to detect potential leaks and not allow them to distort the linking process. At the same time, the panel notes that equivalence of test form dimensionality is an assumption in the linking and on the surface at least, the assessments appear to be multidimensional and non-equivalent across test forms. Additional evidence of the extent to which model assumptions in the linking process are being met and the possible consequences of misfit would be desirable.

12. Score reporting is one of the strengths of the MSPAP. The panel commends the Maryland Department of Education for its openness in reporting and for the clarity in its reports of the MSPAP. Also, we are opposed to making individual scores available with the current assessment design. Finally, reporting gain scores from one year to the next without addressing their limitations seems inappropriate.

Conclusions

MSPAP has its strengths and weaknesses, like all state assessment programs. From a technical perspective, the psychometric panel saw many positive features in MSPAP–procedural steps in test development, steps in scoring the assessment data, initiatives to identify potentially

biased test items, the approach used to set performance standards on the assessments, the linking of forms from one year to the next, and the way in which the scores are reported each year. There is also one major shortcoming and that is the failure, to date, to compile what the panel believes to be a sufficient amount of validity evidence to support the multiple intended uses of the MSPAP. Our panel is not concluding that MSPAP is invalid for its multiple purposes; the panel is suggesting that considerably more work needs to be done to compile validity evidence, and then use the evidence, as appropriate, to continue to improve MSPAP. The panel also suggests a number of additional technical improvements that might be made in the coming years, some are suggested above, and others appear in the full report of the panel.

1. Scope of Study and Study Design

In 1989, the Governor's Commission on School Performance, commonly known as the Sondheim Commission, released a study of what Maryland should do to improve its public schools. One of the most important recommendations in this study was the establishment of a statewide assessment. From this recommendation followed the Maryland School Performance Assessment Program (MSPAP), an annual assessment covering six content areas and administered to students in grades 3, 5, and 8, that serves as a major component of the state's accountability program.

The MSPAP can be described by a number of features, several of which are not common features in statewide assessment:

1. Its format is structured around "tasks," multi-step, multi-question activities built around a central theme. The content of the tasks is derived from a 1990 document called the Maryland Learning Outcomes and Indicators.

2. Subject areas are commonly tested through integrated activities. That is, more than one subject area (domain) is often tested within a single task.
3. Many tasks include "pre-assessment activities." These activities do not produce scored responses, but are intended to provide context for the scored portions of the tasks. In addition to whole-group brainstorming, many of the MSPAP tasks incorporate activities performed by pairs or by small groups of students (e.g., 3-4) arranged at random by the proctor administering the assessment.
4. The MSPAP asks for students to construct their own answers. The assessment contains no selected-response items (e.g., multiple-choice or true-false test items).
5. The assessment is criterion-referenced, and is designed to be challenging for a student to achieve the highest-performance category.
6. The assessment is scored in some instances based on the response patterns of the students as opposed to the number of correct answers given. This is done separately for each content area at each grade.
7. Scores are interpreted separately at each grade level. It is not possible, for example, to compare the score of a third grade student in a content area with the score of a fifth grade student in the same content area.
8. The assessment is given in three forms each year that together cover the content domain. There are approximately 20 tasks for each grade, meaning that each of the three forms consists of approximately 6-7 tasks. The forms are equated or linked to each other by administering them to random groups of students in each school and adjusting scores to achieve approximately equivalent distributions. Equating to the previous year's assessment is achieved in similar fashion, using one of the previous year's forms administered to additional testing groups in a representative sample of schools.
9. Approximately two-thirds of each MSPAP is made up of tasks that have been used previously and about one-third of the tasks are new for that year.

10. All schools have the same performance targets. They are not differentiated based on such considerations as demographic data.
11. The test is a measure of school-wide performance, not individual or classroom performance. Interpretation of individual scores is not recommended.
12. There is an array of both positive and negative consequences for schools based on MSPAP results.
13. The MSPAP is perceived not only as an assessment of achievement but as a vehicle for school improvement, with the intention of having an impact on daily instruction.

The Abell Foundation formed two study groups in the summer of 1999: (1) a technical or psychometric methods study group consisting of psychometricians and experts in assessment, and (2) a content study group consisting of experts in instruction with subject matter expertise. To the extent possible, the Abell Foundation's intent was to have the two study groups hold meetings concurrently. At the single meeting held by the psychometric methods group (in late January of 2000), the two study groups did meet at the same time and in the same hotel in Boston, and did share information and notes. Efforts were made to produce a combined executive summary but premature release of a draft of that document and the rather different conclusions arrived at by the two committees in the few areas where the reports overlapped, made the preparation of a single executive summary impossible.

The new AERA, APA, NCME Test Standards (1999) provided a framework for the psychometric review that was carried out by our panel. The following broad issues were addressed by our panel:

1. The pros and cons of an assessment that reports individual scores versus an assessment that reports school-wide scores, and how the assessment would need to change if a policy decision was made to report individual scores on a wide-scale basis. (See Section 2.1.)
2. An analysis of trends in student performance on the MSPAP compared with other national assessments, the validity of such comparisons, what all the various factors are that contribute to changes in MSPAP scores, and how factors which are not directly related to increases in student achievement can be minimized. (See Section 2.2.)
3. The advantages and disadvantages of the MSPAP's dual role as both an accountability tool and a model for classroom instruction. (See Section 2.3.)
4. Whether or not any existing research on assessment would indicate that the MSPAP might have a greater negative impact on high poverty schools relative to other schools, any more so than any other type of assessment, with particular discussion of:
   - whether or not there is any probable credence to a hypothesis that poor schools reporting a high teacher transfer rate are likely to do less well because the teachers lack familiarity with MSPAP tasks which have appeared previously, any more so than any other type of assessment.
   - whether or not there is any probable credence to a theory that poor students are any less likely than middle class students to take a test seriously that does not report individual scoring. (See Section 2.4.)

Appendix A contains a report that summarizes many of the criticisms of MSPAP that have been expressed in the press over the last three or four years. Clearly, a wide array of questions have been raised about the psychometric merits of MSPAP. Many of the issues and questions that are addressed in the press such as concerns about group versus individual score reports, and the reliability and validity of the MSPAP scores are addressed in this report. Among the features of

MSPAP that have been criticized and are addressed in this report (and see Appendix C also) are the following:

a. Effect on test validity of the use of manipulatives during the assessment and their value to the assessment.
b. Effect on test validity of the practice of preassessment activities involving cooperative grouping, peer review, and brainstorming and their value to the test.
c. The time it takes to complete the test (9 hours), the pros and cons of this amount of time; possible ways to reduce time with articulation of tradeoffs.
d. The 6-month period that it takes to score the test and release results; possible ways to shorten time with articulation of tradeoffs.
e. The choice of grades 3, 5, and 8 for assessment versus a set of different grades or some form of assessment in every grade.
f. Comparison of appropriateness of activities students are asked to perform, and the reasoning and writing levels expected at each grade level.
g. Consistency, objectivity, and clarity of the scoring guides and processes.

In summary, the technical criteria for judging MSPAP were based on such indicators as:

- test score validity and reliability.
- item bias analysis.
- linking of forms in the same year and across years.
- approach used to set passing scores.
- presentation of the results to the public.

In the next section, the four policy-related questions are addressed. Subsequent sections are organized around different technical aspects of the MSPAP. In the final section, all conclusions and recommendations from the technical review are presented.

2. Discussions of Policy-Related Questions

1. Pros and Cons of Assessment Programs Reporting Individual Versus Group Scores

A substantive concern that has been raised about MSPAP is the desirability of reporting individual scores to parents, teachers, and students (see, for example, comments reported in Appendix A). This concern has been expressed as "What are the pros and cons of an assessment that reports individual scores versus an assessment that reports school-wide scores, and how would the assessment need to change if a decision was made to report individual scores? This large question has been subdivided into two issues.

Issue 1: Should MSPAP report usable, interpretable scores for individual students or report scores that permit inferences only for schools, districts, and the state?

Issue 2: What changes would be needed in order for MSPAP to report interpretable individual scores?

Issue clarification: At the present time, individual student scores are reported to schools. The Technical Report (MSDE, et al., 1998) states that scale scores for individual students are not interpretable for a variety of reasons (e.g., lack of comprehensive coverage of outcomes, non-parallelism of content and difficulty across clusters, lack of reliability of individual scores resulting in large errors of measurement, and use of non-standardized group work and manipulatives). The MSDE does not encourage schools to provide the individual scores to parents, but there are some schools that do release these scores to parents, either routinely or at the request of the parent. The MSDE has provided materials to principals to assist them in interpreting these "uninterpretable" scores to parents, but these materials do not include some of

the information recommended by the primary contractor, CTB/McGraw Hill, or the program's National Psychometric Panel. We recommend below and elsewhere in this report that without substantive program changes, the current practice of reporting individual student scores to schools should be discontinued. Thus, the issue is whether to change the program in such a way as to make the individual scores that are reported interpretable in some way.

Discussion of Issue 1. This issue is, for the most part, not a psychometric issue, but rather a policy issue. However, if the decision is made to report interpretable individual scores, there are psychometric implications that will need to be considered. The psychometric implications will depend, in part, on the purposes for reporting interpretable individual scores. Some possible purposes for reporting interpretable individual scores might be one, or some combination of the following:

1. Respond to the concern of taxpayers that, given the large commitment of educational time and financial resources commanded by the program, performance information at the individual student level is warranted.
2. Provide parents with information about their child's performance in subject areas or on learning outcomes deemed important by the state.
3. Provide students with an incentive to perform well, because their parents and teachers will receive performance reports.
4. Provide feedback to schools and teachers about individual student performance to assist them in making instructional decisions about the student (e.g., placement into classroom instructional groups).

The current purposes of MSPAP are to provide accountability information to the public on school, district, and state level performance on important educational outcomes and to provide schools and districts with information to assist them in making school-level instructional decisions. The advantages of an assessment program that reports interpretable scores only at the

school, district, and state levels is cost and efficiency. The use of matrix sampling means that a great deal of information can be collected on a wide array of content, while not requiring all students to respond to all, or even comparable tasks.

It is assumed that the decision to produce interpretable individual student scores will not refocus the program, but will add one or more additional purposes. **Any decision to provide interpretable individual student scores will require changes in the current program.**

If the decision is made to report individual scores to school and parents, there may be some advantages.

Among these advantages are:

1. Parents will have more information about their child's school performance on important educational outcomes.
2. Student motivation to perform well on the assessment may increase. This may be particularly important in schools and districts with students from low-income families.
3. The data may provide information to assist in making instructional decisions about individual students (e.g., placement into classroom instructional groups). The timing of the return of the individual scores is unlikely to facilitate making substantive decisions about a student's instructional strengths and weaknesses (the data would not be available until the fall following the spring administration).
4. School level performance information may be more accurate due to increased information about student performance, providing the total information available to the schools is not compromised (i.e., the number of outcomes assessed in any particular year remains the same).

5. School level performance may be more accurate if student motivation to perform to capacity could be increased. This accuracy may be most likely to be observed at the grade 8 level.

Disadvantages of reporting interpretable individual scores:

1. The decision to report accurate, interpretable individual scores at any level of detail will require substantial changes in the current program. The extent of the change will depend on the nature and potential utility of the individual scores that are desired (e.g., reporting scaled scores for each subject area versus reporting scaled scores for specific learning outcomes). Some of the changes that would be required (e.g., discontinue the group-based pre-assessment tasks) could change the philosophy of the current program. These changes are discussed in more detail under the discussion of Issue 2.
2. Program changes necessary to report interpretable individual scores means that program costs are likely to increase. These costs include both the fiscal costs to the state and the costs in terms of student and classroom time to administer the assessment, and the time required to score the assessment and produce the score reports. Including some selected-response questions (e.g., multiple-choice test items), or other questions that can be scored quickly and objectively, rather than relying completely on the more time consuming constructed-response questions might reduce some of these added costs. If the changes result in adding consequences to the students as a result of reporting individual scores, then some additional quality control strategies during scoring may be needed to insure that all individual scores are accurate.
3. Students, teachers, and principals may feel under more pressure to perform well if individual student scores are systematically reported to parents. Such pressures can lead to undesirable consequences like cheating. (For example, cheating on

the high stakes tests in Massachusetts was reported in the national news on February 25, 2000.)

4. Because of the need for greater information about each student across the content areas and outcomes, there is a substantial risk that the assessment content will be narrowed and this could result in a narrowing of the curriculum.

We are not able to advocate whether MSPAP should or should not report interpretable individual scores to the schools and parents. This decision should be based on the specific needs for such data and after weighing the relative advantages and disadvantages of providing such data. There are a number of concerns, both psychometric and administrative, that feed into any decision about this issue. We do, however, strongly recommend that unless specific program changes are made, that the current practice of reporting individual scores to schools should be discontinued.

Discussion of Issue 2. What changes would be needed in order for MSPAP to report interpretable individual scores?

As pointed out in the discussion pertaining to Issue 1, there are limitations to the current MSPAP design that inhibit the production of interpretable individual scores. These issues include:

1. Non-parallelism of content and difficulty across clusters.
2. Lack of comprehensive coverage of outcomes.
3. A lower than desirable level of reliability of individual scores.
4. Use of non-standardized group work and manipulatives.

If it is decided that interpretable individual scores are desirable, redesign of MSPAP will be needed. Features critical to this redesign are as follows:

1. Non-parallelism of content and difficulty across clusters. The assessments would need to be redesigned so that all students take a comparable set of questions. This

could be accomplished by a) all students taking the exact same assessment, b) students taking different forms of the assessment that are designed to be parallel or comparable, or c) students taking a two part assessment, some of which is common to all students and some of which is matrix sampled so that only a subset of the students respond to that particular cluster of matrix sampled questions. Individual student scores would be based on the common part. Having all of the students take the same assessment (option a) would likely result in an increase in testing time as more of the content would need to be presented in the common test form. In addition, with only one test form, it is likely that the curriculum would be narrowed to focus on the content tested. If students took comparable test forms (option b), many of the same issues would result as with option a, but the narrowing of the curriculum would likely be less pronounced. However, more time and costs in test development would result as multiple parallel forms would need to be administered annually. Option c (common part plus matrix sampling) provides many of the benefits currently experienced with MSPAP, due to the inclusion of matrix sampling, but allows for a common set of content to form the basis of student scores, ensuring score comparability. More time would need to be devoted to testing as the common part would need to represent the content domain. Option c is the current test design in Massachusetts.

2. Lack of comprehensive coverage of outcomes. Sufficient numbers of questions would be needed in order to report interpretable scores. Because the administration already requires nine hours of testing time, it is difficult to expand the number of questions using the same constructed-response format. Instead, the incorporation of more structured questions and answer response formats should be considered. These could include multiple choice questions, but in addition could include table completion, restricted option choices, and more directed answer frameworks. Some may argue that these item and answer types reduce the ability of the assessment to measure higher-order thinking skills, but that is not

necessarily the case. It is possible to design multiple choice questions, for example, that measure higher-order reasoning, analysis, deductive reasoning, and evaluation skills. This change would also likely reduce time needed for scoring as the scoring of these item types is either prescriptive (match to keyed response) or requires less scorer processing and judgment. We note also that the 1989 Governor's Commission on School Performance Report made the case for criterion-referenced testing over norm-referenced testing, but did not recommend the exclusive use of constructed response item formats. Making such changes, though, would have several implications for the MSPAP program, including disturbing longitudinal interpretations.

3. A lower than desirable level of reliability of individual scores. There are several sources of error in individual MSPAP scores that would need to be reduced in order for the reliability levels of individual scores to be high enough to warrant interpretation. Because of time restrictions, students receive only a few tasks, and the set of tasks does not represent a systematic sampling of the content of the subject being tested. Student scores are also affected by the variability among raters. There are several ways that these sources of error could be reduced. One way would be to increase the representation to the content domain of the tasks responded to by the individual student. Again, this would require adding questions and tasks to the assessment. Unless the assessment was redesigned to include some fixed/restricted format type items, additional time would be required to administer and score the assessments. By using more short-answer and fixed format/restricted answer questions, score consistency would likely increase. Individual score reliability would increase which is necessary for interpretable individual student scores.

4. Use of non-standardized group work and manipulatives. Inclusion of group work in the assessment activities affects the validity of the individual student scores. Individual members of the group may be directly influenced by the helpfulness of

other group members. Manipulatives are sometimes used in the administration of the assessment in a non-standard way. The specific manipulatives can vary across teachers, classes, and districts. Advance preparation of the materials and manipulatives is sometimes required and may not be carried out correctly or consistently by teachers. Therefore, both the inclusion of group activities and manipulatives may advantage or disadvantages individual students and therefore is an unwarranted source of individual score differences. They should be discontinued in order for individual scores to be interpretable.

Taken together, these program changes could include incorporating at least some fixed-format, restricted response questions. Group work and manipulatives would be discontinued.

It might be useful to look as these possible program changes in light of the major considerations that helped shape MSPAP's original design. These included:

1. The state assessment should focus on the schools and the school system.
2. The emphasis should be on school improvement.
3. The assessment should focus on the ability of students to apply knowledge, by engaging in higher order thinking.
4. The state assessment should embody and reflect sound instructional practice.

Student scores would form the foundation for school-level results, except that these results would likely be more reliable and valid. They could provide more information (or at least not less content coverage) than is now provided in the school-level scores. The emphasis of the program could still be on school improvement. Nothing needs to be lost at the school level with these changes, as long as they maintain or add to, rather than reduce, the content coverage provided by the current MSPAP.

2. Trends in Student Performance on the MSPAP Compared to Other Indicators

It is a well-known principle in research that findings or results gain credibility when they can be replicated or supported by other evidence. For example, when the validity of the NAEP performance standards was criticized recently as too high, validity evidence was compiled to compare the estimate of the percentage of grade 12 students who were in the Advanced category on the 1996 NAEP Science Assessment with the number of students achieving a score of 3 or higher on the Advanced Placement Exams in Science. As NAEP and AP exams are totally independent, similar findings about the numbers of Advanced level students in the country from these two testing programs would provide external evidence to support the validity of the NAEP performance standards. Different results from the two tests would most likely raise questions about the validity of performance standards on the NAEP because the AP exams are a more established program and the validity of their performance standards is accepted in the field. But of course, it is possible that the results could be different and still the NAEP performance standards are valid. One explanation for any differences may be the content frameworks for the tests (they are certainly very different), or that one testing program is low stakes (NAEP) and the other is high stakes (AP exams). High school students simply do not try as hard on the NAEP as they do on AP exams where they can earn university credits. Another possibility is that obtaining scores of 3, 4, and 5 on an AP exam does not have the same meaning as Advanced on a NAEP Assessment. Recently, the AP exams have even been criticized for having lower standards today than years ago. In this instance, the "gold standard" for judging the validity of the NAEP performance standards is now being challenged too by researchers who probably have no knowledge or even interest at all in the NAEP performance standards debate. The point being made is the use of external evidence is not without its own problems and difficulties in interpretation.

In Kentucky, the achievement gains reported in reading between 1992 and 1994 on the state assessment were compared to gains in reading proficiency on the NAEP (Hambleton, Jaeger, Koretz, Linn, Millman, & Phillips, 1995). The question at the time concerned the possibility that

the achievement gains being reported by the state were inflated. Kentucky was reporting substantial gains in school achievement due to its educational reform initiative known as KIRIS and it seemed important to decide whether the gains being reported were real and due to educational reform or were in some way due to test preparation, that would not reflect the acquisition of new knowledge. This NAEP-KIRIS comparison seemed to have special value because the state's new curricula were based on the NAEP content frameworks. The comparison was informative to policy-makers, educators, and the public in Kentucky but far from conclusive. The content frameworks were not exactly the same, NAEP results were only available on one grade and one subject, testing conditions were different, etc.

Similar questions are being raised about the gains in achievement in Maryland. Are they real or due to one or more artifacts such as test preparation? Other possible factors include coaching of students to perform in ways that mimic high scores on the scoring guides on the performance assessments; narrowing the curricula to focus only on outcomes that are expected on the assessments; ever increasing pressure on students from teachers, administrators and parents to increase motivation to perform well on the assessments; practice to do well on the particular assessment (a skill that may not be related to real achievement gain); and so on. But as was highlighted in the NAEP and Kentucky examples, interpreting external evidence is fraught with problems. Care must be taken to interpret the findings.

There are at least three external variables that might be used to investigate the validity of state gains in achievement: The Scholastic Assessment Test (SAT), other standardized achievement measures (such as one of the major national standardized achievement tests), and the National Assessment of Educational Progress (NAEP). The SAT measures potential for success in college and, though MSPAP assesses children at grades 3, 5, and 8, one of the goals is surely to increase the potential or likelihood of these children going on to college. However, the SAT is limited to assessing broad predictors of grades in college (verbal and quantitative ability); the SAT only uses multiple choice items, writing skills are not assessed; not all students in Maryland

take these tests (only those who are college-bound); and the tests are only administered to juniors and seniors in high school. While high scores on the SAT are valued, the SAT is not an adequate measure of the broader achievement goals in Maryland.

There is one other shortcoming of the SAT. In the case of MSPAP, children who entered the Maryland schools in 1990 when educational reform was initiated are only now reaching the age when they will be taking the SAT. SAT scores available previously are from a select group of students who have had substantially less than 10 years in the schools since the reforms were introduced. Still, some increase of SAT scores might be expected. Note though that it is possible that average scores may go down, if more Maryland students feel that they want to go on to college, another desirable outcome of the educational reform initiative in Maryland. The highly variable participation rates across states, and changes in participation rates over time, contribute to a murky interpretation picture. We did not look at trends in SAT scores in preparing this report.

The use of standardized norm-referenced achievement tests is more promising than the SAT for comparing the educational achievement progress of Maryland students. These tests are available at all grade levels and in many subject areas. At the same time, these tests may not measure more than a fraction of the Maryland curricula and the tests themselves typically use multiple-choice test items. At best these norm-referenced tests provide a limited idea of student progress but the availability of national norms can be quite helpful. Unfortunately, data from norm-referenced achievement tests are not widely available in Maryland but, where they are available, the results may be meaningfully used to provide an external perspective on achievement growth in Maryland.

National Assessment of Educational Progress (NAEP) is potentially the most useful external measure for judging student progress in Maryland. The NAEP assessments measure knowledge and skills that are based on national curricula frameworks; they contain a substantial portion of

performance material, the assessments, like MSPAP, are criterion-referenced in their character. But any comparisons between NAEP and MSPAP results are not without their shortcomings. For one, MSPAP assessments are highly contextualized. Students are provided with advanced information and preparation. This is not the case with NAEP. Second, while students in neither assessment are provided scores, group work is extensive with MSPAP except for the final preparation of answers. Third, NAEP is a mixture of multiple-choice and constructed response items. In MSPAP, constructed response is the only format used and writing skills are important in large parts of the assessment. NAEP takes a maximum of 90 minutes of a student's time. MSPAP may require nine hours. It is almost certainly the case too that "satisfactory" on the MSPAP is not identical to "proficiency" on the NAEP. (This problem though would be less consequential if this comparison was avoided.) NAEP seems to fall far short of being the "gold standard." It does not require as much time and it is not closely aligned with the curricula in Maryland.

Nevertheless, we looked at a comparison between trends on the NAEP assessment and MSPAP. First, we looked at the 1993 to 1998 composite scores on MSPAP as reported by the Maryland Department of Education in 1998:

Year	Percent of Students at the Satisfactory Level
1993	31.7
1994	35.3
1995	39.6
1996	40.7
1997	41.8
1998	44.1

These percents are obtained by averaging the results across subjects, schools, and grades.

Using MSPAP composite scores as the dependent measure, the percent of Maryland students performing at the satisfactory level has increased from 31.7% in 1993 to 44.1% in 1998. This means, and was reported correctly by MDOE, that 12 more students out of every 100 are now performing at the satisfactory level over the five-year period. We will let policy-makers, educators, and the public decide whether this increase and growth rate is satisfactory or not. But we would disagree with the critics of MSPAP who have characterized the growth as "surging." Moving 12 more students of every 100 from the not satisfactory to satisfactory category when the initial level of satisfactory students in 1993 was 31 of every 100 does not seem to indicate very much inflation of results. The results were described by the Department as "strong."

Recently, the Southern Regional Education Board (SREB) released an analysis of NAEP fourth grade reading scores in 1992 and 1998 for 16 states–Alabama, Arkansas, Delaware, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia (Creech, et al., 2000). This time period works well for looking at reading results in Maryland. In 1992, the grade 4 students tested had had two years in the new program. In 1998, many teachers would have been implementing the program for many years, and the grade 4 students would have been exposed to the reading program for all of their years of schooling. At the national level between 1992 and 1998, there has been no change in the combined percentage of students scoring at the Basic, Proficient, and Advanced levels – the rate is 60%. There are three interesting and important findings:

1. In 1992, only three (of the 16) states performed above the national level. In 1998, six of 16 states performed above the national level and Maryland was one of states.
2. Twelve of the 16 states showed an increase in reading performance between 1992 and 1998 (though not all 12 states showed statistically significant increases). Maryland was one of the states showing an increase and the size of the increase was sizable (about 3%).

3. Of these 16 states, only five states showed a bigger increase between 1992 and 1998 than Maryland (and in several of these states, the size of the increase over Maryland was not statistically significant).

Also, SREB provided an analysis of the percent of fourth-grade students who scored at the proficient and advanced levels on the 1992 and 1998 NAEP reading assessments. The proficient level was the goal set by the National Assessment Governing Board for performance on the various NAEP assessments. At the national level, the gain was 3% (from 26% in 1992 to 29% in 1998). Again, there are three interesting findings.

1. In 1992, only two (of the 16) states performed at or above the national level. In 1998, six of 16 states performed at or above the national level. Maryland was one of the six states.
2. Twelve of the 16 states showed an increase in reading performance between 1992 and 1998 (though not all 12 states showed statistically significant increases). Maryland was one of the states showing an increase and the size of the increase was sizable (over 5%).
3. Of the 16 states, only one of the states (Kentucky) showed a bigger increase between 1992 and 1998 than Maryland and of course Kentucky too has been in a major educational reform initiative.

These NAEP results in reading between 1992 and 1998 seem to corroborate reported gains in achievement in Maryland. It is a judgment but the results from MSPAP appear to be greater than the results reported on NAEP but definitely not out of line as some critics have claimed.

Many factors can impact on MSPAP scores and contaminate interpretations of achieve-ment. A number of steps can be taken to minimize the importance of these factors: (a) minimizing the amount of repeated assessment material, (b) continuing to build assessments that tap the broad

curriculum–if you want things taught they need to be on the assessment occasionally, and (c) varying the formats, so that teachers are forced to use other formats in their work.

Perhaps the bottom line is that efforts to compile and interpret external validity evidence should be handled very carefully. External measures should be chosen wisely, their strengths and weaknesses in relation to their intended use should be highlighted.

3. Advantages and Disadvantages of the MSPAP's Dual Role As An Accountability Tool and as a Model for Classroom Instruction

MSPAP's main purpose is to measure school performance in six subject areas at grades 3, 5, and 8 for monitoring improvement over time. MSPAP is focused on the assessment of higher level thinking skills of students that can be demon+strated through finding solutions to applied "real world" problems. A secondary purpose of MSPAP is to provide data to schools and other agencies for guiding school improvement planning.

One of the important features of the current MSPAP design, and a source of considerable discussion, is that the assessments or tests are intended to model what is believed by curriculum specialists in Maryland to be good instruction or instructional practices. This means that the assessments of student performance involve a substantial amount of writing (effectively, every student answer is written, though sometimes only a word or two may be needed); the student tasks on the assessments involve using skills cutting across multiple subject areas; and these tasks often consist of small group activities involving advanced reading materials, student discussions, peer reviews, brainstorming ideas, and cooperation, before independent work by each student on the task begins; the use of calculators and dictionaries is encouraged; and lots more. The basic idea seems to be that if teachers are going to teach to the test, and there is substantial evidence to show that they do (i.e., tests drive instruction), then the advantage of

assessments or tests reflecting good instruction seems obvious. This seems, in principle, like a definite advantage for modeling good instruction in MSPAP.

At the same time, there appear to be a number of potential disadvantages from incorporating good instructional practices into the accountability system. One question that has been raised is, Can a single assessment each year be optimal for both school accountability and modeling good instructional practices? The fact that substantial amounts of time are taken up in the assessment with activities that are preparatory to students indicating their knowledge and skills, means that test time is lengthened. Currently nine hours of testing are needed. It is true that the assessment covers six subjects (that's only 90 minutes per subject at three grades in the first nine years of schooling), but were the preparatory material not needed, or were some of the skills assessed with multiple choice items, two actions that would make the assessment less like instructional practices, substantial testing time could be saved, and one of the major criticisms of MSPAP (the testing takes too much time, especially since individual scores are not available) would be reduced, if not eliminated.

There is also a tension currently between giving and not giving individual scores. A good instructional model involves providing feedback to students on their performance in a timely fashion. The Department will make available individual scores but does not recommend their use. We do not recommend them either with the current design, and our reasons are described later in the report. If it were not for the pressure to provide individual scores, perhaps the state would not make them available at all. But in bowing to some of this pressure, the possibility of responding to another major criticism, the one about excessive test time, cannot be addressed. For example, another design might be to ask students to complete four tasks rather than six or seven. Then five clusters instead of three would be used. There would be no reduction in the content coverage. The only change would be fewer students in each school taking each of the clusters. This would cut test time by more than 33%, i.e., less than six hours compared to the original nine hours. It is doubtful that the quality of the school data would be very much affected

and content coverage would not be reduced. But what would be lost for sure is the opportunity to provide individual student scores. They have questionable value based on six tasks. They have even less value based on four tasks. But were it not for the need to make scores available when they are demanded, a different design might be introduced that could reduce testing test considerably. At the same time, we do recognize the downside to having shorter tests. Shorter tests might, for example, reduce the amount of integration across subjects that is possible, and the presence of five clusters instead of the current three could be problematic in smaller schools.

Another assessment design might be to remain with the nine hours of assessment time but introduce more tasks and proceed with four clusters or forms rather than three. Costs of development and scoring of tasks may increase, but the primary purpose of accountability might be well served with even better content coverage. It could be argued that 20 tasks are not many to cover the application of knowledge and problem-solving skills in six school subjects.

The injection of some multiple-choice items might also allow for better content coverage without sacrificing the goal to focus on the application of knowledge. Certainly multiple-choice items are capable of doing more than assessing factual knowledge. A review of multiple-choice items on many high quality tests will reveal this fact. The purpose of accountability might also be served by more content coverage. But multiple-choice items would be perceived as a violation of the goal of modeling good instruction. There is always going to be a tension in assessment design between efficiency, available test time, breadth of content coverage, etc. for an accountability system on one side and modeling good instruction on the other.

Using tasks in the assessment that require students to demonstrate skills in multiple subject areas, again, probably models good instruction, and may even increase the generalizability of the scores as predictors of later out-of-school performance, but this approach to assessment does appear to make it more difficult to provide pure measures of student and school performance in each subject area. For one, writing skills, become relevant in the assessment of each subject area, and

for another, the subject matter scores are likely to be more highly correlated than they might be were independent assessments of each subject area available. The validity question might be: would the assessment of mathematics skills be different if they were measured separately from other school subjects? And, what are the implications for interpreting subject matter performance in the accountability system?

It seems clear that currently the Maryland Department of Education and its contractor are trying to build a valid school accountability system while ensuring that the tasks model good instruction. The consequence is that to attain what is believed to be satisfactory content coverage, while committing substantial amounts of time to non-assessment-related activities, nine hours of testing time are needed. Were the single focus to be a valid school accountability system, it is likely that testing time might be substantially reduced or testing time could remain the same, with new and potentially more useful assessment designs.

In conclusion, our view of this issue is that it is not possible for a single assessment program to be optimal as an accountability system and as a system for modeling instruction. The imposition of time constraints on the completion of tasks, that the tasks themselves are imposed on the students, subtle pressure or perhaps even not so subtle pressure on students to perform well, introduce some dimensions that may not be present in the typical classroom situation. More importantly, the focus in assessment on tasks that can model instruction, places constraints on the assessment design that reduce its usefulness in an accountability system.

4. Review of Research on MSPAP's Impact on High Poverty Schools

The psychometric panel was asked to address two questions related to the effects of certain characteristics of MSPAP on the probable performance of students in high poverty schools. In particular, the following specific issues were to be addressed:

1. Whether or not there is any probable credence to a hypothesis that poor schools reporting a high teacher transfer rate are likely to do less well because the teachers lack familiarity with MSPAP tasks which have appeared previously, any more so than any other type of assessment.
2. Whether or not there is any probable credence to a theory that poor students are any less likely than middle class students to take a test seriously that does not report individual scoring.

These issues are covered in depth in a paper prepared by Dr. Lisa Smith contained in Appendix B. In this section of the report, a synopsis of the findings prepared by Dr. Smith is presented.

Issue #1: Are poor schools reporting a high teacher transfer rate likely to do less well because the teachers lack familiarity with MSPAP tasks which have appeared previously, any more so than any other type of assessment?

The answer to this question is YES. First, it is well documented that lower SES schools experience greater teacher mobility. In addition, teachers who are new to poor schools tend to be inexperienced teachers, not ones that transfer to the school within the district or to other districts. Maryland does not require assessment training as a component of teacher certification, so new teachers in Maryland will likely have less preparation in assessment than teachers in states providing assessment training in teacher education programs. Therefore, new Maryland teachers most likely will be unfamiliar with performance assessments, and unfamiliar with MSPAP. However, the long history of the MSPAP program may help with this concern. Literature has shown that teacher familiarity with the tasks of an assessment is an important factor in student performance, and is likely to be even more critical for MSPAP because it has tasks that are novel and complex. Thus, the impact on school performance of higher teacher turnover rates in poorer schools will likely be greater for MSPAP than it would be for assessments composed of

multiple-choice questions. Poor performance on MSPAP by students, though, could also be related to poorer teaching skills of these inexperienced teachers.

Issue #2: Are poor students less likely than middle class students to take a test seriously that does not report individual scoring?

The answer to this question is PROBABLY YES. The literature supports a link between consequences of the test for individual students and the level of effort the student puts forth on the test. The research supporting a general relationship between motivation and performance is extensive, but there is less direct evidence regarding the linkage between consequences of a test and performance on the test. In experimental conditions, evidence exists of a strong linkage between consequences of a test and student performance. However, none of these studies were done in such a way that generalizes easily to a statewide assessment program. In some cases, the consequences in the experimental studies were inducements (in terms of monetary prizes and awards). The consequences of student performance on a statewide assessment may not have the same impact on performance as the inducements provided in these experimental studies.

However, the evidence is sufficiently pervasive across these experimental studies to strongly suggest that when there are no direct student consequences, student motivation will likely be lower, particularly for secondary students, than when some student consequences exist. When no individual scores are reported, it is clear that individual student consequences are non-existent. It is not clear to what extent, however, reporting of individual scores will be translated into motivation for students to perform well on the test.

Whether test consequence, in this case the reporting of individual scores, would have a greater motivational influence for poor students as compared to middle class students is again open to question. Studies have shown a relationship between poverty and a variety of student attitudes, motivation, and performance. In addition, racial differences have been noted on the effort

students are willing to expend on tests that do not have direct consequences. African-American students reported less effort than whites on open-ended questions (there were no significant differences in reported effort expended for multiple-choice questions). In another study, urban and suburban student performance was compared on open-ended tasks. Teachers in the suburban schools reported that 100% of their students were actively engaged in the open-ended tasks on the test, whereas teachers in the urban schools reported only 37% of their students showed similar engagement.

Although not particularly strong, the literature does show a probable link between the reporting of student scores and student effort on a test, and this effect seems to be greater in poorer schools with assessments that are composed of open-ended questions. These results held also for African American students as compared to white students.

Summary

Research evidence in general shows a probable relationship between teacher turnover and the performance of students in poorer schools on assessments containing novel and complex tasks, like those comprising MSPAP. These teachers are likely to be less experienced than teachers in more affluent schools. Because they are new to the system they will likely have less experience with MSPAP type tasks. Research indicates that teacher familiarity with assessment tasks, particularly performance type assessments, is a critical factor in student performance.

Research also indicates that students probably do not take as seriously a test that does not have direct personal consequences as they would a test with consequences. There is limited evidence regarding how this relationship might be affected with students from lower SES schools. However, there is evidence to support the contention that students in urban schools, African-American students, and students in poverty schools are likely to expend less effort on open-ended tasks as do middle class students. Taken together, these studies support the hypothesis

that students from poor schools will likely take a test less seriously that does not report individual scores than will students in middle class schools.

3. Choice of Grade Levels for MSPAP

This section provides a brief description of MSPAP and contrasts the grade levels at which tests are administered for MSPAP with similar programs in other states. A discussion of reasons for selecting different grade levels is provided along with comments about MSPAP in terms of these reasons. The Test Standards (AERA, APA, NCME, 1999) provide no specific guidance to the issue of which ages or grades are appropriate for large scale assessment programs other than the general admonition that the test should be appropriate for its designed use.

The purposes of MSPAP are school accountability and instructional improvement. MSPAP represents one component of a school accountability program. MSPAP is administered to students in grades 3, 5, and 8. The rationale for selecting these particular grades was not contained in the program information provided.

Most other states also have assessment programs that are used for school accountability purposes. The other states administer either norm-referenced or criterion-referenced tests at a variety of grade levels in a variety of content areas. A summary of the grade levels and subject areas tested as part of their school or district accountability programs is shown in Table 1. The source of these data is the Council of Chief State School Officers (1997). As can be seen in Table 1 (which follows the references), tests are administered at many different combinations of grade levels and content areas. Most states administer their assessments at elementary and middle school levels for school or district accountability or to provide data for instructional improvement. Some states also include one or more high school grades in their testing program. MSPAP is similar to other states with testing at selected elementary and middle school grades. It does not include a high school testing component. However, there is a high school testing component in the Maryland School Performance Program (the Maryland Functional Test).

Reasons for selecting grade levels for testing

There are at least four different arguments that might be made for selecting the grade levels for testing. These arguments lead to testing at every grade, testing at selected grades based on

organizational or developmental subdivisions, or testing at selected grades based on curricular or instructional subdivisions, or making an arbitrary decision about which grade levels to test.

a. Testing at every grade

The first argument is that instruction occurs at all grades, therefore information needed to improve instruction is relevant to all grades, thus testing should occur at all grades (or a reasonable span of contiguous grades such as 3 – 8). Selecting this option would provide the greatest amount of information on the status of student achievement and on the broad nature of instructional strengths and weaknesses. The advantages of this option for MSPAP would be that all teachers would be accountable for student performance, not just the teachers at the specific grades (i.e., 3, 5, and 8) in which the assessment is administered. Moreover, if instructional improvements are needed the corresponding point in time in the education system where the change is needed could be identified with some degree of accuracy. However, because MSPAP is not designed to provide substantive information on the extent that individual students are achieving the Maryland Learning Outcomes, little information related to providing instructional feedback for individual students is available whether the assessment is administered at every grade or just selected grades.

The disadvantages of assessing students in every grade are many. Assuming the MSPAP program is essentially unchanged the logistics and costs associated with designing, administering, and scoring the assessment would increase to almost impossible levels. The integrated nature of many of the assessment tasks would make test development (including field testing) much more complex and time consuming. Just finding and training enough teachers to develop tasks and to score them after the operational assessment would be a difficult task. Organizing the schools for virtually a full week of testing at all grades (even a smaller span of grades like grades 3-8) would be a monumental task. Hiring enough teachers to do the scoring within a reasonable time frame (and returning the scores to the schools within a reasonable time frame) would also represent a serious ongoing challenge. Finally, the costs associated with test

development, printing, distributing, and scoring would be significantly higher. In short, testing at every grade (or even a selected span of grades) would not be cost effective for the amount of information gained. There are states that test at every grade (or at every grade within a selected grade span), but these states are using commercially available norm-referenced tests that are substantially cheaper to administer and score than are the MSPAP tests.

b. Selecting only certain grades for assessment

If assessing in every grade is ruled out, one reason for deciding to assess in selected grades may be based on logical breaks in the structure of schooling. In this case, the decision of when to test is based on the rationale that there are certain logical organizational break points as students move through the various level of schooling. For example, in many school settings the move from grade 5 to grade 6 represents a shift from the "upper elementary" grades to the middle or junior high grades and this shift reflects an organizational change from mostly self-contained classes to departmentalized classes. A second reason for the selected grades for testing is based on logical curricular break points. These break points represent substantive changes in the curriculum that may or may not be the same as the organizational break points. For example, as students move from grade 3 to grade 4 the focus in reading often shifts from "learning to read" to "reading to learn." Finally, the reason for the choice of grade level may be almost purely arbitrary. That is, grade levels are selected because some grades must be tested and the choice is made on the basis of convenience, tradition, availability of personnel for test development, or other non-educational reasons. Each of these arguments is discussed more fully below in terms MSPAP.

Organizational or developmental break points. Although there are many different organizational patterns used in schools, certain assumptions are often made about the overarching structure of K-12 organizational patterns. One illustration of traditional organizational patterns is summarized below.

For many years the earliest elementary grades were characterized as being the "primary" grades. These grades are used to provide instruction principally in self-contained classrooms with emphasis on teaching the basic skills. Students in these primary grades typically are evaluated informally by the teacher. As students progress through these primary grades more complex content is introduced and greater student independence is expected.

Grade 4 through 6 or 7 represent a shift to even greater independence by the learner and to an expansion of subject content to include more science and social studies, more complex mathematics concepts (long division, fractions), and moving from learning to read to reading to learn. The curriculum is spiraled so that the same concepts are introduced in each grade, but with introductory concepts receiving less attention and more advanced concepts receiving more attention as the grade level increases. Teaching is still largely in self-contained classrooms, or, if students move to a location other than their "home" classroom, all students in the same classroom move as a group.

At grade 6 or 7 (depending on whether the system is a junior high system or a middle school system) even more learner independence is introduced. Teaching is largely departmentalized and students do not stay with the same instructional group throughout the school day. Students are given limited choices in selecting their educational experiences and learners assume even greater independence and responsibility for their own learning. The curriculum is still spiraled, but material is presented at a quicker pace.

Finally, students move into the high school at grade 9 or 10. At this point students are substantially more independent and responsible for their own learning. Teaching is completely departmentalized. The curriculum becomes more focused and essentially no spiraling occurs. Material is presented at a substantially faster pace.

The choice of grade levels to include in a statewide assessment program used for accountability may be based on patterns like the one summarized above or on other predominant organizational patterns found in a state. The utility of selecting either the grade before or after a break point (e.g., grade 3 or 4) may be to provide data that could serve as a "summary" of what has transpired educationally in the grades leading up to the transition grade. If students are not demonstrating the knowledge expected of students at that grade, then the assessment results would provide insights about instructional strengths and weaknesses. The grades assessed in MSPAP reflect such organizational break points in that students are assessed at the end of the primary grades (grade 3), at the end of the upper elementary grades (grade 5), and at the end of the middle school grades (grade 8).

Curricular or instructional break points. Curriculum is not designed to always be congruent with the organizational breaks described above. An examination of the typical school curriculum will show that different content areas typically introduce new content or shifts in the way the previous content was used or taught at different grade levels. For example, a shift in reading instruction often occurs at the beginning of grade 4, when many teachers no longer teach basic reading skills, per se, but require students to use reading in the subject areas as a way to learn the subject area. In mathematics, such a shift in content typically occurs near the end of grade 4 when concepts of long division and conducting operations with fractions and decimals become a dominant focus. In science grade 5 is often a transition point. Schools using different instructional materials (from different publishers) may have different curricular transition points. With the use of spiraled curriculum materials many concepts are introduced and reintroduced many times between grades 4 and 8.

Because of the variation in curricular transitions across the different content areas, if the objective of a statewide assessment program were to reflect on students' readiness to move through these transitions, then it would be logical to assess different content areas at different grade levels. The decision of MSPAP to assess at grades 3, 5, and 8 in all content areas will

result in assessing some of the curriculum transitions at the point in time just prior to when they happen. Other transitions will be assessed in the grade in which they happen, or in the grade following the transition. From a purely accountability perspective, the choices of grades 3, 5, and 8 may work well. From the perspective of making instructional decisions assessing at grades 3, 5, and 8 may be less than optimal. Choosing more optimal grade levels, however, may create substantial logistic and administrative confusion because of the number of different grade level and content area combinations that would be required.

Arbitrary selection of grades. This method of determining the grade levels to be tested has many advantages if the program is strictly for purposes of accountability. Some restrictions on the choice of grade level are applicable, but beyond those restrictions the choices of which grade levels to test are wide open. The restrictions are related to such things as a) the reasonableness of administering tests to very young children (e.g., under age 8), b) the amount of testing time that might be required in terms of students' attention span, c) the efficacy of collecting useful data from unmotivated students (especially high school students), and d) the timing of test administration in terms of students' opportunity to learn (e.g., testing knowledge of dividing fractions at the end of grade 3 when most students are not exposed to this until the middle of grade 4).

Because MSPAP is administered in grades 3, 5, and 8 some of the restrictions have been avoided (testing very young children). Moreover, because MSPAP is designed to focus on Maryland's Learning Outcomes that are grade-specific, the opportunity to learn restriction may be of limited relevance.

Summary, Conclusions, and Recommendations

It is suggested that the most useful assessment program is one that tests at virtually all grades (or a reasonable span of contiguous grades such as 3-8). The disadvantages of this strategy, however, for MSPAP are that such a program would be extremely unwieldy, extremely

expensive, and would require more energy from the teachers for development, administration and scoring. The time required for scoring and obtaining feedback would likely be increased over the current program.

Because it is likely that the different organizational structures vary greatly across the state, and because the instructional and curricular variations may also vary from district to district and school to school, to the extent that grades 3, 5, and 8 represent break points in either organization or curriculum, these are reasonable grades in which to assess.

Conclusions. The choice of testing in grades 3, 5, and 8 for MSPAP is reasonable from the perspective of providing data on school accountability. In terms of providing information that may be useful in improving instruction, the choice of grades also may be reasonable, but it may not be optimal.

Recommendations. We recommend that some justification for selecting grades 3, 5,and 8 be elaborated. The justification should indicate the extent that the content of the assessment is reflective of the curriculum of the grade level in which the assessment is administered.

4. Test Development Procedures

According to the Test Standards (AERA, APA, NCME, 1999) there are four basic components of test development:

1. Delineation of the purpose(s) of the test and the scope of the construct or the extent of the domain to be measured;
2. Development and evaluation of the test specifications;
3. Development, field testing, evaluation, and selection of the items and scoring guides and procedures; and
4. Assembly and evaluation of the test for operational use.

In this section, each of these components is addressed sequentially. For each component, the relevant test development features and procedures are summarized followed by an evaluation of how well this component meets the 1999 Test Standards.

1. Purpose of MSPAP

MSPAP is designed to provide information for instructional improvement and for school and district accountability. The assessment requires students to construct responses to single content area and interdisciplinary tasks that call for the application of basic skills and knowledge to real life problems. The content areas include reading, writing, language usage, mathematics, science, and social studies.

The content of the tests is tied to the Maryland Learning Outcomes that were approved by the State Board of Education in 1990. Not all of the Maryland Learning Outcomes for these content areas are assessed by the MSPAP, but those that are assessed are delineated in the test documentation and test development guides. The goal of the assessment is to measure critical

thinking, higher-order reasoning, and problem-solving skills. This is accomplished by asking students to respond to questions or directions that lead to the solution of a problem, a recommendation or decision, or an explanation or rationale for their responses.

Evaluation of MSPAP on Delineation of Purpose. Test Standard 3.2 states that the purposes of the test, definitions of the domain, and the test specifications should be stated. The philosophy of MSPAP, to provide school instruction and accountability information through the use of integrated performance assessments that measure application of knowledge and skills, is clearly stated in program reports and documentation. Matrix sampling is used so that, overall, the results provide information on the set of content learning outcomes that are specified to be assessed through MSPAP. There is some question about the appropriateness of the Maryland Learning Outcomes, but that is not pertinent to this part of the evaluation (and is being addressed by the Content Review Committee).

2. Test Specifications

According to the test specifications, tasks are to be developed by Maryland teachers. Teachers work in content/grade level teams and are assigned specific learning outcomes to be assessed by the tasks they construct. Tasks are then assigned to integrated clusters.

A total of three "clusters" of tasks are developed for each of the grade levels, Clusters A, B, and C. The content of these clusters is prescribed directly in the Specification and Procedures Manual for Test and Task Design and Development (MSDE, 1999). Clear detail is provided regarding the learning outcomes that are to be addressed and how integrated tasks should be constructed. The exact number of questions per cluster or task is not specified, nor are the desired psychometric properties. However, a minimum number of measures per outcome is specified in order for outcome scores to be reported. The total time for testing is 9 hours spread over 5 half-days. In the assembly of the tasks into clusters, guidelines are presented that pertain to the distribution of hands-on tasks, short and long writing tasks, and sequence of tasks. A

separate document details test administration and directions. Directions for examinees are included in the student booklets and may be read aloud by the administrator (depending on the grade level).

Some tasks require the use of tools and/or manipulatives. One step in the test development process is to consider the reasonableness (including costs) of the proposed tools and manipulatives. Six-member Tools and Manipulatives Feasibility Committees for each grade level meet to review the tools and manipulatives and how they are being used in the assessment. These committees meet yearly to review the materials lists for all tasks. Materials are also reviewed during the peer review of task development, during the field test endorsement meeting, and during the final endorsement meetings.

Specifications are also provided for scoring of the tasks and are detailed in the Specifications and Procedures Manual for Test and Task Design and Development. Score descriptors are prepared by the teacher teams during task development. Early in task development, these teachers identify the outcomes they believe are being assessed. These are expected to mirror those outcomes assigned to the test development team. After the tasks have been drafted they are subjected to a scorability review where each activity is examined to determine if it contains any miscuing or cross-cueing. Also, as part of the test development process, all draft tasks and preliminary scoring information receive a formative review. Scoring experts, administrative experts, and content representatives identify problems, revisions, etc. that need to be addressed by the task development team. Finally, a summative review is conducted following these revisions.

Current specifications detail that the operational test will be made up of no more than 50% scorable units from new tasks. By 2000, the test will be composed of one-third new, one-third banked (used operationally two or more times; virtually unchanged from a previous administrations), and one-third rollover tasks (used operationally only once previously; may be

modified slightly or not from previous administration), with the additional specification that reused tasks must rest at least 2 years.

Evaluation of MSPAP Test/Task Specification. Test Standard 3.3 states the test specifications should be documented, along with their rational and the process by which they were developed. Much thought and care appears to have guided the development and documentation of the test/task specifications. The document Specifications and Procedures for Test and Task Design and Development is clear and thorough, providing excellent documentation of test and task specifications. Standard 3.3 also stipulates that the test specifications define the proposed number of items, the item formats, the desired psychometric properties of the items, and the item and section arrangements. Although the specifications do not give the specific numbers of items, they do give the minimum number of items needed in order to produce student-level Outcome Scores. The amount of time devoted to testing is specified and procedures are delineated for test administration and scoring. An external review by relevant experts, as called for in Standard 3.5, focusing on the test specifications is desirable. As will be pointed out in the validity section, there is evidence that some of the components stipulated in the test specifications are not being carried out. The external review may be able to pinpoint areas of weakness in the test specifications and their implementation.

3. Test/task Development, Field Testing, Selection of Operational Tasks and Scoring Guides

Test development follows a 24-month cycle, beginning with the development of the overall design for the assessment and concluding with the final endorsement of the tasks and scoring tools.

As noted above, teams of Maryland teachers develop tasks. Over 100 grade and content area teachers are recruited and trained for test/task development. These teacher teams are assigned learning outcomes for their task development.

The Specifications and Procedures Manual for Test and Task Design and Development gives the specifications and procedures to be used in the development and evaluation of tasks. At several points in the development process, reviews are undertaken. Prior to field testing, the tasks and scoring tools will have been reviewed at least twice (formative and summative reviews) by MSDE staff, administrators, and content representatives. Additionally, there are reviews for Bias and Sensitivity and Developmental Appropriateness. The Manual does not specify who should participate in the Bias and Sensitivity Review. The Developmental Appropriateness Review occurs twice during each task development cycle, once after the content review and again prior to the field test. Teams of grade level appropriate teachers and non-classroom professionals who have expertise in cognitive and psychological development review grade level clusters. The reviews are guided by a set of "Guidelines and Checklist for Creating Developmentally Appropriate Tasks."

A Field Test is conducted to collect information on the feasibility of administering the tasks in a classroom setting, clarity of the directions, utility of the tools and materials, appropriateness of task timing, and scorability of tasks. For the 1998 MSPAP, schools in the Inter-borough School District in southeastern Pennsylvania served as the field-test site. These schools were chosen because their student populations were deemed to be a close match to Maryland's population and because they used collaborative learning and hands-on tasks as part of their instructional programs. All new tasks were administered to two classrooms of 25-30 students. As a result of

the field test, some tasks were modified slightly to adjust timing, clarify directions, or address confusing questions. Following the revisions, a post field-test review was conducted to assure that the tasks were now ready for operational use. In previous years, the field test was conducted with a representative sample of Maryland students in the next higher grade (4^{th}, 6^{th}, and 9^{th}).

Evaluation of Test Development and Field Test. Several reviews are conducted during the task development process that identify areas where changes are needed. It is not clear how these review committees document their recommended changes. There are follow-up reviews, suggesting that issues and concerns raised earlier are accounted for in the subsequent review. The names of the reviewers were not given in the materials provided, however, in most cases their qualifications were stipulated (except for the Bias and Sensitivity Review Committee). For the most part, these procedures are consistent with Standard 3.7 which states that the procedures used to develop and review items should be documented.

Field tests are conducted to gather information about the appropriateness of task time allocation, clarity of directions, and evidence of the completeness of the scoring criteria. As indicated before, prior to 1998, out of grade field tests were conducted by using 4^{th}, 6^{th}, and 9^{th} grade students to tryout tasks developed for students in 3^{rd}, 5^{th}, and 8^{th} grades, respectively. This procedure had the strength of retaining Maryland's curricular reform philosophy which emphasizes constructive learning and cooperative groups, for example, but had the weaknesses of less than perfect grade level alignment and the potential for less the optimal control over task security.

In 1998, field tests were conducted by administering the tasks in a limited number of classes in Pennsylvania. The Pennsylvania students who participated in the field test were represented as matching the race/gender distributions in Maryland's schools, but specific percentages were not provided. Although it was documented that there was congruent instructional practices involving cooperative learning teams and hands-on activities, no information was provided regarding the reasons given to the students for why they were taking these tasks. No information was provided regarding the motivation of these students for performing well on these tasks. The results from the operational administration in Maryland are not used for student decisions, but they are used for school accountability purposes. There is likely to be some level of motivation for the students to perform well, even though there are no direct ramifications for individual students based on their scores. No information was provided regarding motivation of students in the Field Test. The degree to which there is a difference in motivation levels between the field test students and the students in the operational administration is open to speculation. Standard 3.8 states that for item tryouts and field tests, procedures used to select the sample(s) of test takers and the resulting characteristics of the sample(s) should be documented. When appropriate, the sample(s) should be as representative as possible of the population for which the test is intended.

4. Assembly and Evaluation of the Test for Operational Use

Following the post field-test review, tests are prepared for operational use. Three assessment books are used in the operational assessment: Examiner's Manual, Answer Book, and Resource Book. All scorable responses are found in the Answer Book. The Resource Book contains all readings, graphs, illustrations, and tables that the students use during the assessment. The Examiner's Manual provides detailed directions for administration. Tasks are often introduced

by pre-assessment activities, designed to orient the students to the task. Sometimes these activities are teacher-directed, where the teacher elicits ideas from the students; other times these activities are group centered where groups of 3-5 students engage in an activity together. These groups of students are to be formed randomly in advance of the test administration. None of these pre-assessment activities are scored. However, test administrators may be more or less skilled in the delivery of these pre-assessment tasks. Therefore, quality and usefulness of the pre-assessment activities varies across sites. Total task administration time allows for flexibility during pre-assessment activities and directions, but dictates fixed timing for scored activities. All scored activities are accomplished independently by the students.

Some of the tasks require additional tools or manipulatives and may require advanced organization and planning by the test administrator. Schools are responsible for acquisition of these materials. Most schools obtain these materials from an external provider. The teacher/administrator must be familiar with the tasks and needed materials in advance of test administration. For some tasks, preliminary actions on the part of the test administrator need to occur (e.g., cutting tape into sections).

Following each administration, a Technical Manual is produced that documents the extent to which the content domain of the test questions are congruent with the test specifications. Psychometric properties of the assessment are also documented in the technical manual.

Evaluation of Assembly and Evaluation of the Test for Operational Use. Through the use of sequenced reviews, evidence is gathered regarding the appropriateness of the tasks (consistent with Standard 3.11 which specifies that content of the test should represent the defined domain

and test specifications) and scoring tools (consistent with Standard 3.14 which states that the criteria used for scoring test takers' performance on extended-response items should be documented). Field-test results are also considered in the assembly of the operational form. There is some concern whether the results from the field test will provide good indications of student performance on the operational administration due to lack of representation of the field test sample in terms of demographics and motivation (Standard 3.8).

Consistent with Standard 3.1 which states that tests and testing programs should compile and document evidence bearing on test development, a technical manual is prepared subsequent to administration documenting the technical quality of the assessment. There is some unevenness in the delivery of the pre-assessment activities that could affect the preparedness and engagement of some groups of students in the operational assessments. This appears to be in conflict with Standard 3.19 that states that directions for test administration should be presented in sufficient clarity and emphasis so that it is possible for others to replicate adequately the administration conditions.

Conclusions and Recommendations

1. Test development appears to be a strength of MSPAP. Clear documents detail the steps for task developers to use in creating the tasks. However, there is evidence that some of the procedures, although clearly documented, may not be carried through in operation.

2. Multiple reviews are used to identify potential problems and steps are taken to make adjustments and corrections prior to field test. Tasks are reviewed for developmental appropriateness and for bias and sensitivity during test development.

Recommendations: Test Development

1. External reviews by persons outside the test development process should be conducted on the test specifications. These reviews should focus on where there may be weaknesses in the process that would allow for non-compliance with the stated test specifications and procedures.

2. Better documentation is needed of the membership of review committees, their recommendations, and changes made in response to their recommendations.

3. Better standardization is needed of the assessment materials and manipulatives, both in their acquisition and in their content. It has been suggested that teacher-prepared materials may be of uneven quality and may therefore jeopardize the standardization of the test experience for students.

4. Test administrators should be given a standardized orientation to the tasks they will be administering and the pre-assessment activities for which they are responsible. Standardization of the pre-assessment preparation is needed.

5. More information is needed about the selection of field test sites. Using sites that are as comparable as possible to Maryland school curriculum, instructional philosophy, grade level, and student motivation enhances the usefulness of the field test information.

5. Scoring

In this section, issues such as consistency of scoring, objectivity of scoring, the clarity of scoring procedures, and the time required to complete scoring are the principal focus.

The Test Standards (1999) devote substantial attention to scoring in general and scoring procedures in particular. The most relevant standards to the MSPAP scoring procedures require that a) scoring procedures and criteria should be prepared at the time of test development (3.3, 3.14, 3.22), b) scorer selection and training should be documented (3.24, 12.8) and training should insure a high degree of consistency and reliability throughout the scoring process (3.23, 3.24), c) procedures for scoring and scoring criteria are presented in sufficient detail to maximize accuracy of scoring (5.8), d) rubrics specify criteria for scoring and that scoring accuracy is monitored (5.9, 3.23), d) the frequency of scoring errors be monitored and reported (3.24), and e) the scoring procedure is well documented (most standards related to scoring require documentation). The scoring procedures employed in the MSPAP program are summarized below. As will be clear, these procedures tend to be consistent with the Test Standards.

The Scoring Process

The scoring process begins at the time of test development. Because all tasks are constructed response, rather than multiple-choice, the teachers responsible for task development develop the scoring criteria as the tasks are written. The scoring is refined after the tasks are pilot tested. After the operational administration, a sample of student papers is scored immediately by the scoring contractor to verify the scoring guides. Maryland teachers who are trained as scorers score the remaining tests. The details of the process are described in more detail below. The process is described in the 1998 Maryland School Performance Assessment Program Scoring Report (Measurement Incorporated, 1998) and the 1998 Maryland School Performance Assessment Program Technical Report (MSDE, CTB/McGraw Hill, Measurement Incorporated,

University of Maryland Baltimore County, Center for Educational Research and Development, and Westat, Inc., 1999).

When all tests are scored, reports are prepared. The main report is produced by the MSDE. The results of the May, 1998 assessment (1997-98 school year) were reported in Fall 1998. The reports included data for each student tested (returned to schools with the recommendation that it not be shared with parents), summary reports for each school, and a summary report for public consumption on the results for each school system and the state as a whole. The latter report, called the Maryland School Performance Report 1998 State and School Systems (MSDE, 1998), summarizes the background and purpose of the program. A more detailed description of the reporting is included in a later section of this report.

The scoring process involves several major (and myriad minor) steps. These major steps include developing the initial and final scoring rubrics and scoring guides, organizing the tests for scoring, selecting the sites for scoring, hiring and training Maryland teachers to serve as scorers, and conducting the scoring. Each of these steps is described briefly below.

a. The Time Frame for Scoring

A time frame for scoring is shown in the scoring report prepared by Measurement Incorporated (MI, 1998). This schedule for the 1997-98 MSPAP is assumed to be representative of the schedule for all MSPAP administrations since 1998. The schedule indicates that the scoring tools are reviewed in October of the year prior to the assessment and the draft scoring guides are prepared immediately following this review. Beginning in January, the scoring sites are selected and beginning in March the scorers, Team Scoring Coordinators, and Team Leaders are hired by MI. Between mid-May (when the tests were administered) and June, MI prepared final training materials. When training materials were completed, the Team Scoring Coordinators and Team Leaders were trained (a week-long process). Coordinators and Team Leaders then trained scorers at the scoring site to which they were assigned (a two-day process). After the tests are

administered, most completed test booklets are shipped to the CTB/McGraw Hill headquarters in CA for processing and then shipped back to Maryland to be scored. Approximately July 1, the scoring begins. Approximately 5 weeks later, initial scoring is completed and scan sheets are delivered to CTB. Between the end of the first week in August and the first week in September a process of "clean up" occurs (missing score sheets are located and scored and other problems resolved).

Because the scoring process depends on the availability of Maryland teachers who are not available until after the end of their contract period shortly after the close of school, it is difficult to see how this process could be shortened without making some major changes in the program. At most perhaps a week or two might be saved by quicker preparation of the training materials and employing more scorers to conduct the scoring process. Additional time might be saved by making the assessment tasks and items easier to score, that is by using more objectively scored questions that can be responded to in a multiple choice or short answer format. Shipping the completed booklets to CA for processing and then returning them to Maryland for scoring seems quite inefficient. During that period MI is refining the scoring guides and training materials, so eliminating the shipping and doing the processing in Maryland would not shorten the overall time frame unless the development of scoring guides and training materials was done faster. This would save, at most, a few days in the overall process.

Once the data are delivered to CTB in September, the analyses begin (e.g., equating, converting student raw scores to various scale scores, estimating school, district, and state level performance). The nature of the program and the number of analyses (and we assume, cross checks) that are undertaken are very time consuming. After the analyses are completed and verified, the various score reports are prepared. The production of the reports is also a time consuming process. Because of the importance of the program and the potential consequences to schools, releasing score information prior to the complete reports being available to schools is inappropriate.

The time frame from test administration to the release of score reports is approximately 6 months. During this time, scorers are trained, booklets are prepared for scoring, hand scoring of almost 290,000 tests is completed, data are analyzed, and reports are prepared. It may be possible to reduce the turnaround by a week, perhaps two weeks without making changes in the characteristics of the program (e.g., using more objectively scored items), but to expect greater efficiencies is to invite errors due to attempts to hurry the process. The process is described in more detail below.

b. Developing the Rubric and Scoring Guides

Scoring rubrics are designed at the time tasks are initially developed by Maryland teachers. They are refined after the tasks are field tested. At field testing, Benchmark responses (responses that reflect each score point accurately and are used to train scorers) are also selected. These procedures are described more fully in the Test Development section of this report. The rubrics and scoring guides[1] are refined again immediately following the initial administration of the task in the operational testing. At this time the final scoring guides and scorer training materials are produced by Measurement Incorporated (MI), the scoring contractor. This finalization is accomplished by having a small sample of the completed test booklets "hijacked" and sent directly to the scoring contractor for initial scoring. The hijacked test booklets are typically selected from schools that tend to be high scoring (to insure the comprehensiveness of the scoring key) and from schools that have diverse demographics (e.g., minority or LEP students).

[1] Scoring guides are produced by the scoring contractor and are used in training readers. The guide is made from a combination the scoring tools (produced at the time of test development) with actual student responses from the field-testing and hijack sampling. The guides contain all the scoring tools, answer cues, guidelines, and sample responses at each score point.

c. Organizing the Test Booklets for Scoring

Except for the test booklets that are hijacked all completed test booklets are shipped directly to CTB/McGraw Hill where the booklets are coded with student identification numbers and coded scan sheets are inserted in each booklet. Except for the coding the booklets are essentially stripped of personally identifying information (to facilitate blind scoring). The booklets from the 24 counties and Baltimore City are then organized into nine "bins." Within a bin booklets are put in random order. Bins are then shipped back to Maryland for scoring. In 1998 approximately 288,000 test booklets were scored.

d. Scoring Sites

All response booklets are scored at one of four sites. The MSPAP uses matrix sampling. Typically, the sampling process results in having three non-overlapping clusters of tasks at each grade level. Each cluster of tasks is contained in a test booklet. A cluster may consist of tasks to be scored in each of the six different subject areas within a grade level. There are nine grade/cluster combinations (three clusters, A, B, and C at each grade 3, 5, and 8). All tests from a cluster are scored at the same site. Site locations may, or may not, change annually. In 1998, site 1 (Mattawoman Middle School) scored all of clusters 3A and 8A. Site 2 (Grasonville Elementary School) scored only Cluster 5A. The remaining two sites scored Clusters 3C, 5B, and 8B, and Clusters 3C, 5C, and 8C, respectively. Teachers who serve as scorers are typically hired from the geographic area near the scoring sites.

There is another site, unspecified in the reports, where a fourth, equating, cluster is scored. This cluster is a set of tasks that were administered in both the current and the previous year and is used for purposes of year-to-year equating. Priority is given to hiring teachers who were employed as scorers in the previous year to score this cluster. Thus, the equating cluster may be scored by few inexperienced scorers. New clusters that contain newly developed items may be scored by a substantially higher percentage of inexperienced scorers.

e. Hiring and Training Scorers

Maryland teachers score all test booklets. In 1998, over 800 teachers were employed during the summer to participate in the scoring process. Many of these teachers had been employed previously as scorers, but many were first-time scorers. MI hires and pays all scorers. Some scorers receive supplemental funds from their school district.

Scoring is accomplished using a team approach. Each team has a Team Scoring Coordinator and Team Leader who are experienced scorers and competent trainers who are selected and hired by MI. These individuals are trained by MI staff and, after being trained, are responsible for training the scorers on their team (scorers are called "readers"). Teams vary in size. In 1998 the teams ranged from 13 to 29 readers. An attempt is made to employ scorers who were certified and experienced teachers in the grade level of the clusters they were scoring; however this is not always possible. It is assumed that teachers at grades 3 and 5 are qualified to teach all content areas and therefore are competent to score all areas. An attempt is made to assign scorers of the grade 8 clusters scoring responsibilities only in the content in which they specialize. Scoring training is extensive. Readers are provided scoring guides for each task. Training sets of papers are used to familiarize readers with the types of responses that define different score levels, and then readers are tested on pre-scored "qualifying sets" of tasks. Readers are trained to a criterion of 70% perfect agreement on at least one of the three qualifying sets of tasks. Readers who do not meet this criterion are dismissed.

During scoring, the scorers score sets of tasks (called "check sets" and "accuracy sets") that are used to insure that readers and teams of readers do not "drift" in their scoring. These check sets and accuracy sets are sets of tasks that were pre-scored and copied for use to verify that scorers are maintaining their scoring consistency and accuracy. In general the check sets are administered on Monday mornings. If inconsistencies or inaccuracies are detected, the check sets are used to recalibrate either the entire team or individual readers. Accuracy sets are generally administered on Tuesday and Thursday mornings to detect reader drift. If reader drift

is detected, the reader is retrained. Drift is defined as a reader having less than 70% perfect agreement with the original scores on the accuracy set of tasks. By the end of scoring, the overall average agreement on the accuracy sets is typically substantially higher than the criterion. In 1998 the range of average agreement on the accuracy sets was from a low of 71% to a high of 97%, with most values between 80% and 90%.

It is assumed, but not explicitly described in the training process, that such factors as language structure, spelling, etc. are ignored except for those items for which a language usage icon are displayed. It is assumed that handwriting is also not a factor in scoring under any circumstances as long as the handwriting is legible. Systematic inclusion of papers with poor language structure, spelling, and handwriting should be a part of the training and quality control process to insure that when content is the focus of scoring that scoring bias is not introduced by irrelevant factors.

f. Conducting the Scoring

Readers at each grade level are assigned to one of four teams for scoring. Each site has multiple teams. A team scores all the scorable units for one content area. The four teams are assigned to score Mathematics (both content and process), Science, Social Studies, or Writing/Reading/Language Usage. As noted above, there is an attempt to have members of a team score the items that are in the content area of the reader's certification or specialization. That is, an eighth grade mathematics teacher is typically assigned to score the mathematics items rather than science items.

When booklets are returned to Maryland from CTB/McGraw Hill, each one contains four score sheets coded with the student's ID, one sheet for each team. Readers complete these scannable score sheets for the content area they score. When a booklet is completely scored, the score sheets are examined to ensure completeness. The completed score sheets are sent to MI's home office for scanning. Various accuracy checks are also undertaken at the time of scanning.

Quality control strategies include the use of check sets and accuracy sets of tasks that are described above. In addition, Scoring Coordinators and Team Leaders spot-check (rescore) papers periodically. After the score sheets are verified at the scoring site, they are sent to MI for scanning where additional checks take place.

When all score sheets have been scanned and any coding or scoring errors resolved, the data are sent to CTB/McGraw Hill. CTB/McGraw Hill completes the final scoring processes [described more fully in the 1998 Technical Report (MSDE, et al. 1999)]. These processes include converting student raw scores to scale scores using an IRT two-parameter partial credit model.

Because the administration of the assessment takes place over a five-day period and some students may be absent for part of the time, an algorithmic strategy is applied to estimate scale scores for some students who do not complete the full assessment. If a student has completed 60% or more of the responses in a content area and a minimum of eight independent measures (the criteria are different for writing and language usage or other "short tests" like mathematics process), then algorithmic scoring is performed to estimate a student's ability level. In 1998, algorithmic scoring was undertaken for approximately 15, 000 students.

Once final scale scores in each of the six content areas have been determined for all students, ability estimates are made for students for the Learning Outcomes[2]. The student outcome scores are projected as if a student had taken all items that measure an outcome. Once student scale and outcome scores are produced, several other scores are computed. These scores are a) scale scores for schools in the six content areas, these scores range from about 350 to about 700, with a mean of about 500 and a standard deviation of about 50; and b) school outcome scores. There are two types of outcome scores: Outcome Scores and Outcome Scale Scores.

[2] Students take at most one-third of the test thus not all students take all items that are associated with an outcome. At least four measures of the outcome have to be present on the test form for a student outcome score to be computed. Student outcome scores are aggregated to produce school outcome scores.

The Outcome Scores are intended to indicate the proportion of the knowledge and skills associated with a Maryland Learning Outcome (MSDE, et al., 1998) that have been mastered. They are not comparable across content areas or years due to differences in difficulty across content areas and differences in content coverage within a content area from year-to-year. The intended use of these scores is by school improvement teams to aid in assessing a school's relative instructional strengths and weaknesses.

The Outcome Scale Scores are on the same scale as the Scale Scores (ranging from 350 to 700) and thus are directly comparable across outcomes in the same content area across years. They are also tied to the MSPAP Proficiency Levels in each content area. When all scoring has been completed and scale scores and outcome scores developed, score reports are prepared.

Summary and Conclusions Related to the Scoring Process

As the above summary of procedures makes clear, the MSPAP scoring process is, for the most part, consistent with the Test Standards. There are, however, several criticisms that may be offered. These criticisms include the basis for estimating student ability for Maryland Learning Outcomes. Ability estimates may be based on student responses to a cluster that includes less than one-third of the outcomes. Moreover, the tasks within clusters are not parallel to tasks in other assessment clusters, but that are related to the same outcome. An additional potential criticism is that the only systematic scoring quality control checks for drift occur at regularly scheduled times all of which are in the mornings when scorers may be fresher. Finally, the description of the process, as described is consistent with the standards, however, there is no systematic independent review to assure that these procedures are actually translated into good practice.

In addition to evaluating the scoring process in terms of its consistency with the Test Standards (1999), questions about the consistency, objectivity, and clarity of the scoring guides and

processes have been raised. Other questions relate to the relationship between scoring for correctness of response versus scoring for correctness of process (particularly in mathematics), the extent that testing reinforces classroom instruction of writing mechanics and structure (e.g., scoring for grammar, spelling, and punctuation), and the level of objectivity in the scoring guides relative to political or cultural bias.

Consistency, Objectivity, and Clarity of Scoring Guides and Process

Some of these issues have been addressed above. Specifically, the consistency of scoring is such that readers must meet standards of performance that require 70% exact agreement with pre-scored responses that are used to audit reader accuracy. On average, this standard is met or exceeded for readers at all grade levels and content areas. This level of scoring accuracy implies that training in the scoring process results in a high degree of scorer agreement. Such a high degree of agreement implies at least a moderate degree of consensus on the meaning and interpretation of scores. Because each paper is scored by only a single reader (except for papers that are scored a second time by Scoring Coordinators or Team Leaders as part of the quality control procedures used to verify scorer accuracy) some slippage in scorer consistency may be difficult to detect and may result in a "halo effect." Moreover, it is possible that the reported degree of scorer accuracy may be inflated because the schedule of quality control checks is known and is performed only on certain mornings. Finally, the clarity of the scoring guides and processes, by inference, seems to be high. Extensive efforts are made by the scoring contractor and MSDE to insure the clarity and accuracy of the materials used for both training all the Maryland teachers who participate in the scoring. These efforts seem to result in the production of clear and effective materials used in the scoring process.

These issues of consistency and accuracy of scores are addressed by Koretz (1997). He stipulates that consistency is a necessary but not sufficient condition for validity. He suggests that readers may show a consistent bias against examinees whose handwriting is poor. Should

such a condition occur, the scores would be consistent, but would be a source of construct irrelevancy unless the score was not intended to reflect on the quality of the handwriting. Only a careful review of the scoring guides (including the rubrics, Benchmark papers, and other relevant scoring materials) and oversight of the scoring process may reveal the extent that the scoring is objective, accurate, and valid.

The Content Review Panel[3] indicated that they examined all items and scoring guides that have been used over the past five years. They reported to us that they found many rubrics to be vague, to the extent that similar answers may receive very different scores. They said that many rubrics contained the phrase "constructing meaning from…" but they contend that this notion is vague and that the cues used to assess this phrase seem to be valued more than the correct answer. They observed cases in which constructing-meaning indicators such as student's responding "I think…" or "I feel…" are scored higher than are responses that contain more substantive (and correct) answers.

The Content Review Panel also looked at the match between the rubric and the scoring process by examining Benchmark papers used to define various score points. The panel noted several instances of inconsistency between the rubric and the Benchmark papers that were used to determine how papers were actually scored. The panel members provided several examples of these inconsistencies, especially in the areas of Social Studies and Science. Other than the Content Panels' report of potential problems associated with mismatches between the rubric and the actual scoring process, there are no specific data reported that could be used to make direct inferences regarding objectivity. Such data need to be collected and examined routinely during the scoring process.

[3] A panel of psychometric specialists prepared this section of the report. There was also a panel of content experts that provided insights about other aspects of the MSPAP program. In January, both panels met together for a short period to help provide some insights to each other's areas of responsibility.

i. Process Versus Correctness

The practice of scoring for process versus correctness is not unique to the MSPAP. However, the MSPSP test-development specifications for mathematics require that each task will include at least three outcomes (combining process and content) and at least one item asking for a process to be explained. Thus, both correctness of response and the process used to obtain the response are considered in scoring mathematics tasks. The relative weighting of the correctness of the answer versus the correctness of the process is not specified and may vary from task to task and from year to year.

j. Congruence Between the Assessment Process and Classroom Practice

The correspondence between classroom practice and scoring of responses for correctness of grammar, punctuation, and spelling is not clear. First, it is not known what classroom practice is. The MSPAP practices are clear. Some tasks are scored explicitly for language usage (e.g., grammar, punctuation and spelling) and others are not. A language usage icon denotes items that are scored for language usage. Students are advised that all such items will be scored for both content and language usage. The testing process seems to make available such items as dictionaries, calculators, and other aids that are routinely available in the classroom for all items including those that are scored for language usage.

k. Taking Language Use into Account for All Responses

The practice of not scoring all items for proper language usage is an issue related to the validity of the score interpretation. The validity concern is that if a response is scored for more than one construct then score interpretation is problematic. For example, if one wants to know about a science outcome, then the correctness of the science response should not be contaminated with information regarding the correctness or incorrectness of language usage. If such contamination were present, then it would not be possible to know how to interpret a low score. Low scores might reflect perfect language usage and substantial lack of science knowledge, or they may

reflect a high degree of science knowledge, but a poor use of language to express that knowledge. Such ambiguity is clearly a threat to the validity of the interpretation of the score.

Another consideration is that because all items require some sort of written response, if all responses were scored on language usage, then more test data related to language usage would be collected than is needed to obtain an accurate score for students in that content area. Moreover, if all items required the use of complete sentences, correct punctuation and spelling, and correct grammar, then it would take students even longer to complete the test, suggesting the potential for lower scores due to fatigue and resentment from teachers for using even more instructional time for testing. A final consideration is that, if every response was scored separately for language usage, then the scoring process would take longer and be substantially more expensive.

Thus, it is not appropriate to score every item for language usage as that would represent a substantial overweighting of those outcomes in terms of the importance associated with those outcomes. However, if it is decided that language usage should be considered for every task, then that element of the student's answer should be scored separately from the substantive content the task is intended to assess. To do otherwise would compromise the validity of the score.

l. Objectivity of Scoring Guides

The final issue, related to the objectivity of the scoring guides in terms of political or cultural bias, is partly covered in the test development process. Both tasks and scoring rubrics are subjected to a bias and sensitivity review at the time they are developed. These bias reviews are specified in the Specifications and Procedures Manual for Test and Task Design and Development (MSDE, 1999). To determine if the scoring guides were less than objective and contained the possibility of political or cultural bias, they would have to be examined individually by content experts who were trained to identify such problems. All bias and sensitivity reviews are undertaken under the auspices of the MSDE staff.

Overall Summary, Conclusions, and Recommendations

The scoring process begins with the development of tasks and items. The time line for scoring includes the development of initial scoring guides, scoring site selection, hiring Scoring Team Coordinators, Team Leaders and scorers, training scorers, preparing response booklets for scoring, scoring, shipping data to CTB and back to Maryland, developing the final student, school, district, and state scale scores and reports. Once the personnel are hired, this process takes about six months to complete. Reducing by more than a week or two the time required to complete all tasks under the current program is unlikely. Time savings could be accrued by modifying the assessment program to include more items that could be scored more quickly (e.g., multiple choice items or very short answer items).

Conclusions

The scoring process in general is consistent with the Test Standards (1999). However, there are some concerns that need to be addressed to provide evidence of scorer consistency, accuracy, and objectivity:

1. The amount of time between the administration of the MSPAP (mid May) and the release of score reports (late fall) seems consistent with the level of work required. It may be possible to shorten the time by a minimal amount (one-two weeks) without making substantial changes in the nature of the program, however, to try to shorten the time line by more than a week or two might introduce the possibility of errors in reporting that would be damaging to the program.

2. The various reports of the procedures suggest a high degree of compliance with the Test Standards. However, these descriptions do not provide assurance that the procedures were carried out in ways that resulted in high levels of consistency, accuracy, or objectivity of the resulting scores. Specifically

    a. The consistency and accuracy may be overestimated because quality control checks are done at scheduled times, all of which are in the mornings when scorers are fresh.

    b. A single scorer scores all responses for a content area for any student. This may result in a halo effect, which would distort reported levels of

consistency (and have an impact on the magnitudes of the reliability estimates). There is no systematic rescoring by Scoring Coordinators or Team Leaders for all scorers to check for this halo effect, scorer drift, or inaccuracy.

c. The Content Panel noted discrepancies between the rubric and how the rubric was interpreted in the selection of Benchmark papers.

Recommendations

The following recommendations are for the most part minor in terms of their impact on the program. However, recommendation two is of the highest priority.

1. Attempts to compress the time line for the scoring and reporting should be approached cautiously. Change to the time line could be most effectively undertaken by making substantive changes in the nature of the program. Such changes could include using more objectively scored items (e.g., multiple choice or very short answer). This change would make the program different so it should not be made simply for the purpose of shortening the time between administration of the program and the dissemination of score reports.

2. Additional quality control checks should be introduced.

   a. Quality control checks (check sets and accuracy sets of tasks) should be performed at other than the scheduled times. At least some of these checks should be made at random, unannounced times. Moreover, papers with poor language structure, spelling errors, and poor handwriting should be included to insure that scoring bias is not introduced by irrelevant factors.

   b. Scoring Team Coordinators and Team Leaders should undertake to perform random checks on all scorers on their team at random intervals to verify scorer accuracy.

   c. There should be a systematic review of the extent that rubrics are sufficiently specific in what is required to achieve specific scores. The selection of Benchmark papers needs to be consistent with the Maryland Learning Outcomes that are being referenced. To accomplish this an independent panel of subject matter experts should cross validate the quality of rubrics and the appropriateness of the selected Benchmark papers.

6. Validity Evidence

Because validity refers to the degree to which evidence and theory support interpretations of test scores dictated by the purposes of the test, it is essential to focus validity evidence directly on these articulated purposes. The stated purposes of MSPAP are to a) provide information on school and district accountability and b) encourage instructional improvement. Further, the assessment was designed to meet goals of instructional/school reform by utilizing performance, constructed-response formats that are tied to real life situations and that measure the application of knowledge and skills through critical thinking, higher-order reasoning, and problem solving.

The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) identify several potential sources of validity evidence. This section first summarizes information from MSPAP pertaining to these sources of validity evidence and then critiques the quality of that evidence for supporting the intended interpretations of scores from MSPAP.

Sources of Validity Evidence

Five categories of validity evidence are presented in the Test Standards: a) evidence based on test content; b) evidence based on response processes; c) evidence based on the internal structure of the test; d) evidence based on the relationships of scores on the test with other variables; and e) evidence based on the consequences of testing. Each of these categories of validity evidence is addressed in turn, with an evaluation of the quality of the validity evidence from MSPAP. An integration of the quality of MSPAP validity evidence is provided in the concluding section.

a. Evidence Based on Test Content

MSPAP was designed to measure Learning Outcomes that were adopted in 1990 by the Maryland Board of Education. (New outcomes were adopted in 1999, but MSPAP has not yet been revised to reflect these new outcomes.) These Learning Outcomes identify targeted

learning goals for Grades 3, 5, and 8 in content areas of Reading, Writing, Language Usage, Mathematics, Science, and Social Studies. MSPAP currently measures only a subset of these learning goals. In addition, MSPAP was also designed to further school and instructional reform by using real-life performance tasks that assess the application of knowledge and skills through problem solving, critical reasoning, and higher-order thinking.

There is little information available on the process that was used in the identification of these Learning Outcomes. It is not clear whether the Learning Outcomes were developed by administrators or teachers and to what degree they were subjected to internal or external review prior to adoption by the State Board of Education. The 1998 Technical Manual states that the Learning Outcomes were based on national curriculum standards and learning theories.

Test content varies from year to year and from form to form. Not all of the Learning Outcomes are assessed each year. A matrix sampling procedure is employed so that within any year, designated Learning Outcomes are distributed across the three forms. Individual students take one of these forms. Thus, although all designated outcomes are assessed across all students who participated in a particular year's assessment, individual students taking specific test forms are assessed only on a limited number (about 1/3) of the designed outcomes.

Teachers from relevant content areas and grade levels develop the tasks comprising the assessment. During the test development process, several reviews are conducted to verify the alignment of the assessment tasks with their intended Learning Outcomes.

One issue regarding test content pertains to whether the content, although in the curriculum, has in fact been delivered to the students by the time of testing. If the students have not been exposed to the content, then low school-based scores should not be interpreted as meaning that instruction was ineffective. If it is unknown whether the content has actually been taught, then interpretation of low test scores becomes problematic. The most direct interpretation of low

scores is that the students are not able to perform the skill. However, it is not clear why the student cannot perform the skill as this may be due to inadequate instruction or due to lack of instruction. In 1991, a survey was administered as part of the MSPAP. Teachers were asked to indicate whether the tested content was covered to-date in the curriculum, whether it is in the curriculum but not yet covered, or whether it is not a part of current content area curriculum. At Grade 3, for example, over 50% of the teachers indicated that the material related to mathematics Tasks 2 in Session 6, Tasks 1 and 3 in Session 7, and Task 2 in Session 8, would be taught later in the school year. Similar results were found in the teachers' responses for the timing of curriculum delivery at Grades 5 and 8.

One unique feature of MSPAP is the administration of the assessment to randomly formed groups of students. Students perform pre-assessment activities in these groups, often directed by the teacher/test administrator. It is presumed that the purposes of the pre-assessment tasks are to a) provide all students with the pre-requisite information so that the assessment can measure application of knowledge and skills, b) provide background information relevant to the assessment task, and c) serve as a motivator for students to engage in the assessment task. All scored student responses, though, are from independent student work. Another unique feature is the incorporation of hands-on activities as part of the assessment. These tools and manipulatives are presumed to serve to improve the real-life relevance of the assessment and to improve student engagement in assessment activities.

As part of the test development process, a Bias and Sensitivity Committee reviews tasks. Further, specifications are designed to ensure that MSPAP is in compliance with MSDE's regulations regarding multiculturalism, avoids biases against race/ethnic, religious, cultural, groups and biases against gender age, disability, or sexual orientation, and avoids issues that are sensitive or that may offend a large segment of the population. Differential item functioning has been regularly addressed since 1994.

Evaluation of MSPAP Validity Evidence for Test Content. It appears that good efforts are made to ensure that the tasks comprising the assessments do align with the intended Learning Outcomes, consistent with Standard 1.6, which stipulates that the procedures followed in specifying and generating test content should be described and justified in reference to the construct the test is intended to measure. Also relevant is Standard 13.3 which states that when a test is used as an indicator of achievement in an instructional domain or with respect to specified curriculum standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. However, as raised by the Content Review Committee, there are instances where there are errors in the factual content of some of the tasks or that the skills required by the task (or scored according to the scoring guides) are not congruent with the target Learning Outcome. This degrades the validity of the scores as representing the intended learning outcome.

Further, it is not clear how well these Learning Outcomes are aligned with the curriculum. There is some indication that these Learning Outcomes were identified to further goals of instructional/school reform (and may therefore not be in direct alignment with the curriculum at the time of their adoption). However, sufficient time has passed to allow for the revisions of curriculum to align with these articulated Learning Outcomes. A survey was administered in 1997 to over 400 teachers system-wide by the Maryland State Teachers Association to address this question. Nearly 75% of responding teachers from kindergarten to eighth grade indicated that the curriculum was geared to MSPAP performance.

More information is needed regarding how well the timing of the curriculum matches with the scheduling of the MSPAP administration. Based on the survey of teachers as part of the 1991 administration, teachers indicate most of the content is delivered in the curriculum. However, there are notable places where the delivery of content occurs after the date of the assessment. Thus, the timing of the assessment precludes measuring how well students have learned that content. It is curious, though, that there is so much of the curriculum yet to be delivered after the

May administration of the assessment. The phrasing of the question does not clarify whether the content is in the curriculum and not yet covered in this school year or whether it will be covered in the next school year. This survey is no longer given as part of MSPAP. A periodic survey of teachers regarding curriculum alignment of the MSPAP assessment tasks should be undertaken.

The use of student groups and pre-assessment activities is both a strength and a potential weakness for the validity of score interpretations. On the positive side, it is possible that the use of these pre-assessment activities could serve to remove from score interpretations confounds because students do not have the needed basic information to demonstrate their ability to apply knowledge and skills through problem-solving, critical reasoning, and higher-order thinking. If the students do not have the pre-requisite information, they may not be able to demonstrate their ability to perform these more sophisticated tasks. Through the pre-assessment tasks, it is presumed that background information and task motivation will be enhanced, again adding to the validity of score interpretation.

However, there are also serious potential threats to the validity of score interpretations that may be introduced by the use of these pre-assessment activities. The degree to which the successful completion of the pre-assessment tasks impacts the accuracy or quality of individual student work on the operational assessment could be another source of score invalidity. Some of the pre-assessment tasks are lead by the teacher/test administrator. These teachers/test administrators therefore have a substantial role in the successful delivery of the pre-assessment activities. Less than successful delivery of these teacher-lead pre-assessment activities could affect individual student performance on scored tasks.

Further, the pre-assessment tasks are administered to groups of 3 to 5 students. With randomly formed groups, the variability of student abilities within and across groups will likely be large. Research indicates that the composition of student groups with respect to individual student ability affects how members of the group perform on tasks (Webb, Nemer, Chizhik, & Sugrue,

1998). This will create inequities in the quality of the pre-assessment experiences that could influence how individuals perform on the scored parts of the assessment. Standard 13.11 states that test users should ensure that any test preparation activities and materials provided to students do not adversely affect the validity of test score inferences.

Similar comments and concerns can be raised about the use of tools and manipulatives during the operational tasks. Designed to enhance the "real-life" nature of the assessments and to potentially increase students' engagement, their use could add to the validity of score interpretations. However, the inclusion of these materials as part of the assessment could add a source of inequity across assessment administration. Although teacher/test administrators have early access to the assessments they will be working with, and are instructed to be familiar with the tools and manipulatives needed for successful administration of their assessments, this may be not achieved uniformly across the schools. Some of the manipulatives require pre-administration preparation (e.g., pouring specified amounts of liquid into containers, cutting strips of tape into specified lengths). If these tasks are not done in advance, or done incorrectly or even differently by some teachers/test administrators, students will have an unequal opportunity to perform successfully on these tasks. It should be noted, however, that the selection of tools and manipulatives is considered in the test development process. Special Tools and Manipulatives Committees meet to review the materials to be included in all tasks. Grade level committees meet to observe the tools and manipulatives and how they are being used in the assessment. Even with the careful review and screening of these manipulatives and tools, there is the concern that they may not be prepared or administered in a fair, consistent, and equitable manner, thus introducing the possibility of invalid score interpretations.

During test development, the Bias and Sensitivity Committee reviews assessment tasks for issues related to race, gender, ethnicity, and other pertinent factors. The specifications that guide the development of tasks and the review by the Bias and Sensitivity Committee are clear and thorough. These procedures are consistent with Standard 7.4 which states that test developers

should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups. The addition of empirical investigations of differential item functioning strengthen the evidence of validity of score interpretations for Black, Hispanic, Asian, and female students and is consistent with Standard 7.3 which indicates that research on differential item functioning should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups. However, currently the analyses for differential item functioning are conducted on data from the operational administration. This does not provide many options for substitution of tasks or questions if any demonstrate evidence of differential item functioning.

Evidence Based on Response Processes

The intent of this category is to provide evidence on the fit between the response processes employed by the students and the construct to be measured. In the case of MSPAP, the intent is to measure higher-order reasoning using constructed response tasks. There is limited evidence in the MSPAP materials to indicate how well the cognitive processes used by the students are in fact congruent with the use of higher-order reasoning skills. On the surface, the tasks appear to be focused on the application of knowledge and skills to solve problems, reason critically, and make decisions, predictions, or recommendations. These task demands are consistent with those skills judged to require the use of higher-order reasoning.

Developmental appropriateness of the tasks and response processes will influence the validity of the scores. As part of the test development process, task developers use the "Guidelines and Checklist for Creating Developmentally Appropriate Tasks." In addition, all tasks are reviewed at several points in the test development process for compliance with these guidelines. Grade level teachers teamed with non-classroom professionals who have expertise in cognitive and psychological development review grade-level assessments.

Evaluation of MSPAP Validity Evidence for Response Processes. There is only limited information about the cognitive processes used by students when they are completing MSPAP tasks. Because of the intent of the assessment program to measure the application of knowledge and skills through higher-order reasoning, it is very appropriate that validity evidence directed to response processes be gathered. Standard 1.8 specifies that if the rational for a test use or score interpretation depends on premises about the cognitive operations used by examinees, then theoretical or empirical evidence in support of those premises should be provided.

Another critical feature in the meaning of scores from MSPAP is the requirement that all responses be written. This again relates to the appropriateness of the cognitive processes required by students, however in this case, the requirement that all answers be provided in writing may serve to confound the interpretation of student performance in content areas other than writing. For example, a student who has a low score in science may in fact have the ability to perform the skill being measured but may lack the verbal or writing skills needed to express the answer in writing.

In addition, although the question may on the surface ask students to use higher order thinking skills such as making predictions or drawing a conclusion, if the scoring criteria or rubrics do not appropriately reflect this, the scores will not be valid for the intended interpretation. Likewise, if the sample of student work that is used in the training of scorers is inappropriate in how it illustrates the scoring of the student work for the intended performance, again, the scores will not be valid for the intended interpretation. Therefore, evidence to support the meaning of test scores should include results from efforts to document that the skills being elicited by the tasks are in fact requiring the intended cognitive processes, that the scoring rubric operationalizes these intended cognitive processes in appropriate ways, and that the training materials for the scorers provides adequate and appropriate evidence of the intended student performances.

The Maryland State Teachers Association commissioned a MSPAP survey in 1997 and 1998. According to the Report to the 1998 MSTA Representative Assembly, less than one-quarter of Maryland teachers agreed with the statement, "The MSPAP test is developmentally appropriate for their students." Overall 67% either disagreed or strongly disagreed with this statement. Therefore, there seems to be concern, at least with the majority of the teachers, about the developmental appropriateness of MSPAP tasks. Further, student performance, particular at the lower grade levels, is low. It is not clear why, but one contributing factor could be that the tasks or response strategies are too difficult for the students or that they call for cognitive skills that are developmentally inappropriate. Evidence should be gathered to clarify the degree to which low performance by students at the lower grade levels is related to overly demanding tasks as opposed to inadequate instruction.

Evidence Based on Internal Structure

There are two issues related to internal structure that need to be addressed: factorial structure and unidimensionality. Factor structure evidence would support the inference regarding the uniqueness of content areas whereas unidimensionality would address the degree to which the Item Response Theory model being used is appropriate.

Evidence of factor structure comes from two sources, factor analyses and between content area correlations. Factor analyses were conducted on the 1991, 1992, and 1996 version of the MSPAP. Several changes in configuration have occurred since the 1991 version of MSPAP. In the 1991 assessment only Reading, Writing, Language Usage, and Mathematics were assessed and the test was administered in intact classrooms. The factor analysis from the 1991 MSPAP indicated that Reading, Writing, and Language Usage loaded on a single factor (although it was hypothesized that Reading would load on one factor and Writing and Language Usage on another). Further, Mathematics showed a unidimensional structure (although two factors were hypothesized, one for Mathematics Content and one for Mathematics Process).

In 1992, Social Studies and Science content areas were added to MSPAP and it was administered to randomly formed groups consisting of 3 – 5 students. Therefore, the 1992 MSPAP is similar in content and administration to the current MSPAP. Green (1993) reported that factor analyses by content area generally supported a unidimensional structure.

Tippetts and Michaels (1997) compared the factor structure for Reading and Language Usage on the 1996 MSPAP for students with accommodated and non-accommodated administrations. Basic equivalence in factor structures was found, supporting the generalizability of scores across these administration modes. At Grades 3 and 5, a two-factor solution was reported; a unidimensional structure was reported for Grade 8.

More information on the internal structure is available through content area score correlations. This information was presented in the 1995 and 1998 Technical Manuals. In 1998, these correlations are presented by grade level statewide. These correlations are based on all students at a particular grade level score. The content area correlations, across schools statewide, are all moderate to strong (low of .57 to a high of .82). Mathematics and science had the highest correlations for Grades 3 and 5. For eighth grade the highest correlation was between language usage and writing. The smallest correlations for each grade level were between reading and writing for Grades 3 and 5 and between language usage and mathematics for Grade 8.

Evaluation of Validity Evidence for MSPAP Internal Structure. Unidimensionality was addressed using the data from the 1991 assessment (upon which the initial IRT scaling was based). Using the results of the factor analysis it was determined that reading, language usage, and writing formed an essentially unidimensional scale as did mathematics content and process. Likewise, factor analyses for the 1992 MSPAP generally supported unidimensionality within content areas. This is in part discussed in Standard 1.11, which indicates that evidence concerning the internal structure of the test should be provided if the rationale for test use or interpretation depends on premises about the relationships among part of the test. Because item

response theory is used in scaling, a unidimensional internal structure needs to be documented. On the other hand, because scores are reported separately for content domains, the internal structure of the test must also support their interpretation as well. This presents a challenge for the test developers as they must demonstrate sufficient unidimensional structure for the mathematical model underlying scaling, yet at the same time support multidimensional score interpretations.

The magnitudes of the correlations calculated on the 1998 MSPAP results were very high. If these correlations were to be corrected for attenuation they would approach 1.0. Therefore these content area correlations are remarkably high. The reasons for these excessively high correlations are not clear but may result from lack of content area independence due to the interdisciplinary nature of the tasks. The degree to which such high intercorrelations impact on the validity of score interpretations is not clear. It is obvious that at the school level, the rank order of scores across content areas has little variation for all students. This raises questions about whether these separate content area scores are meaningful at the school level and also about the validity of the multidimensional internal structure underlying the test. Standard 1.12 addresses this point in stating that when interpreting subscales, relevant evidence needs to be presented to document such interpretations.

Evidence Based on Relations with Other Variables

In addition to MSPAP, Maryland third and fifth grade students also are administered CTBS/4. Correlational evidence showing the relationship between student performance on the CTBS/4 and MSPAP is available from studies conducted in 1991 and 1995. Correlations for Grades 3 and 5 ranged from the mid .50s to mid .80s. Correlations were also calculated between MSPAP scores and scores on the Maryland Functional Literacy Test (MFLP). These correlations ranged from .42 to. 61.

In 1991, a study was conducted to determine the relationship between teacher ratings of student performance and the student's performance on MSPAP. A total of 300 students per grade were randomly selected from 12 LEAs to be part of the study. Teachers rated their students' performance in relation to the relevant grade level curriculum as being below, at, or above grade level. Student Reading Proficiency ratings had a correlation with MSPAP Reading of .47 at Grade 3 and of .60 at Grade 5. Writing Proficiency rating correlations with MSPAP writing were .46 at Grade 3 and .43 at Grade 5. The correlations for Grade 3 and Grade 5 were .53 and .46 for MSPAP Total Mathematics and Mathematics Proficiency ratings. A similar study was conducted in 1997. Correlations between teachers' perception of students' proficiency levels and these students' MSPAP performance ranged from .41-.49 and Grade 3, .43-.53 at Grade 5, and .43-.61 at Grade 8.

Evaluation of MSPAP Evidence for Relations to Other Variables. The evidence that is provided shows moderate to strong relationships with MSPAP to other variables. It would be useful to repeat some of these studies, due to the changed nature of the MSPAP over time. Also, because of the purported intent of measuring higher-order thinking, studies examining relationships of MSPAP scores to those from other tests that measure related (and unrelated) constructs would be useful. The degree to which higher order reasoning, and not knowledge acquisition and skill level, is being assessed should be studied. These studies would help to clarify the meaning of scores from MSPAP. Such studies are consistent with those called for in Standard 1.14, which in part states that patterns of associations between and among scores on the instrument under study and other variables should be consistent with theoretical expectations.

Evidence Based on Consequences of Testing

New to the 1999 Test Standards, evidence of the consequences of implementing an assessment program is identified as an additional source of validity information. Such consequences could be planned and positive, such as progress toward instructional reform as a result of the

implementation of the assessment program. Consequences could also be unplanned and be both positive and negative, although most unplanned consequences fall into the negative category. Evidence for the consequences of MSPAP is reported in several studies, for example Firestone, Mayrowitz, & Fairman, 1997; Koretz, Mitchell, Barron, & Keith, 1996; Lane, Parke, & Stone, (1998); Lane, Ventrice, Cerrillo, Parke, & Stone, (1999).

For MSPAP, one of the articulated goals of the assessment program is to institute instructional reform by using more cooperative learning, hands-on curriculum experiences, and a movement toward more authentic classroom assessment practices. Surveys sponsored by the Maryland State Teachers Association in 1997 and 1998 provide some evidence regarding the realization of these intended consequences. Across teachers from Kindergarten to Eighth Grade, more than 50% of the teachers either agreed or strongly agreed with the statement that "MSPAP helps the education of children." In addition, more than 50% of the teachers indicated they agreed or strongly agreed with the statement that "experience with MSPAP has raised expectations for student performance." Nearly 75% of the teachers indicated that "the curriculum was geared to MSPAP performance." These results support the conclusion that movement toward curriculum reform has occurred as a result of MSPAP.

Lane et al. (1998) surveyed mathematics teachers, administrators, and students from a total of 59 elementary and 31 middle schools, stratified on a) percent free and reduced lunch, and b) MSPAP performance gains from 1993-1995. They reported that elementary teachers in Maryland indicate they place a higher emphasis on mathematics instructional reform than did middle school teachers. Similar results were reported by teachers in Reading, Writing, and Science (Lane, et al., 1999).

Another intended consequence of the MSPAP was the use of results for instruction. Over 50% of the teachers who responded to the 1997 survey agreed or strongly agreed with the statement that "MSPAP data were useful for new lesson plans" and nearly 50% agreed or strongly agreed

with the statement that "new lesson plans generated higher MSPAP scores". Results from the Lane et al. studies (1998, 1999) indicate that teachers view MSPAP as a useful tool for improving instruction.

As part of the Lane et al. studies, teachers were asked to submit sample lesson plans and classroom assessments. Lane, et al. reports that 50% of the lesson plans were consistent with MSPAP tasks, and 15% were very similar.

Firestone, Mayrowitz, and Fairman (1997) interviewed eighth grade mathematics teachers in Maine and Maryland on the impact of performance assessments on classroom instructional practices. Although teachers in Maryland reported the use of reform-based instructional tasks in their mathematics instruction, Firestone, et al. evaluated these instructional tasks as superficial rather than meaningful in mathematics instructional reform. They argue that more fundamental change is needed to support meaningful mathematics instructional reform than simply the implementation of a statewide assessment program, even one with moderate to high stakes.

Koretz, Mitchell, Barron, & Keith (1996) also studied the consequential validity of the MSPAP program. They conducted phone interviews with 224 5^{th} and 8^{th} grade teachers in 1994-95 (prior to the administration of the 1995 MSPAP). An additional 186 teachers in these grade levels received mailed surveys. Similar to the results from Lane, et al. (1998, 1999), Koretz, et al. reports that teachers generally support MSPAP as an instrument for instructional reform. However, the teachers also reported shifting time away from subject areas that are not addressed in MSPAP.

Teachers in the Koretz, et al. study also were asked about their test preparation activities. Teachers indicated that they spend considerable instructional time in preparation for MSPAP-like tasks. Some teachers reported giving practice assessments with the goal of improving their

students' performance on MSPAP (without necessarily increasing the students' content knowledge being assessed).

Other evidence of unintended consequences has also been demonstrated. The teachers who reported that they had made adjustments to their curriculum to emphasize MSPAP content also reported reducing the attention paid to non-MSPAP content areas. In a 1999 survey of principals, 43% of the principals surveyed indicated teacher morale was worse since the implementation of MSPAP.

Evaluation of MSPAP Evidence on Consequences. Limited evidence is available to date regarding the intended and possible unintended consequences of MSPAP. Although the evidence shows that many of the intended consequences may have been realized, this same evidence could also be related to unintended consequences. Because nearly half of the teachers report adjustments of lesson plans and curriculum re-alignment, the curriculum delivered to students may be narrowed by reflecting only those learning outcomes targeted by MSPAP. Further, because schools are purportedly evaluated by how their students' MSPAP performance has improved over time, redirecting the curriculum and lesson plans to MSPAP-type activities will likely increase student performance across years. This could give an inflated indication of overall student improvement if these gains are not related to student improved knowledge.

Evidence of unintended consequences is called for in Standard 1.24 which states that an attempt should be made to investigate whether such consequences arise from the test's sensitivity for characteristics other than those it is intended to assess or to the test's failure to fully represent the intended construct. Further, Standard 7.9 stipulates that test developers should fully and accurately inform policymakers about consequences of test use. Survey results indicate that MSPAP has contributed to low teacher morale, teachers' expressions of stress and their decisions either to refuse to teach at the grade levels where MSPAP is administered or to leave teaching altogether. There are also concerns about validity of the interpretations of score results when

there are no direct consequences for students. This may mean that these students are less likely to give their best effort on the test, leading to lower student performance than might have occurred under a testing environment when student scores were reported and used.

Other Issues Related to Validity of Scores from MSPAP

In addition to the general sources of validity evidence identified by the Standards for Educational and Psychological Tests (AERA, APA, NCME, 1999), some additional issues related to validity of score interpretations for MSPAP need to be addressed. One issue relates to the possible differential impact on test performance of students from schools with low socio-economic status, of higher teacher mobility, and lack of motivation due to limited individual consequences of test performance. A report prepared by Dr. Lisa Smith, Kean University, especially commissioned for the MSPAP evaluation (see Appendix B) addresses these issues. Her review of relevant research showed a probable lowering of student test performance a) as a function of teacher mobility (which was documented to be higher in schools in urban and lower socio-economic areas), and b) by students having less than optimal motivation due in part to the limited personal consequences of test performance. Again, the literature supports the probable relationship between performance of students from schools in urban/lower socio-economic areas and lower motivation without personal consequences.

On a related matter, in districts where there is high teacher turnover, teachers who are new to the Maryland program will likely have limited exposure to the multi-discipline, constructivist philosophy that heavily undergirds the MSPAP. In addition, they will not be familiar with the unique aspects of the administration of MSPAP, particularly the need for advanced preparation of materials and the emphasis on pre-assessment activities. If new teachers are naïve or ineffective in these features, the scores of their students may not appropriately reflect their level of instruction or the students' actual achievement levels.

Reporting of individual student MSPAP scores is officially discouraged by MSDE. Reporting and using individual student scores is inappropriate, and unwarranted, given the intended purposes of MSPAP, its design, and its psychometric properties. MSPAP documents clearly state that individual scale scores should not be interpreted or reported. This is consistent with Standard 1.3 that says that if an interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations. However, individual student scores are provided to schools and reported in some cases to parents. MSDE even provides interpretative guides for the use of individual student scores for making comparisons of percents of students who meet or exceed certain proficiency levels. This presents a confusing situation for consumers of MSPAP results and is in violation of the Test Standards. If the intent is to use MSPAP scores for school accountability and instructional improvement, only group level results should be made available and reported. If the intent were to also use MSPAP scores for making decisions about individual students, then reporting of individual scores would make sense. However, the design of the assessment program (using a matrix sampling approach) and levels of psychometric quality of the scores do not support use of MSPAP results for use with individuals. Further, administration of MSPAP to groups of students who have participated in cooperative pre-assessment activities confounds the interpretation of individual student scores.

One point that should be made relates to the policy regarding what students are included in MSPAP. The state policy of including all students whose educational plans are consistent with the Maryland Learning Outcomes, providing accommodations as needed for the assessments, is appropriate and consistent with current federal and state legislation.

Conclusions and Recommendations

Conclusions

1. Review committees evaluate how well the targeted Maryland Learning Outcomes are incorporated in the assessment tasks. This provides evidence that the assessments are aligned with the Maryland Learning Outcomes.

2. The timing of the administration of the assessments does not necessarily align with the delivery of the relevant content.

3. Although the intent of MSPAP is purportedly to measure the application of knowledge and skills through tasks that require higher order thinking, there is no direct evidence that the tasks do in fact meet this goal. In addition, there is no evidence that the scoring rubrics or scored student work reflect the emphasis on higher order thinking skills.

4. All student responses are constructed. This presents a potential confound in the interpretation of scores as a low score could mean either that the student did a poor job in answering the question or that the student did not have the necessary verbal and/or writing skills to communicate the answer to the question.

5. Pre-assessment tasks are administered in small randomly formed student groups. It is probable that some groups will consist of all high ability or all low ability students. The pre-assessment tasks are used to introduce important information and knowledge that is needed by the individual students to perform well on the assessment tasks. If the groups are not equal in their performance on these pre-assessment tasks, the ability of the individual students to perform well on the operational, scored tasks is compromised. This could have implications for the validity of score interpretations as low scores may result from less that optimal information yielded by the pre-assessment activities that interfered with the students' acquisition of the necessary information to succeed on the task.

6. Many of the assessment tasks require advanced acquisition of and preparation of materials and manipulatives. If these pre-assessment preparations are not done correctly and consistently, students will not receive a fair administration of the tasks. The validity of student and school scores is dependent in part on the fair and equitable administration of the assessment, including the appropriate acquisition and preparation of materials needed for the assessment.

7. Even though the tasks are reviewed for developmental appropriateness, teachers report that the tasks are not developmentally appropriate, especially for the lower grade levels (3 and 5). Performance of students at these grade levels is consistent with the conclusion that the tasks may be overly difficult for these grade level students. It is not clear whether the levels of cognitive challenge (either in the presentation of the tasks or the way responses are made) or the interdisciplinary nature of the tasks may be contributing to low student performance and teacher perception of inappropriate development level of the tasks.

8. Factor analyses have been completed by content area and in most cases are consistent with a unidimensional structure. However, factor analyses across content areas have not been reported. Such analyses would help support the validity of reporting and interpreting separate content area scores.

9. Excessively high between content area correlations (especially when a correction for attenuation is applied) suggest that the interpretation of the scores by content area may be misleading, if not inappropriate.

10. There is reason to believe that MSPAP scores may be differentially valid for students from lower socio-economic groups due to higher teacher turnover and lower student motivation.

11. Teachers new to Maryland may not be well versed in the instructional philosophy that undergirds the MSPAP. Further, they may not be as versed with the kinds of pre-assessment activities for MSPAP as teachers who have experience with MSPAP. Students with new teachers may have lower scores because of lack of teacher familiarity with MSPAP rather than because of ineffective instructional practices.

12. The inclusion rules for students with disabilities and who are limited in their English language skills are appropriate and consistent with state and federal laws.

13. Reporting of individual student scores is not supported by MSPAP. However, these scores are reported to schools, and on occasion to parents and students. This practice is inconsistent with the intent and design of MSPAP.

14. Gains in student performance across years could be attributed to increased student learning due to instruction or to other possible sources. The scaling and equating procedures should remove changes in test difficulty as a potential reason for increased school performance. However, it is not clear whether gains in performance are in fact due to increased student learning from improved instruction, or possibility to factors such as enhanced test preparation, familiarity with reused assessment tasks, or increased exposure to integrated performance assessment tasks.

15. Evidence is provided that many of the intended consequences of the MSPAP are being realized, especially with regard to instructional reform. Some evidence also shows that unintended negative consequences have also occurred, including lower teacher morale, less emphasis on content areas not covered in MSPAP, and increased teacher stress at the grade levels at which MSPAP is administered. Therefore, evidence indicates that both intended and unintended consequences have occurred due to MSPAP.

Recommendations

1. Information is needed to address how well the Maryland Learning Outcomes are incorporated into the curriculum. Evidence about when these skills are addressed in the curriculum could be gathered by review committees who map the Maryland Learning Outcomes to curriculum guides and other curricular materials.

2. Evidence of when the content measured in the assessment is delivered in instruction needs to be gathered. In earlier versions of MSPAP, a questionnaire addressing this issue was administered to teachers. This questionnaire is no longer administered. Reinstating a questionnaire of this type would provide current information about whether students have the opportunity to learn the content assessed in MSPAP. Other materials, such as a review of curriculum guides, lesson plans, student work, and classroom assessments would also provide evidence of opportunity to learn.

3. Evidence needs to be gathered to support the inference that MSPAP assessment focuses on application of knowledge and skills through tasks that require higher order thinking. In addition to evaluating the tasks for whether they elicit higher order skills, the scoring rubrics and actual scoring of student work should be evaluated to ensure that they support the conclusion that higher order skills are being assessed.

4. The heavy emphasis on writing as the sole means for a student to communicate his or her answers should be re-considered due to the potential confound in score interpretation. It could be the case that a student has the cognitive skills and knowledge to correctly answer a question, but is limited in the writing skills needed to communicate the answer effectively. This confound prevents direct interpretation of low scores in a content area.

5. The use of group-based, pre-assessment activities should be discontinued as they introduce a potential source of invalidity in the individual student performance on MSPAP tasks.

6. The incorporation of manipulatives in the assessments should be re-considered. If these materials and manipulatives cannot be administered in a more standardized manner, they should be discontinued.

7. More evidence is needed to support the conclusion that the tasks and response requirements are developmentally appropriate, especially at grades 3 and 5. More information about the criteria used in the review of the tasks for developmental appropriateness should be provided, in addition to seeking additional evidence that what is asked and how it is asked of the students is consistent with their cognitive level of development and educational experiences.

8. Additional empirical evidence should be presented on the internal structure of the integrated assessments. Current analyses consider internal structure by content area, rather than for a full assessment.

9. Additional evidence is needed to defend the meaning of the separate content area scores, particularly in light of the high correlations exhibited between content area scores.

10. Because high teacher turnover has been identified as a possible reason for low scores for students from schools with concentrations of lower socio-economic groups, special efforts should be made to retain teachers in these schools and to provide additional orientation and training for these new teachers on MSPAP tasks and curriculum integration.

11. Teachers new to Maryland should be given special orientation to the instructional philosophy that undergirds MSPAP and be provided with training on the pre-assessment activities required to administer MSPAP.

12. The decision to include all students whose educational program aligns with Maryland Learning Outcomes in the assessment, with appropriate administrative accommodations, should be maintained as it is consistent with current law.

13. According to current policy and assessment design, it is inappropriate to report individual student scores to parents and students. Therefore, unless the design of the assessment is changed, reporting of these scores to schools should be discontinued.

14. Additional information should be gathered to support the inference that the gains realized by schools across years is due to improved student achievement and better instruction. In order to strength this inference, evidence should be gathered to counteract the conclusion that these gains are the result of enhanced test preparation, familiarity with reused assessments, or increased exposure to integrated performance assessment tasks, and not due to real gains in student achievement and teaching quality.

15. More evidence is needed to document both the intended and unintended consequences of MSPAP, including evidence that supports the outcomes of MSPAP in directing instructional reform. No evidence has been provided regarding revamping of the curriculum to include interdisciplinary content. Teacher and principal surveys suggested that unintended negative outcomes have occurred in teacher morale and narrowing of the curriculum. These issues, and others, should be addressed in a planned research program on the consequences of the assessment.

7. Standard-Setting Procedures

Three Test Standards (AERA, APA, NCME, 1999) are related to standard setting. The first of these Standards (4.19) indicates that when score interpretations involve cut scores, the rational and procedures used to establish the cut score(s) should be clearly documented. The second Standard (4.20) states that "When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria" (p. 60). Finally, Standard 4.21 suggests that when cut scores define proficiency categories that are based on direct judgments about the adequacy of item or test performances or performance levels, judges should be able to bring their knowledge to bear in a reasonable way.

Overview and Introduction

In the MSPAP program three sets of cut scores have been set. One set of cut scores is designed to assign students to one of five performance levels. A second cut score classifies student performance as unsatisfactory, satisfactory, or excellent. The third cut score is used to classify schools as being unsatisfactory, satisfactory or excellent. The methods used to set each set of cut scores are described below. The focus of this section is on the processes used for setting the various performance standards for both schools and students. These processes are described and critiqued.

Procedures for Setting Institutional Performance Standards

The process for setting institutional[1] performance standards for all variables in the Maryland School Performance Program is described in Thorn et al. (1990). The process for setting the

[1] School and district performance standards are expressed in terms of the percentage of students who pass, dropout, etc. This description applies to the method used to set standards on variables in the Maryland School Performance Program other than MSPAP. However, an

institutional performance standards for MSPAP is described in the MSDE (1994) Score Interpretation Guide. The methods used to set the MSPAP institutional performance standards are reported to be essentially the same as the methods described in Thorn et al. (1990).
The process described in Thorn, et al. (1990) involved a committee of 20 educators (the Maryland School Performance Standards Committee) who developed generic definitions of two levels of performance for schools: Satisfactory and Excellent. Satisfactory performance was characterized as "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" (p. 7). Excellent performance is defined as being "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" (p. 7).

Once the global definitions of Satisfactory and Excellent were established, an 11-member subcommittee of the Maryland School Performance Standards Committee was formed to recommend ranges of student performance that would initiate discussion of operational performance standards. The MSDE decided to set operational standards for Satisfactory performance. Excellent performance would be determined statistically by setting a performance level that was "significantly" higher than the operational standard for Satisfactory.

A Delphi-like process was used to set the standard for Satisfactory performance for schools. The initial step in the process was for subcommittee members to review the definition for Satisfactory performance and then exchange their initial convictions about what constituted Satisfactory performance. Upon revealing their personal convictions, they undertook the preliminary consensus round. This was followed by an examination of empirical data. Personal convictions were then revisited and shared among the subcommittee members, followed by a final consensus

---

MSDE (1994) publication specifies that these same procedures were used to set the MSPAP standards to determine if schools and districts are performing at satisfactory and excellent levels.

round. The final consensus round resulted in a percentage range within which the Satisfactory performance standard could be selected.

The members of the larger committee of educators (excluding the subcommittee) reviewed the process used by the subcommittee and the proposed ranges of student performance that defined Satisfactory performance for schools (e.g., 60% to 80% of the students achieving at a satisfactory level on MSPAP).

The School Performance Standards Committee recommended a range of 60% to 80% of students achieving at the Satisfactory level as the standard for schools to be classified as being Satisfactory. After a review process as described above, the final standard that defines a school or district as being Satisfactory was passed by the State School Board. To be classified as Satisfactory a school (or district) must have 70% or more of its students score at a student performance level of Satisfactory. To be considered Excellent, a school (or district) must have "70% or more of its students achieve at satisfactory or above and 25% or more of its students achieve at the excellent level (MSDE, 1998)."

In order to evaluate the performance of a school (or district) on MSPAP the committee also had to define what level of student performance constituted Satisfactory or Excellent performance. Student performance is reported on a five-point scale (each point is called a level). The scale defines 5 as the lowest performance level and 1 as the highest. The final, approved criterion for student performance used to determine the performance of institutions is that Satisfactory student performance is represented by a score of 3 or better on a 5-point scale and a score of 2 or better is Excellent student performance. Thus, for example, a school is determined to be Satisfactory in Grade 3 Reading if 70% or more of its students attain Reading scores at Level 3, Level 2, or Level 1.

Critique of Institutional Standard-Setting Procedures

The reports from which the summaries of the standard-setting procedures were obtained contained fairly detailed information about the process. However, because information about the process is found in several reports, it was difficult to understand fully certain details about the procedures used. Not all the information needed to evaluate these cut scores or the processes used to set them are found in a single report. This criticism also applies to the procedures used to set Student Performance Standards described below.

The various reports do not make clear in what order institutional and student standards were set. However, the MSDE has indicated that the standards defining Satisfactory and Excellent performance for students (both the cut scores for each level of performance, and the performance level descriptions) were set prior to setting the standards for institutional performance. This sequence is appropriate.

The inclusion of multiple groups and multiple reviews and public hearings about the standards is to be commended. The availability of reports that describe the process, albeit there are some problems with the reports, is another positive aspect of the program.

Setting Student Performance Standards

There are two standards for student performance. One standard is a global standard that classifies a student's performance as being either Unsatisfactory (not at Level 3 or better), Satisfactory (at Level 3), or Excellent (at Level 2 or 1). This standard is described above and it is used to make evaluative ratings of schools and districts. The second standard of student performance classifies the student into one of five different performance levels (Level 5 being the lowest level and Level 1 being the highest level). This standard is used to describe student performance in terms of the knowledge and skills students demonstrate at the different performance levels. This information may aid schools and districts in making broad instructional

decisions based on performance on the MSPAP measures. It is the procedures for setting the cut scores for the five levels of student performance that is the focus of this section of the report.

The procedures for setting performance standards for student-level performance on the MSPAP tests were somewhat more complex than were the procedures for setting institutional performance standards. These procedures occurred over a several year period beginning in 1991. These procedures are described in several documents, CTB/Macmillan/McGraw-Hill (1992), Westat, Inc. (1993), and Atash (1994). These procedures resulted in a system for classifying students into five performance levels thus requiring four cut scores. The methods used to set these four cut scores are summarized below. As noted above, in addition to the five performance levels, students are classified as performing at the Unsatisfactory, Satisfactory or Excellent level, based on their classifications using these five performance levels.

Procedures For Setting Student Performance Standards

Setting cut scores defining student performance levels was undertaken in several stages. The first stage entailed determining the scale scores that define the cut score between each of the five levels. The second stage involved elaborating the knowledge and skills that students at each level demonstrate. The outcome of the second stage is called the Proficiency Level Descriptions. Thus, each of the five performance levels is defined in terms of a range of scale scores and also in terms of the student demonstrated knowledge and skills associated with those performance levels.

Each subject area and grade level has different scale score[2] ranges that represent each performance level. For example, the scale scores that define Level 5 performance in Grade 3 reading range from 350 to 489, whereas the range of scale scores for Level 5 performance in Grade 3 Social Studies ranges from 350 to 494. For some content areas score ranges at the

[2] Scale scores range from 350 to 700 with an approximate mean of 500 and a standard deviation of 50.

extremes (Levels 5 and 1) are combined with their adjacent level because there were insufficient data for panels to review at those levels (i.e., there were too few items in the extremes to define the proficiency categories). In the early stages of the program there were more of these undefined performance levels than there are currently. The cut points that separate the performance levels are not necessarily the same for each grade level and subject area.

The standard setting process was initiated with the 1991-92 MSPAP administration. As noted above, the order in which the standards were set is not clear. That is, it is not clear if the determination of the score levels that constituted a schools' being classified as Satisfactory and Excellent were defined before or after determining the cut scores for the five levels of student performance, or before or after setting the cut scores that classified student performance as Satisfactory or Excellent. It appears that the cut scores for the five student performance levels were set last. The principal reason for classifying students into performance levels was to establish Proficiency Level Descriptions that might have some diagnostic value for schools.

The test development contractor (CTB) undertook the first step in the process. Using IRT technology, CTB calibrated each item on the overall scoring scale (about 350 to about 700). The standard setting process involved CTB selecting items that provided maximum information within various score levels. Not all items were selected in each content area at each grade level. The CTB report indicated that some content areas (most notably writing) had few items overall, and that all items were selected because each item provided substantial information. The percentage of items used ranged from as low as 35% (Grade 5 reading) to as high as 100%, with most grade/subject areas using between 67% and 75% of the total number of items available. The items that were not used were examined to insure that the content coverage did not leave gaps in the skills and knowledge tested that year. If there had been gaps it might have resulted in a bias in the development of the Proficiency Level Descriptions.

After the final determination of items to include in the standard setting process was made, frequency distributions of the overall locations of the items across the score levels were produced. An initial set of scale score ranges that corresponded to the five proficiency levels was determined by examining the score distributions and finding approximate natural breaks. The two extreme categories were defined first. These cut scores for the high (dividing Level 1 and Level 2) and low (dividing Level 4 and Level 5) categories occurred close to scale scores of 490 for Level 5 and 620 for Level 1. The remaining two cut scores (between Levels 2 and 3 and between Levels 3 and 4) were around 580 and 530, respectively.

The process as described in the CTB (1992) report is not at all clear because the referenced tables are either incorrectly numbered or incorrectly referenced. In the narrative description, it states that "proficiency levels were intended to increase the interpretability of the scores" (p. 11-2). The CTB report goes on to state that it was deemed important to use the same scale values for all grades and subject areas. In short, the initial cut scores that defined the five performance levels were set by CTB by examining the score distributions of items that provided substantial "information."

The next step in the process shifted from CTB to committees of Maryland educators. Committees of educators were formed for each grade level and subject area. Committees were provided all relevant information needed to develop the performance descriptions. Specifically, each committee had student response books for each form, the scoring guide, and the score level of each item. The objective of this process was to arrive at Proficiency Descriptions that were based on the nature of the content of items within the score ranges for each performance level.

This process used to set the cut scores changed somewhat in 1992 and 1993. One major change was that a new contractor (Westat) was employed to set the cut scores and to establish the proficiency descriptions. The 1992 and 1993 procedures Westat used are described in their reports (Westat, 1993; Atash, 1994). The 1992 procedures represented a reexamination and

reconsideration of the 1991 cut scores and proficiency level descriptions. In 1993 a similar procedure was undertaken as a "refinement" of the 1992 results. In both years the Proficiency Level Descriptions were also reviewed and updated.[3]

The 1992 content and grade level committees were formed. All committees included Maryland educators and content specialists. Committee members were directed to examine the 1991 proficiency level descriptions and conceptualize what students at higher levels can do that students at lower levels cannot do. Working independently committee members examined test activities (items) that had been pre-classified as being at certain performance levels (i.e., test activities were grouped and described as being at Level 1/2, Level 2/3, Level 3/4, and Level 4/5 based on 1991 cut scores). Committee members then matched the items provided in each category to the 1991 Proficiency Level Descriptions. Test activities were then rated as to performance level. For example for test activities pre-classified as being Level 1/2 items, committee members indicated if the item was "Clearly Level 1," "Clearly Level 2," or "borderline."[4] Because items had been ordered from most to least difficult within each grouping, committee members only classified items until they had identified a "run" of three or more consecutive items at each level, disregarding items classified as borderline. Once the two runs had been identified, the cut score that divided the two levels was determined by using a table look up to convert the item scale score for the "bottom" Level 1 item and the "top" Level 2 item in the respective runs. This was done for each level. The final cut score that defined each of the performance levels was the average across all the committee members for the items at the appropriate levels. Confidence ratings were made by each committee member at the conclusion of making their initial judgments of the cut scores.

---

[3] The updating is necessary because not all Maryland Learning Outcomes are assessed in any one year, so to maximize the coverage of the proficiencies across the Outcomes; it was necessary to review the items over a several year period.

[4] Borderline for Level 1/2 was defined as ("could be level 1 or 2 can't tell" in Westat, 1993, Appendix D).

A second round of judgments was made following discussion of the committee's initial cut scores for each level. That is, after discussion of the results of the process described above, committee members independently reexamined the test activities and either verified or adjusted items classified as clearly one level or the other and borderline. The next cut score recommendation was "the midpoint of the items verified/adjusted as "borderline" averaged across committee members. The results of this process were shared with the members of the committee and discussed. The committee determined either that convergence had occurred or if another iteration was needed. If another iteration was needed it followed the process used in the second round and the average of the items classified as borderline was used as the cut score between performance levels. The Westat (1993) report does not provide information on the number of times a third round of judgments was needed.

Because there were too few test activities in Language Usage and Writing for the committee to evaluate, the cut scores and proficiency level descriptions from 1991 were adopted for 1992. Only one other content area at one grade level had too few test activities at the extremes of performance for a cut score to be set. In Grade 5 Social Studies Levels 4 and 5 were combined into a single level of performance because there were no items classified at Level 5.

Once the cut scores had been set, panelists examined carefully the test activities associated with each proficiency level and independently tried to characterize the knowledge and skills required to complete the activities successfully and that were common to all or most of the activities at that level. As a group, these knowledge and skill descriptions were elaborated, discussed, and refined to attain consensus on the interpretation of the descriptions of the proficiency levels. The knowledge and skills increase in complexity as the level increases from 5 to 1. For example, the grade 3 mathematics Level 5 descriptions suggests that students are able to "color models to demonstrate the meaning of fractional parts" and to "describe a number as odd or even." The Level 1 students are, among other things, able to "develop and apply problem-solving strategies

to solve open-ended problems" and "use basic probability concepts to make predictions in an abstract setting."

Analyzing the content of the activities (items) that measured selected Outcomes in the various subject areas resulted in the Proficiency Level Descriptions. Because not all Outcomes are measured every year, the proficiency level descriptions may not be sufficiently inclusive of the knowledge and skills associated with all the Outcomes until enough time passes that all Outcomes are assessed and proficiency level descriptions are relatively complete. The 1993 MSPAP results measured some Outcomes that were not part of the 1992 assessment. Therefore, the process for setting the cut scores and for determining the Proficiency Level Descriptions was revisited (with some procedural revisions to make the process smoother) to refine the proficiency level descriptions in reading, mathematics, science, and social studies. Revisiting the cut scores was done to correct possible "discrepancies resulting from inconsistencies in item location values … as we obtain more evidence concerning the location of these items." (Atash, 1994, p. 5)
In addition to refining the Proficiency Level Descriptions in the four content areas specified above, writing and language usage cut scores and Proficiency Level Descriptions from 1991 were also revisited because additional test activities in these content areas had been included on the 1993 MSPAP. Even though test activities in these areas had been added, there were some levels that still had too few test activities for a cut score to be set (e.g., grade 3 writing and language usage combined levels 4 and 5).

Critique Of Standard-Setting Procedures For Student Performance

Some of the same criticisms and positive statements made about the institutional standard setting procedures also apply to the procedures for setting cut scores for student performance. Specifically, the number of different reports and the focus of the reports sometimes made finding information difficult. The existence of these reports, however, is commendable and provides an historical record that is necessary for this program. Recall that the major reason for setting the four cut scores was to provide a basis for developing Proficiency Level Descriptions that could

be used to help schools and districts in making institutional instructional decisions (e.g., to help decide what areas of instruction in a school need to be strengthened). It is assumed that after 1993, no substantial revisitation of the cut scores was undertaken because elaborate procedures were used to equate performance across years and across forms within any particular year. Thus, one assumes that, if the scores on each form each year were perfectly reliable, that students classified as performing at Level 3 would perform at that level on any form from any year. This is a stringent assumption and suggests that misclassifications of students (and therefore schools and districts) is possible. As consequences are attached to the MSPAP program, such misclassifications can have serious implications for the schools and districts in Maryland.

Summary, Conclusions and Recommendations

There are three sets of cut scores set in the MSPAP program. Two of these are used to determine if a school's performance is Satisfactory or Excellent; the third is used to classify students into one of five performance levels. The intended use of these five performance levels is to classify students into categories that have been defined in terms of knowledge and skill levels in order to help schools diagnose the need for instructional changes. Large numbers of Maryland educators and patrons were involved in the process of setting the various cut scores and in developing the Performance Level Descriptions.

The MSDE has indicated that the five student performance levels were the first to be defined, followed by the definitions of satisfactory and excellent student performance, and finally the definitions of satisfactory and excellent school performance. At the time these cut scores were set, in the period from 1991 to 1993, the process of setting cut scores in programs such as MSPAP was not as well developed as it is in 2000. The process used was consistent with the state of the art of setting cut scores for complex performance tests like MSPAP.

Conclusions

1. The procedures used to set the various cut scores involved many Maryland educators and school patrons. This level of involvement is commendable. Also

commendable is the extensiveness of the reporting of the procedures used to set the various cut scores, albeit not in a single comprehensive report.

2. The cut scores that were set in the 1991-93 period were consistent with the testing standards and technology of setting cut scores for complex performance tests at that time. These procedures, in general, remain consistent with the current Test Standards (AERA, APA, NCME, 1999), but not with current practice in setting cut scores for complex performance tests.

3. Measurement experts often recommend that cut scores should be revisited whenever there is a substantive change in the program or on a periodic schedule (e.g., every 5 years). Because of changes in the program since 1991 and the adoption of new content standards by the Maryland State School Board, cut scores for all uses should be reset.

Recommendations

1. When cut scores are reset as many Maryland educators and patrons as possible should be involved in the procedures appropriate to their expertise. These procedures should be documented and the results accurately reported in a variety of documents, including a summary in the Technical Report.

2. When new cut scores are set, the current literature in standard setting should be reviewed and state of the art methods selected that will not be construed as being capricious.

3. Studies to examine the validity of the interpretations of the cut scores should be undertaken. Specifically, validity studies should provide evidence that a) students performing at different performance levels are really different in their knowledge and skills, b) that the characterizations of students as performing at satisfactory and excellent levels differ, and that c) schools performing at satisfactory and excellent levels actually differ. Methods and data for evaluating a set of performance standards can be found in the measurement literature. The results of these studies should also be summarized in the Technical Report.

8. Detection of Potentially Biased Test Items (or Differential Item Functioning) Test Development Activities

Issues of bias and sensitivity are important in testing. Professionals have been acutely aware of this for many years, and the strategies for minimizing and detecting items which are potentially biased and/or sensitive are well known. Contractors are aware of these issues, and MSPAP has given considerable attention to the issues of bias and sensitivity.

At the test development stage, task writers were trained with respect to bias and sensitivity issues. Details regarding this can be found in "Chapter 9: Bias and Sensitivity" in the MSPAP Specifications and Procedures Manual for Test and Task Design and Development (Maryland State Department of Education, July, 1999). An early statement in this chapter states that it is basically an adaptation from Bias Issues in Test Development by NES (NES, 1990).[5] Guidelines are provided to test developers regarding language usage, stereotyping, representational fairness, treatment of race and cultures, treatment of the sexes, and treatment of persons with disabilities. Sensitive topics and topics to be avoided are listed. We believe the original NES booklet and the adaptation for the MSPAP are quality documents and we are pleased that they have been used.

In addition to the training of item developers with these materials, MSDE specialists and the task writers also participate in a review and revision process. One of the review criteria relates to controversial and sensitive topics, where assessment and content staff "examine tasks for controversial language, stereotyping, and treatment of minorities, genders, and persons with disabilities" (Technical Report, May, 1999, p. 8).

Statistical Analyses

[5]They do not give a full reference for this document. We assume they mean a booklet entitled Bias Concerns in Test Development published by National Evaluation Systems in 1990.

A differential item functioning (DIF) analysis is performed on the test items (see pages 33-34 of the Technical Report). The measure of DIF is generalized from the Linn-Harnisch procedure (Linn & Harnisch, 1981), which is recognized as an appropriate IRT-based procedure by experts in the profession. The procedure has been especially helpful in situations where the majority group is large, and the minority groups may be considerably smaller. Performances of African-Americans, Asians, and Hispanics are compared to Caucasians, and the performance of females is compared to males. For the DIF analyses, for each item the difference between the predicted and observed examinee success rates (D) was calculated separately in each decile. DIF was defined in terms of the absolute value of D. Items were flagged if this value was greater or equal to 0.10. Also, items were flagged if either the sum of the positive D values or the negative values was greater or equal to 0.10.

A table in the Report shows there were no items flagged as exhibiting DIF for or against African Americans. Only a few items were flagged for the other DIF analyses. However, we would have preferred to see the total distribution of D values. This would allow us to see how many more (or fewer) items would have been flagged under more (or less) stringent flagging rules. Also, it would be useful to have an empirical study showing what impact (if any) a different flagging criterion would have had on the calibrations.

While we are generally pleased with the DIF analyses and the results, we would ideally prefer to have such an analysis at the field test stage so items showing DIF can be reviewed and a decision can be made whether to keep them in the assessment. With the current practice, there would be some reluctance to discard items showing DIF from the scaling process because it could leave the actual assessment with an unbalanced and short assessment. We understand that an earlier DIF analysis is being planned and we approve such a proposed change in test development. Also, we were unable to find whether any tasks showing DIF are in the set of tasks that are reused across years for the purpose of linking. Are they automatically excluded from further

assessments, do they go back to a committee for review, or what? This information should be provided.

Bias Due to Teacher Mobility/Experience

Another issue we have not seen addressed is whether there is a bias against the scores of students (and therefore schools) who have the tasks administered by new teachers. Clearly the administration of the set of tasks is complicated. It seems plausible that new teachers are less able to structure the administration of the tasks in as effective a manner as more experienced teachers. Thus, there is probably a "bias" against the schools that have more turnover of teachers. Certainly the information provided by Lisa Smith in Appendix B of this report would suggest there may be some bias due to teacher mobility.

Fairness as Judged Against the Test Standards

The 1999 Test Standards (AERA, APA, NCME, 1999) have a separate chapter on "Fairness in testing and test use." Many of the relevant standards in this chapter have been met in MSPAP. However, we do note the following standards and make some comments about whether they have been met.

> Standard 7.7 states that
>
> *In testing applications where the level of linguistic or reading ability is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct. (p. 82)*

It is our general impression that this standard has not been met -- at least we could not find sufficient evidence in the materials we studied to convince us it had been met.

Standard 7.11 states that when a construct can be measured in different ways there should be evidence of mean score differences across subgroups when deciding which test to use. We do

not believe this standard was followed when determining what format to use in writing the test questions. While the original Report of the Governor's Commission on School Performance suggested the statewide assessment "not be restricted to machine-scorable tests" (1989, p. 15), the final test only contained constructed response items. We found no evidence that this decision was informed by any evidence of mean score differences across subgroups using the two formats.

Standard 7.12 states that

The testing or assessment process should be carried out so that test takers receive comparable and equitable treatment during all phases of the testing or assessment process. (p. 84)

We do not believe this standard has been met for two major reasons: (1) teachers are not equally competent to administer the test and new teachers are likely at a disadvantage, and (2) there is group interaction to set the context for a test item. Students almost certainly do not receive comparable treatment across groups within a classroom, and across schools.

Conclusions and Recommendations

We are pleased with the attention that MSPAP has paid to the issues of potential bias and sensitivity. We essentially applaud their procedures at the test development stage. This good attention may well account for why so few items have been identified as having differential item functioning. We are also generally pleased with the DIF analysis. They used an acceptable procedure and the results showed there were very few items showing any DIF.

Conclusions

We do have some conclusions that may require changes in current practices for test development and analysis. We list these conclusions below and provide concomitant recommendations in the next section. These conclusions, however, should not detract from our overall positive conclusion regarding how MSPAP proceeded in this important issue of avoiding and detecting potentially biased items.

1. A DIF analysis at the stage of field testing would be preferable to doing it following the actual administration of the test. This would allow the test developers to drop and/or revise items if further review suggested this was warranted.[6]

2. We would have preferred to see the actual distribution of DIF statistics. In addition, it would be helpful to have a study showing the impact on the calibrations of using different flagging rules. Also, the most current report does not show the sample sizes for the various demographic groups.

3. We were unable to discern what was done with items that showed DIF. Were they sent back to a committee? Were these items reused on subsequent tests?

4. We have reason to believe that there is some lack of fairness due to differential levels of experience and expertise among the teachers in administering the assessments. This has not been addressed in any research study we have seen.

5. There is likely some unfairness in individual student and school aggregated scores due to the group interactions that are intended to set the context for a test item.

6. There is likely some confounding of linguistic ability with the assessment of other constructs.

7. While the original case made around the country for performance assessment often suggested that such assessments would decrease the mean difference in performance across various demographic groups, this has not turned out to be the case in most assessments. The MSDE needs to study this issue. It is possible the use of only constructed response items has increased the mean differences across various demographic groups.

Recommendations

1. Carry out a DIF analysis of items at the field test stage. (We understand this is planned for future test development activities.)

2. In future technical reports, provide more information about the distribution of the "D" statistic; present information about the sample sizes used for DIF analyses; and do studies showing the impact of keeping or removing items at different levels of D has on scaling and equating.

3. Make explicit what policy and procedures will be followed regarding items that show significant DIF. We prefer they go back to a committee. They should not be automatically discarded. Items with DIF should have low priority for reuse.

[6]We understand this is being planned for future assessments. Obviously such a plan would have implications for the design of the field test. This is discussed elsewhere in this report.

4. Conduct a study on the effects of teacher experience and expertise in administering the MSPAP.

5. Drop the group pre-assessment activities. They add to potential unfairness. (This recommendation has also been made under the validity section of this report.)

6. Study the effects of the confounding of linguistic ability with the assessment of other constructs. In this section, we recommend this due to the fairness issues, so the study should investigate differential effects across demographic groups.

7. Conduct a study on the differential demographic differences across different item formats.

9. Reliability and Standard Error Measurement

Nine hours of testing over a five-day period results in scores in reading, writing, language usage, math total, math content, math process, science, and social studies. Some of the tasks in the various forms (clusters) assess one content area, but other tasks assess multiple content areas. Activities comprising the tasks may be group or individual activities, but "Group interaction ends before students begin work in their Answer Books, which is always done individually" (MSDE et al., Technical Report, May, 1999, p. 11). According to the Technical Report:

> *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. (May, 1999, p. 8)*

Reliability Estimates

Although the focus of the assessment is on schools, individual scores are available and individual reliabilities and standard errors are reported. In one section of the Technical Report, it is concluded that "scale scores . . . for individual students are not interpretable because each student takes only one-third of the total test" (1999, p. 38).

A variety of methods are available to estimate the degree of consistency or reliability of the scores. One notion of consistency is the consistency of the scoring process. We discuss that in section 5; it is not a part of the discussion of reliability in this section.

Another estimate of reliability is to use an internal consistency estimate such as coefficient alpha. Alpha provides a lower bound estimate of the coefficient of precision. Based on the 1998 Maryland Technical Report, coefficient alpha reliabilities are estimated from a calibration sample of approximately 7,500 students for each content area for each cluster for each grade (grades 3, 5, and 8). Table 18 (p. 69) indicates these reliabilities range from .65 (Cluster A, Grade 3 writing) to .93 (language usage in clusters B and C in Grade 3). As the Technical Report states, the alpha coefficients are generally around .85 except for writing which is generally around .70. Also, the mathematics process scales have lower alphas than the other scales.

Three issues regarding the reliability estimates deserve comment. First, it is our understanding that all the scoring for an individual student for the subject matter for a cluster is done by the same person. Thus, the internal consistency reliabilities reported above may be influenced somewhat by a halo effect (a lack of intra-individual differentiation) in the scoring.

Second, "algorithmic scoring" was done "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" (1999, p. 17). "Algorithmic scoring uses a maximum-likelihood estimation which is a general method of finding good parameter estimates in a model" (p. 17). Although such algorithmic scoring would necessarily increase the internal consistency of the scores for which it is used, it is unclear to us how much (if at all) this scoring may have inflated the coefficient alpha estimates. The Final Technical Report for the 1991 assessment (CTB Macmillan/McGraw-Hill, June 1992) stated that to be included in the calibration analysis "the student had to be present for all days of testing for that book and not require an administration accommodation" (1992, p. 4-1). However, the 1999 Technical Report states that "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. . . . Beginning with the 1995 MSPAP, CTB McGraw-Hill scored all students algorithmically" (1999, p. 17). As a result of this process, "more than 15,000 more scores were computed using algorithmic scoring" (pp. 17-18). We are not told how many (if any) of these scores were in the various calibration samples. However, this is not a large number, and even if they were in the calibration sample they probably would not inflate the alphas to any great extent.

A third issue is that although there are no testlets of dependent items, and although a dependency statistic was computed, we have no way to estimate whether there was an alpha inflation due to dependency. In the 1992 Final Technical Report there was a discussion of the issue of local dependence. If Q3 (a measure of local dependence) was greater or equal to 0.20, a pair of items

"was considered to exhibit local dependence" (1992, p. 4-13). If the items appeared in the same task, they were formed into a testlet so long as the testlet did not produce a "substantial increase" in the misfit statistic. In the 1999 Technical Report it is stated that:

The Q3 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. (Technical Report, 1999, p. 22)

However, there do not have to be testlets specifically and intentionally constructed in order to find local dependence, and no information regarding the results of looking at the Q3 statistic was reported in the Technical Report.

Standard Errors of Measurement

Standard errors of measurement for individual scores are provided for the proficiency level cut scores (Tables 19 - 21)(1999, pp 70 - 72). These SEMs are estimated through the two parameter partial credit model. On a scaled score metric with a mean of 500 and a standard deviation of approximately 50, the SEMs for the various cut scores range from a low of 10 (Cluster 8C for Reading at Level 4/5) to a high of 76 (8C, Reading, Level 1/2). (We remind our readers that there are five levels. Level 1 is the highest and level 5 is the lowest. A student must score 3, 2, or 1 in order to be considered satisfactory or better. Scores of 1 or 2 are considered excellent performance.) Generally the SEMs are higher for Reading and Writing, and are higher for the point Level 1/2. It should be noted that the SEMs vary considerably across the clusters. For example, for Grade 3 Math Process Level 1/2 the SEM was 37 for Cluster 3B and 23 for Cluster 3C.

These standard errors need to be interpreted in light of the range of scaled scores within a level. For example, in grade 3 reading, a score of 490 is the lowest possible score for level 4, a score of 530 for level 3, a score of 580 for level 2, and a score of 620 for level 1. The respective standard

errors for scores at the cut scores for cluster 3A are 16 (level 4/5), 15 (level 3/4), 20 (level 2/3), and 29 (level 1/2). Thus, we are 68% confident that an individual who scores 530 (the 3/4 cut) has a "true" score between 515 and 545; we are 68% confident that an individual who scores 580 (the 2/3 cut) has a true score between 560 and 600; and we are 68% confident that an individual who scores 620 has a true score between 591 and 649.

Standard Errors of Percent Above the Cut-Score (PAC)

One of the MSPAP standard reports is the percent of students at satisfactory and excellent levels of performance. A percent above cut (PAC) is the "percent of students in a school who perform at proficiency level 3 or above (satisfactory or better) in a content area" (Use of confidence intervals . . ., no date, p.1). Standard errors of the PAC are reported. For the 1998 MSPAP, these standard errors range from 3.23 for grade 8 reading for extra large school sizes (n=270) to 9.87 for grade 3 writing for a small (n=30) sample size. The report suggests setting a 90% confidence interval (adding and subtracting 1.65 standard errors around the estimated PAC). The authors use the term "margin of error" to represent 1.65 times the standard error. These range from 5.33 to 16.29. Thus, for example, if a school with a sample size of 30 had a PAC of 50% in grade 3 writing, the PAC 90% confidence interval would be 50 +/- 16.29 or approximately 34% to 66%. If a school with a sample size of 270 in grade 8 reading, the PAC 90% confidence interval would be approximately from 45% to 55%.

The same report also presents the standard error of the difference between PACs. For comparing two PACs, this standard error is the square root of the sum of the squares of the two standard errors. For example, the standard error of the difference between a small and a large elementary school in grade three writing is approximately 12. This can be extended for comparing more than two PACs. Composite indices are averages of percents satisfactory or better. The report provides a table giving margins of error in comparing schools' composite indices. These margins of error range from 3.57 to 6.25. As we point out in Section 11, this across school comparison seems somewhat inappropriate since the purpose of the program is to make within school

comparisons across years. We believe some cautions should have been written into this section of the report.

Reliability and Standard Errors as Judged Against the Test Standards

We commend CTB/McGraw Hill for the information provided about coefficient alphas, standard errors at the cut scores, and the standard errors of the PACs. In comparing the information provided to what is suggested in the Test Standards (AERA, APA, NCME, 1999) we conclude that most of the relevant information suggested in the standards is provided. There are a few standards that are not completely met, but we do not view these omissions as very important. However, for the record, we comment on some standards that seem somewhat relevant yet not totally met.

Standard 2.2 calls for the standard error of measurement to be reported in both raw score and derived scores. We believe only the derived score standard errors are reported. We see no reason to report the standard error of the raw scores, and, indeed this may be confusing to readers.

Standard 2.7 suggests that a stratified alpha would be preferable to the alpha reported. This is not a problem for us. The stratified alphas would be at least as high as the alphas reported.

Standard 2.10 suggests that:

> *when subjective judgment enters into test scoring, evidence should be provided on both inter-rater consistency in scoring and within-examinee consistency over repeated measurements. (AERA, APA, NCME, 1999, p. 33)*

The second portion of this seems unrealistic to us in the extant situation.

Standard 2.15 states that:

> *when a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure. (p. 35)*

In the MSPAP situation, students scores are categorized as above or below a cut score, but the emphasis is on the percent of students in a school above the cut. The MSPAP program does present the standard error of the PAC, but does not present an estimate of the percentage of examinees who would be classified in the same way. We think such an estimate should be provided, as long as individual scores are available upon request.

Conclusions and Recommendations

The reliability estimates we have reported are internal consistency estimates. The information provided suggests reasonably high reliabilities and reasonably low standard errors -- although the standard errors for a few cut points are actually larger than the standard deviation of the scores and many are 40% to 60% as large as the standard deviation of the scores.

Conclusions

Many aspects of both the reporting about and the results of the reliability and standard errors please us and we believe the test scores are sufficiently reliable for their original stated purpose (i.e., to provide information at the school level). Nevertheless, we do have some conclusions below that relate to some recommendations that also follow.

1. The reliability estimates are internal consistency estimates. We do not know the impact of the halo effect of a single reader for a student on these estimates.

2. Algorithmic scoring will inflate internal consistency reliability estimates. We are unsure of the degree of this impact due to the relatively low proportion of students receiving algorithmic scores.

3. We are unsure what any possible item local dependency effects may have had on the alpha estimates though we suspect that local dependencies in the data tend to inflate reliability estimates.

4. In many cases, the standard errors for the individual scores are reasonably large in comparison to the standard deviations and the range of scores within a level.

5. We commend the use of confidence intervals to interpret standard errors of PAC. We wish the technical details would become part of future technical manuals. We believe the reporting of the standard error of the differences between two schools invites potential misuse.

6. We have not been provided information about the percentage of examinees who would be classified in the same way on two applications of the procedure (Standard 2.15). We believe such information would be of value.

Recommendations

1. Conduct a study that would inform users of the impact, if any, of using a single reader for a student's paper on the internal consistency estimates.

2. Provide readers with information about the number of papers (if any) in the calibration sample that were scored algorithmically. Consider excluding such papers from the calibration sample if that is not the current practice.

3. Report the Q3 values and provide readers with an estimate of the changes in reliability estimates if locally dependent items were scored as testlets.

4. Alert readers to the relationship between the size of the standard errors at various cut scores and how these relate to the range of scores within a level.

5. Place the information about the standard errors of PAC in future technical manuals. Insert cautions about the use/misuse of reporting differences between two schools.

6. Provide estimates of the percentage of examinees who would be classified in the same way on two applications of the assessment.

10. Linking of Test Forms Within and Across Years

Linking of test forms from the same test administration to a common score scale and/or linking of test forms from one year to the next are essential for test scores to be comparable. Linking (or sometimes called "test equating") of test forms is always technically challenging and made more difficult in MSPAP because performance assessments create special problems: They may violate the unidimensionality assumption; they often require the use of complex modeling and difficult-to-use item response theory software; reliability of individual scores may not be high (which complicates ability and item parameter estimation), parameter estimation for seldom-obtained item scores is problematic; and dependencies may exist among the test items (which can lead to over-estimation of score reliability), and more. We are going to assume that the technical aspects of the equating are being handled correctly–that is, the actual processing of the test form data, because we will not be able to check on the technical accuracy ourselves. We will limit our attention, instead, to the validity of procedures that are actually in place, and that are described in publications that the Maryland Department of Education and the test contractor have published Relevant standards from the Test Standards (AERA, APA, & NCME, 1999) for our work include 4.10, 4.11, 4.12, 4.14, 4.16, and 4.17.

With the MSPAP, three non-parallel test forms (or clusters) are used each year at each grade level, with each test form consisting of between six and seven tasks. Within a school, the three forms are assigned randomly to students, and careful IRT analyses are carried out prior to proceeding with the linking or equating of forms. This is an excellent equating design for forms within a given year, especially given the very large number of participating students who are used in the actual equating. Horizontal equating is a standard procedure for making scores across test forms comparable. But there are at least three assumptions that are required by these equating methods–(a) that the dimensionality of the test forms are equivalent and preferably unidimensional, (b) that the statistical model fits the test form data, and (c) that the examinees administered each form are equivalent. Assumption three does not appear to be a problem as

long as samples used in the linking remain large, and test forms continue to be randomly distributed within schools.

At the same time, related assumptions one and two seem to be more problematic. One of the concerns whenever equating is done with item response modeling, is the question of model data fit. The evidence in the 1998 Technical Manual is clear and supports the use of a two-parameter polytomous response IRT model. Still, what is not discernible from the Manual are the sizes of the "statistical flags" that are used in detecting any problematic items. Do these "statistical flags" tend to overidentify problematic items, or were they set to identify only the worst offending items? In view of the small numbers of test items that are discarded, the latter may be the case; we can't tell from the information available. What are the practical consequences of these choices for test score equating, score reporting, test development, and so on? We also worry about the extent to which these three test forms used in a particular year are multidimensional, and more importantly, have different dimensional structures from each other. Under these conditions, linking the test forms becomes problematic.

Most performance assessments to some extent are multidimensional, by design really, because states use performance assessments when they aspire to measure many things in an integrated way–notably factual knowledge, higher-level thinking, problem-solving, organizational skills, and writing. By definition, test multidimensionality is increased, and in MSPAP, this multidimensionality may be different across forms because no attempt is made to make these forms "parallel." It is stated in the Test Manual that the IRT model fits the test data , but these findings do not necessarily mean that the test forms are unidimensional. Sometimes the multidimensional structure is hidden by assigning low discrimination indices to the offending test items. This improves the fit, but distorts what the test forms were designed to measure. We have great confidence in the contractor, but recognize that the psychometric challenges being placed on them in equating the forms is substantial. More evidence in the Technical Manual that addresses our concern would be desirable.

More problematic is the technical challenge of linking forms from one year to the next. If this is not done properly, it is impossible to sort out change or growth that reflects changes in student, school, and district performance versus changes due to non-equivalent test forms being used. For example, let's suppose there is a 5-point increase in scores between 1998 and 1999. Are students performing better on the test forms in 1999 because they have learned more, or are the changes due to the use of an easier test in 1999?

The current linking or equating design for test forms across years seems excellent. It consists of two parts. First, adjustments are made in the scores for a given year if it is discovered that the scorers are more or less lenient than scorers in the previous year. This adjustment is based on the rescoring of 1500 student papers from the previous year by a sample of current scorers. These scorers are retrained in a comparable way to the scorers on the test forms the previous year and then asked to score a sample of student papers from the previous year. The implications of any differences in scoring are played out through the actual scaled scores, and then if differences are noted, adjustments to the current scores are made to remove the rater effect.

Whereas representative samples of students are not needed to do successful linking across test forms in a year, or across forms from one year to the next, it does seem important here to have representative samples of scorers from previous and current years participating in this type of study because if differences are observed or not observed (as was the case in 1998), the effect, whatever it is, then is generalized to all scorers in the current year, and used in adjusting scores. More information about the sampling procedure used in selecting scorers would be valuable information for the review process. Also, the design assumes equivalent training of scorers in both the current and the previous year. That the same company or even the same training materials are used, may not be sufficient evidence on which to "hang the hat" of the accountability system in Maryland. Would a stronger design result, if a representative sampling of scorers from the previous year were trained on current material, and then asked to score

current papers? Shouldn't the same effect be observed? If both studies show a 2-point leniency effect for raters or scorers in the current year, the validity of the adjustment would be confirmed with the double design. When the two estimates of the effect size are substantially different, the appropriate action is less clear, but important information has been learned.

The second part of the linking design across years involved administering one of the test forms from the previous year at each grade level to 2500 students in the current year. Then a careful matching is done to find another equivalent sample in the current year to the 2500 students who took one of the previous year's forms. With equivalent samples then, taking the previous and current forms, the two forms can be linked, and then, subsequently, all of the new test forms from the current year are linked into the MSPAP reporting scales. Again, the suitability of all of the linking or equating is dependent on the equivalence of the dimensional structure of the test forms. Multidimensionality is a problem, but even more problematic is different dimensional structures across test forms. More evidence on the structural equivalence of the test forms and the consequences of any lack of equivalence on the validity of the linking would be desirable.

One of the most problematic aspects of the current test design work is that 2/3 of the tasks on an assessment that are used in a given year appeared at some time on past MSPAP administrations. As these tasks are memorable, it is difficult to eliminate the possibility that improvement on the task is not due to prior familiarity. Of course, one way to spot the problem would be if students showed more improvement on one task than another. This might suggest prior exposure. Is prior exposure of tasks a problem in the assessment and more importantly for our work here, is the possibility of prior exposure to tasks taken into account in the linking and scoring processes?

Conclusions and Recommendations

The 1998 Technical Manual excluding tables and references is 40 pages in length. About 25% of this Manual addresses a single topic--test form linking. The Maryland Department of Education and the contractor clearly recognize the importance of this topic, as does their national

psychometric panel. There is every indication that the linking of test forms within and across years is being carefully done, and carefully evaluated. At the same time, there appears to be a major tension between the curriculum specialists who want to generate rich, interesting, and cognitively challenging activities for the assessments and the psychometricians who must carry out the linking or equating of forms to hold the accountability system together. Were the test forms more equivalent in content and statistical characteristics, we suspect that the test form linking would be more accurate. One potential side benefit of matching forms more carefully (or main benefit, depending upon your position) might be the chance to report student scores that could be meaningfully reported in terms of a common set of performance standards.

Four recommendations seem to follow from our review:

1. Because of the pragmatic way in which the tasks are put together each year, often cutting across many subjects and skills, the question of test dimensionality is always a concern. On-going checks each year of model fit and test unidimensionality are essential and are being carried out. But we suspect from our review of many of the assessments themselves and statements about what it is these assessments are intended to measure that the test forms are not strictly unidimensional, and, more importantly, test form dimensionality is somewhat different across forms within a given year, and over years. Linking of forms within years and across years would be stronger if steps were taken at the test specifications and test development stages to construct forms that are more equivalent in their dimensionality and preferably unidimensional. In other words, some of the creativity at the earlier stages needs to be reigned in to permit more valid linking of test forms. In this way, the accountability role of MSPAP will not be compromised. Also, more reporting of the structural equivalence of test forms and the consequences of non-equivalence for linking needs to be reported.

2. Research should be continued to evaluate the "statistical flags" used to allow test items to remain in the assessments that may contribute to less than optimal test score equating and score reporting?

3. Continue to monitor the design for the effects of scorers from one year to the next and their impact on scoring. Also, a "double link" should be considered.

4. More information on "repeated tasks" needs to be reported, especially as it relates to linking of test forms. For example, would conclusions about growth be the same based on old versus new tasks?

11. Score Reporting

Numerous Test Standards (AERA, APA, & NCME, 1999) focus on interpretation and reporting of test scores. Not all these standards are relevant to MSPAP. The most relevant standards related to score reporting focus on such things as the (a) reporting of scores is consistent with intended use of the program (Standard 1.2), (b) equivalence of meaning for different groups when disaggregated scores are reported (Standard 7.8), (c) year-to-year consistency in timing the release of scores (Standard 11.17), (d) inclusion of appropriate interpretative information about released scores, including gain scores (Standards 5.10, 11.18, and 13.17), and (e) availability of data on the psychometric quality (validity and reliability evidence) of the scores (Standard 5.12). Standard 1.2 and 5.12 are particularly relevant because the purposes of MSPAP are to provide school-level accountability and information for instructional improvement. Standard 1.2 states, in part, that "the test developer should set forth clearly how test scores are intended to be interpreted and used" (p.17). Similarly, Standard 5.12 states, in part, that "Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores has been established." (p. 65). Many of the Test Standards related to score interpretation are discussed in Section 6 of this report that evaluates the validity of MSPAP. Because year-to-year comparisons are reported and emphasized in press releases and schools are encouraged to attend to annual gains, Standard 13.17 is also deserving of attention. Standard 13.17 states that "When change or gain scores are used, such scores should be defined and their technical qualities should be reported" (p. 149).

Several score reports are released annually. According to the web page that describes MSPAP, the following types of reports are available: Maryland School Performance Standards Reports, Proficiency Level and Participation Reports, Outcome Score Reports, and Outcome Scale Score Reports. These types of reports are made available for the state, for each district, and for each school.

Score reports are typically released in the fall of the year following the administration of MSPAP. Thus, the 1997-98 school year test results were released in the fall of 1998. This is consistent from year to year. Some of these reports are available at the MSPAP web site. The only publicly released score report made available for examination was the Maryland School Performance Report 1998, State and School Systems (MSDE, 1998). In addition to these types of public reports, the scores of individual students are reported to the schools.

To help interpret the score reports, there is a Score Interpretation Guide (MSDE, 1997a) that is referenced in the Technical Report (MSDE, 1998). A supplemental report for principals was also released in 1997 to aid school principals in communicating with the public about MSPAP (MSDE, 1997b). This supplemental report included questions and answers that could be used to assist principals or teachers when interacting with parents about the interpretation of their child's individual scores.

In addition to the various reports and interpretation guides, press releases are also prepared and distributed by the MSDE. These releases report a variety of results from the program. Some of the results contained in the press releases emphasize particular information that is contained in one or more of the formal reports. For example, a December 1998 press release focuses on "Across-the Board" gains in student performance.

Finally, research reports are also produced, but not necessarily disseminated widely, or publicly. An example of such a report is one that was prepared to aid in developing confidence intervals to interpret standard errors of the percent of students above the cut score (prepared by CTB/McGraw Hill, but no author or date was provided).

Score Reports

The MSDE did not provide a comprehensive list of all reports that are routinely produced, although a list of the types of reports is on the web site. A description of these reports is also shown in the 1998 Technical Report (p. 40) and is summarized below.

Standards reports provide results at the school, district, and state level on the percentage of students who have achieved at the satisfactory or excellent level for each grade level and content area. There are also reports that provide these data disaggregated for race and gender.

Proficiency level and participation reports provide data on the percentage of examinees who score at each of the five proficiency levels and the percentage of students who completed the assessment and received a scale score. These data are also disaggregated by grade level, content area, race and gender.

Outcome score reports provide the average outcome score, or percentage of mastery of an outcome. Also reported is the percentage of students in each of four outcome ranges: 0-25, 26-50, 51-75, and 76-100. These show the percentages of students who have demonstrated little or no mastery of an outcome (0-25) and also the percentage who have demonstrated near complete mastery (76-100). These scores are not comparable across grades or content areas due to difference in the difficulty level of the tests across content areas.

Outcome Scale Score reports contain median outcome scale scores for each learning outcome. These scores are on the same metric for all content areas, thus they can be compared across content areas within grade levels.

In addition to these regular types of reports, other reports are also produced. For example, the 1997 March Supplement (MSDE, 1997b) provides graphic displays of district-level longitudinal data for the 1993-1996 period. The graphs are labeled "MSPAP System Composites." No other

labeling or definition of what the data are, or how they can be interpreted, is provided. One assumes that these graphs reflect the aggregate percentage of students (across all subject areas and grade levels) who have achieved at the satisfactory level. However, an article copied from the December 12, 1996 MSDE Bulletin (shown in MSDE, 1997b) provides graphs for each grade level that show the percentage of students meeting one or more standards and is labeled "MSPAP Gains among Students." There is a similar graph intended to show gains among schools that shows the percentage of schools meeting and "approaching" one or more standard. The interpretation of these graphs and the emphasis of the article are on gains in student performance on MSPAP from 1993 to 1996. A December 8, 1998, press release also focuses on year-to-year gains.

Although not listed as a report, student scale scores and outcome scores are returned to the schools. The 1998 Technical Report (MSDE, et al., 1998) states clearly that "scores for individual students are not interpretable because each student takes only one-third of the total test." (p. 38) The 1997 March Supplement (MSDE, 1997b) provides answers to questions that might be asked by parents about the interpretation of individual student scores. One of the questions is "My child is a good student – Why didn't she/he score high on the MSPAP?" Three possible reasons are provided.

A report Use of Confidence Intervals to Interpret Standard Errors Above the Cut (unnamed author, undated) provides tables for estimating a confidence interval for a school's performance above the cut (PAC) and for comparing a school's year-to-year performance. This report makes an excellent contribution to the assessment program and to the assessment literature in general. The contents of this report should be incorporated into the Technical Report.

Of some concern is that in addition to providing the means to make within school statistical comparisons, methods are provided for comparing two different schools on performance for content areas and for composite scores. This discussion seems inappropriate. The introduction

to the Maryland School Performance Report 1998 (MSDE, 1998) states "the performance of school systems and individual schools are judged against their own growth from year to year, not against growth in other school systems or in other schools" (p. 3). Many of the reports of results at the district and state jurisdictions are constructed in such a way to make cross-district comparisons difficult. The comparison of schools is discouraged. Thus, the illustration of how to use the PAC to compare schools, although it may be accurate, should not precede the illustration describing the process for making the appropriate within school comparisons, which are ostensibly the purpose of the program.

A second problem is that some of the standard errors of the PAC are quite large and are related to school size (the standard errors for large schools are smaller than for small schools). It is not clear in the press releases or in other reports that emphasize year-to-year gain that the standard error of the PAC is taken into account. For example, the 1998 press release cited above contains statements such as "Eighty schools across Maryland recorded a composite score of 70% or better – one-third more schools than last year…" and "Van Bokkelen Elementary School in Anne Arundel County . . . saw strong improvement, with 19.1 percent of its students now performing at the satisfactory level" [up from 8.3 percent in 1995]. Depending on the size of this elementary school, the margin of error for comparing its 1998 composite index with its 1997 composite may be as small as 4.51 (if it is a large school) or as high as 6.25 if it is a small school. If this is a small school and the standard errors for comparing 1996 with 1997 composite scores are similar to the 1997-98 standard errors, then the growth reported for this school (and others like it) could be well within the standard error and the reported "gain" could be due to chance! This possibility is not disclosed in the press release.

It appears that there are several areas of the Test Standards (1999) that are not met in terms of score reporting. These areas related to a) releasing of individual student scores to schools and permitting the schools to provide those scores to parents, and b) reporting and interpreting gain scores without providing appropriate explanation and cautions. Although not necessarily

inconsistent with the Test Standards, providing the means to make cross-school comparisons, which is contrary to the stated purposes of the program, also seems to be a questionable policy.

Summary, Conclusions, and Recommendations

MSPAP provides a number of formal reports that are disseminated through a variety of sources. These reports provide summary information on the status of schools, districts, and the state on the performance levels of students (percent above satisfactory or excellent), on the relative strengths of schools (by reporting summary performance on Maryland's Learning Outcomes), and on the year-to-year gains of individual schools and districts in terms of overall student performance. Individual student scores are provided to schools, although MSDE recommends that these scores not be passed on to parents due to their lack of accuracy. There are also a variety of technical reports that document the program's psychometric qualities and provide information regarding the methodology used to produce the assessment, score the assessment, and evaluate the program.

The production of the public and technical reports is commendable. The broad dissemination of these reports using a wide variety of dissemination strategies, including the Internet, is also commendable. Finally, the public reports we had the opportunity to review were well done, simple to read, and had the potential to be very useful to the various audiences.

Conclusions

We commend the MSDE for its openness in reporting and for the clarity in its reports of the MSPAP program. There are, we believe, some areas that could benefit from closer scrutiny and change.

1. In general, the score reporting in the MSPAP program is in compliance with the Test Standards (AERA, APA, NCME, 1999). However, there are some areas that are questionable in this regard. Specifically, providing individual scores to schools that might be reported to parents is not consistent with the stated policies of MSPAP, it is not consistent with the admonitions in the Technical Report (that indicate these individual scores are not valid), or with the Test Standards. (See Standard 5.12, which indicates that "when group-level information is obtained by aggregating the results of partial tests taken by individuals" that individual's scores should not be reported without evidence of the validity, comparability, and reliability of such scores.)

2. Reporting gain scores without adequately describing the limitations of the interpretation of those scores is not consistent with the recommendations found in the Test Standards. (See Standard 13.17, which indicates that the technical qualities of such scores should be defined and reported if such scores are to be used.)

3. The availability of a methodology to estimate a standard error of the percentage of students above the cut score is a positive advance in the measurement literature. However, the use of this statistic to make between-school comparisons is inconsistent with the stated purposes of the program. (See Standard 1.2, which indicates that the "test developer should set forth clearly how test scores are intended to be interpreted and used." Also relevant is Standard 13.15, which suggests that "reports of group differences should be accompanied by relevant contextual information.")

4. Our impressions are that the approach to score reporting is well done. The materials appear to be clear, comprehensive, and useful. We were impressed with the clarity and simplicity of the public reports, which do not contain complex tables, or extraneous caveats that detract the reader from making reasonable interpretations of the results.

Recommendations

1. The Technical Manual states explicitly that individual scale scores are not interpretable. To the extent that the program continues as it is presently designed (a matrix sampling such that students take only a small portion of the total assessment), these scores should not be reported to the schools. Thus, either the program should be modified in such a way that individual scale scores can be interpreted, or stop reporting these scores to schools.

2. More caution should be exercised in reporting gain scores. Any reported gain scores should not be likely to have occurred by chance.

3. The report describing the methodology for computing a standard error above the cut score should be incorporated into the Technical Report. The method for estimating the standard error that permits comparing two schools should be relegated to an appendix and its use discouraged.

4. Exploration of ways to keep the public reports easily read and understood by the various constituencies for whom the reports are prepared should be continued.

12. Conclusions and Recommendations

What follows are the conclusions and recommendations from sections 3 to 11 in the report.

Choice of Grades

Conclusions

1. The choice of testing in grades 3, 5, and 8 for MSPAP is reasonable from the perspective of providing data on school accountability.

2. In terms of providing information that may be useful in improving instruction, the choice of grades also may be reasonable, but it may not be optimal.

Recommendations

1. Justification for selecting grades 3, 5,and 8 should be elaborated. The justification should indicate the extent that the content of the assessment is reflective of the curriculum of the grade level in which the assessment is administered so that the "burden of responsibility" that may be associated with low performance may not fall exclusively on teachers at the tested grade levels.

Test Development

Conclusions

1. Test development appears to be a strength of MSPAP. Clear documents detail the steps for task developers to use in creating the tasks. However, there is evidence that some of the procedures, although clearly documented, may not be carried through in operation.

2. Multiple reviews are used to identify potential problems and steps are taken to make adjustments and corrections prior to field test. Tasks are reviewed for developmental appropriateness and for bias and sensitivity during test development.

Recommendations

1. External reviews by persons outside the test development process should be conducted on the test specifications. These reviews should focus on where there may be weaknesses in the process that would allow for non-compliance with the stated test specifications and procedures.

2. Better documentation is needed of the membership of review committees, their recommendations, and changes made in response to their recommendations.

3. Better standardization is needed of the assessment materials and manipulatives, both in their acquisition and in their content. It has been suggested that teacher-

prepared materials may be of uneven quality and may therefore jeopardize the standardization of the test experience for students.

4. Test administrators should be given a standardized orientation to the tasks they will be administering and the pre-assessment activities for which they are responsible. Standardization of the pre-assessment preparation is needed.

5. More information is needed about the selection of field test sites. Using sites that are as comparable as possible to Maryland school curriculum, instructional philosophy, grade level, and student motivation enhances the usefulness of the field test information.

Scoring

Conclusions

1. The amount of time between the administration of the MSPAP (mid May) and the release of score reports (late fall) seems consistent with the level of work required. It may be possible to shorten the time by a minimal amount (one-two weeks) without making substantial changes in the nature of the program, however, to try to shorten the time line by more than a week or two might introduce the possibility of errors in reporting that would be damaging to the program.

2. The various reports of the procedures suggest a high degree of compliance with the Test Standards. However, these descriptions do not provide assurance that the procedures were carried out in ways that resulted in high levels of consistency, accuracy, or objectivity of the resulting scores. Specifically

    a. The consistency and accuracy may be overestimated because quality control checks are done at scheduled times, all of which are in the mornings when scorers are fresh.

    b. A single scorer scores all responses for a content area for any student. This may result in a halo effect which would distort reported levels of consistency (and have an impact on the magnitudes of the reliability estimates). There is no systematic rescoring by Scoring Coordinators or Team Leaders for all scorers to check for this halo effect, scorer drift, or inaccuracy.

    c. The Content Panel noted discrepancies between the rubric and how the rubric was interpreted in the selection of Benchmark papers.

Recommendations

1. Attempts to compress the time line for the scoring and reporting should be approached cautiously. Change to the time line could be most effectively undertaken by making substantive changes in the nature of the program. Such changes could include using more objectively scored items (e.g., multiple choice or very short answer). This change would make the program different so it should not be made simply for the purpose of shortening the time between administration of the program and the dissemination of score reports.

2. Additional quality control checks should be introduced.

    a. Quality control checks (check sets and accuracy sets of tasks) should be performed at other than the scheduled times. At least some of these checks should be made at random, unannounced times. Moreover, papers with poor language structure, spelling errors, and poor handwriting should be included to insure scoring bias is not introduced by irrelevant factors.

    b. Scoring Team Coordinators and Team Leaders should undertake to perform random checks on all scorers on their team at random intervals to verify scorer accuracy.

    c. There should be a systematic review of the extent that rubrics are sufficiently specific in what is required to achieve specific scores. The selection of Benchmark papers needs to be consistent with the Maryland Learning Outcomes that are being referenced. To accomplish this an independent panel of subject matter experts should cross validate the quality of rubrics and the appropriateness of the selected Benchmark papers.

Validity

Conclusions

1. Review committees evaluate how well the targeted Maryland Learning Outcomes are incorporated in the assessment tasks. This provides evidence that the assessments are aligned with the Maryland Learning Outcomes.

2. The timing of the administration of the assessments does not necessarily align with the delivery of the relevant content.

3. Although the intent of MSPAP is purportedly to measure the application of knowledge and skills through tasks that require higher order thinking, there is no direct evidence that the tasks do in fact meet this goal. In addition, there is no evidence that the scoring rubrics or scored student work reflect the emphasis on higher order thinking skills.

4. All student responses are constructed. This presents a potential confound in the interpretation of scores as a low score could mean either that the student did a poor job in answering the question or that the student did not have the necessary verbal and/or writing skills to communicate the answer to the question.

5. Pre-assessment tasks are administered in small randomly formed student groups. It is probable that some groups will consist of all high ability or all low ability students. The pre-assessment tasks are used to introduce important information and knowledge that is needed by the individual students to perform well on the assessment tasks. If the groups are not equal in their performance on these pre-assessment tasks, the ability of the individual students to perform well on the operational, scored tasks is compromised. This could have implications for the validity of score interpretations as low scores may result from less that optimal information yielded by the pre-assessment activities that interfered with the students' acquisition of the necessary information to succeed on the task.

6. Many of the assessment tasks require advanced acquisition of and preparation of materials and manipulatives. If these pre-assessment preparations are not done correctly and consistently, students will not receive a fair administration of the tasks. The validity of student and school scores is dependent in part on the fair and equitable administration of the assessment, including the appropriate acquisition and preparation of materials needed for the assessment.

7. Even though the tasks are reviewed for developmental appropriateness, teachers report that the tasks are not developmentally appropriate, especially for the lower grade levels (3 and 5). Performance of students at these grade levels is consistent with the conclusion that the tasks may be overly difficult for these grade level students. It is not clear whether the levels of cognitive challenge (either in the presentation of the tasks or the way responses are made) or the interdisciplinary nature of the tasks may be contributing to low student performance and teacher perception of inappropriate development level of the tasks.

8. Factor analyses have been completed by content area and in most cases are consistent with a unidimensional structure. However, factor analyses across content areas have not been reported. Such analyses would help support the validity of reporting and interpreting separate content area scores.

9. Excessively high between content area correlations (especially when a correction for attenuation is applied) suggest that the interpretation of the scores by content area may be misleading, if not inappropriate.

10. There is reason to believe that MSPAP scores may be differentially valid for students from lower socio-economic groups due to higher teacher turnover and lower student motivation.

11. Teachers new to Maryland may not be well versed in the instructional philosophy that undergirds the MSPAP. Further, they may not be as versed with the kinds of pre-assessment activities for MSPAP as teachers who have experience with

MSPAP. Students with new teachers may have lower scores because of lack of teacher familiarity with MSPAP rather than because of ineffective instructional practices.

12. The inclusion rules for students with disabilities and who are limited in their English language skills are appropriate and consistent with state and federal laws.

13. Reporting of individual student scores is not supported by MSPAP. However, these scores are reported to schools, and on occasion to parents and students. This practice is inconsistent with the intent and design of MSPAP.

14. Gains in student performance across years could be attributed to increased student learning due to instruction or to other possible sources. The scaling and equating procedures should remove changes in test difficulty as a potential reason for increased school performance. However, it is not clear whether gains in performance are in fact due to increased student learning from improved instruction, or possibility to factors such as enhanced test preparation, familiarity with reused assessment tasks, or increased exposure to integrated performance assessment tasks.

15. Evidence is provided that many of the intended consequences of the MSPAP are being realized, especially with regard to instructional reform. Some evidence also shows that unintended negative consequences have also occurred, including lower teacher morale, less emphasis on content areas not covered in MSPAP, and increased teacher stress at the grade levels at which MSPAP is administered. Therefore, evidence indicates that both intended and unintended consequences have occurred due to MSPAP.

Recommendations

1. Information is needed to address how well the Maryland Learning Outcomes are incorporated into the curriculum. Evidence about when these skills are addressed in the curriculum could be gathered by review committees who map the Maryland Learning Outcomes to curriculum guides and other curricular materials.

2. Evidence of when the content measured in the assessment is delivered in instruction needs to be gathered. In earlier versions of MSPAP, a questionnaire addressing this issue was administered to teachers. This questionnaire is no longer administered. Reinstating a questionnaire of this type would provide current information about whether students have the opportunity to learn the content assessed in MSPAP. Other materials, such as a review of curriculum guides, lesson plans, student work, and classroom assessments would also provide evidence of opportunity to learn.

3. Evidence needs to be gathered to support the inference that MSPAP assessment focuses on application of knowledge and skills through tasks that require higher order thinking. In addition to evaluating the tasks for whether they elicit higher order skills, the scoring rubrics and actual scoring of student work should be

evaluated to ensure that they support the conclusion that higher order skills are being assessed.

4. The heavy emphasis on writing as the sole means for a student to communicate his or her answers should be re-considered due to the potential confound in score interpretation. It could be the case that a student has the cognitive skills and knowledge to correctly answer a question, but is limited in the writing skills needed to communicate the answer effectively. This confound prevents direct interpretation of low scores in a content area.

5. The use of group-based, pre-assessment activities should be discontinued as they introduce a potential source of invalidity in the individual student performance on MSPAP tasks.

6. The incorporation of manipulatives in the assessments should be re-considered. If these materials and manipulatives cannot be administered in a more standardized manner, they should be discontinued.

7. More evidence is needed to support the conclusion that the tasks and response requirements are developmentally appropriate, especially at grades 3 and 5. More information about the criteria used in the review of the tasks for developmental appropriateness should be provided, in addition to seeking additional evidence that what is asked and how it is asked of the students is consistent with their cognitive level of development and educational experiences.

8. Additional empirical evidence should be presented on the internal structure of the integrated assessments. Current analyses consider internal structure by content area, rather than for a full assessment.

9. Additional evidence is needed to defend the meaning of the separate content area scores, particularly in light of the high correlations exhibited between content area scores.

10. Because high teacher turnover has been identified as a possible reason for low scores for students from schools with concentrations of lower socio-economic groups, special efforts should be made to retain teachers in these schools and to provide additional orientation and training for these new teachers on MSPAP tasks and curriculum integration.

11. Teachers new to Maryland should be given special orientation to the instructional philosophy that undergirds MSPAP and be provided with training on the pre-assessment activities required to administer MSPAP.

12. The decision to include all students whose educational program aligns with Maryland Learning Outcomes in the assessment, with appropriate administrative accommodations, should be maintained as it is consistent with current law.

13. According to current policy and assessment design, it is inappropriate to report individual student scores to parents and students. Therefore, unless the design of

the assessment is changed, reporting of these scores to schools should be discontinued.

14. Additional information should be gathered to support the inference that the gains realized by schools across years is due to improved student achievement and better instruction. In order to strength this inference, evidence should be gathered to counteract the conclusion that these gains are the result of enhanced test preparation, familiarity with reused assessments, or increased exposure to integrated performance assessment tasks, and not due to real gains in student achievement and teaching quality.

15. More evidence is needed to document both the intended and unintended consequences of MSPAP, including evidence that supports the outcomes of MSPAP in directing instructional reform. No evidence has been provided regarding revamping of the curriculum to include interdisciplinary content. Teacher and principal surveys suggested that unintended negative outcomes have occurred in teacher morale and narrowing of the curriculum. These issues, and others, should be addressed in a planned research program on the consequences of the assessment.

Setting Performance Standards

Conclusions

1. The procedures used to set the various cut scores involved many Maryland educators and school patrons. This level of involvement is commendable. Also commendable is the extensiveness of the reporting of the procedures used to set the various cut scores, albeit not in a single comprehensive report.

2. The cut scores that were set in the 1991 – 93 period were consistent with the testing standards and technology of setting cut scores for complex performance tests at that time. These procedures, in general, remain consistent with the current Test Standards (AERA, APA, NCME, 1999), but not with current practice in setting cut scores for complex performance tests.

3. Measurement experts often recommend that cut scores should be revisited whenever there is a substantive change in the program or on a periodic schedule (e.g., every 5 years). Because of changes in the program since 1991 and the adoption of new content standards by the Maryland State School Board, cut scores for all uses should be reset.

Recommendations

1. When cut scores are reset as many Maryland educators and patrons as possible should be involved in the procedures appropriate to their expertise. These procedures should be documented and the results accurately reported in a variety of documents, including a summary in the Technical Report.

2. When new cut scores are set, the current literature in standard setting should be reviewed and state of the art methods selected that will not be construed as being capricious.

3. Studies to examine the validity of the interpretations of the cut scores should be undertaken. Specifically, validity studies should provide evidence that a) students performing at different performance levels are really different in their knowledge and skills, b) that the characterizations of students as performing at satisfactory and excellent levels differ, and that c) schools performing at satisfactory and excellent levels actually differ. Methods and data for evaluating a set of performance standards can be found in the measurement literature. The results of these studies should also be summarized in the Technical Report.

Bias and Sensitivity

Conclusions

1. A DIF analysis at the stage of field testing would be preferable to doing it following the actual administration of the test. This would allow the test developers to drop and/or revise items if further review suggested this was warranted.

2. We would have preferred to see the actual distribution of DIF statistics. In addition, it would be helpful to have a study showing the impact on the calibrations of using different flagging rules. Also, the most current report does not show the sample sizes for the various demographic groups.

3. We were unable to discern what was done with items that showed DIF. Were they sent back to a committee? Were these items reused on subsequent tests?

4. We have reason to believe that there is some lack of fairness due to differential levels of experience and expertise among the teachers in administering the assessments. This has not been addressed in any research study we have seen.

5. There is likely some unfairness in individual student and school aggregated scores due to the group interactions that are intended to set the context for a test item.

6. There is likely some confounding of linguistic ability with the assessment of other constructs.

7. While the original case made around the country for performance assessment often suggested that such assessments would decrease the mean difference in performance across various demographic groups, this has not turned out to be the case in most assessments. The MSDE needs to study this issue. It is possible the use of only constructed response items has increased the mean differences across various demographic groups.

Recommendations

1. Conduct a DIF analysis of items at the field test stage. (We understand this is planned for future test development activities.)

2. In future technical reports, provide more information about the distribution of the "D" statistic; present information about the sample sizes used for DIF analyses; and do studies showing the impact of keeping or removing items at different levels of D has on scaling and equating.

3. Make explicit what policy and procedures will be followed regarding items that show significant DIF. We prefer they go back to a committee. They should not be automatically discarded. Items with DIF should have low priority for reuse.

4. Conduct a study on the effects of teacher experience and expertise in administering the MSPAP.

5. Drop the group pre-assessment activities. They add to potential unfairness. (This recommendation has also been made under the validity section of this report.)

6. Study the effects of the confounding of linguistic ability with the assessment of other constructs. In this section, we recommend this due to the fairness issues, so the study should investigate differential effects across demographic groups.

7. Conduct a study on the differential demographic differences across different item formats.

Score Reliability

Conclusions

1. The reliability estimates are internal consistency estimates. We do not know the impact of the halo effect of a single reader for a student on these estimates.

2. Algorithmic scoring will inflate the internal consistency estimates. We are unsure of the degree of this impact due to the relatively low proportion of students receiving algorithmic scores.

3. We are unsure what any possible item local dependency effects may have had on the alpha estimates though we suspect that local dependencies in the data tend to inflate reliability estimates.

4. In many cases, the standard errors for the individual scores are reasonably large in comparison to the standard deviations and the range of scores within a level.

5. We commend the use of confidence intervals to interpret standard errors of PAC. We wish the technical details would become part of future technical manuals. We believe the reporting of the standard error of the differences between two schools invites potential misuse.

6. We have not been provided information about the percentage of examinees who would be classified in the same way on two applications of the procedure (Standard 2.15). We believe such information would be of value.

Recommendations

1. Conduct a study that would inform users of the impact, if any, of using a single reader for a student's paper on the internal consistency estimates.

2. Provide readers with information about the number of papers (if any) in the calibration sample that were scored algorithmically. Consider excluding such papers from the calibration sample if that is not the current practice.

3. Report the Q3 values and provide readers with an estimate of the changes in reliability estimates if locally dependent items were scored as testlets.

4. Alert readers to the relationship between the size of the standard errors at various cut scores and how these relate to the range of scores within a level.

5. Place the information about the standard errors of PAC in future technical manuals. Insert cautions about the use/misuse of reporting differences between two schools.

6. Provide estimates of the percentage of examinees who would be classified in the same way on two applications of the assessment.

Test Score Linking

Conclusions

1. There is every indication that the linking of test forms within and across years is being carefully done, and carefully evaluated.

2. There appears to be a major tension between the curriculum specialists who want to generate rich, interesting, and cognitively challenging activities for the assessments and the psychometricians who must carry out the linking or equating of forms to hold the accountability system together. Were the test forms more equivalent in content and statistical characteristics, we suspect that the test form linking would be more accurate.

Recommendations

1. Because of the pragmatic way in which the tasks are put together each year, often cutting across many subjects and skills, the question of test dimensionality is always a concern. On-going checks each year of model fit and test unidimensionality are essential and are being carried out. But we suspect from our review of many of the assessments themselves and statements about what it is

these assessments are intended to measure that the test forms are not strictly unidimensional , and more importantly, test form dimensionality is somewhat different across forms within a given year, and over years. Linking of forms within years and across years would be stronger if steps were taken at the test specifications and test development stages to construct forms that are more equivalent in their dimensionality and preferably unidimensional. In other words, some of the creativity at the earlier stages needs to be reigned in to permit more valid linking of test forms. In this way, the accountability role of MSPAP will not be compromised. Also, more reporting of the structural equivalence of test forms and the consequences of non-equivalence for linking needs to be reported.

2. Research should be continued to evaluate the "statistical flags" used to detect potentially flawed test items. Are the "statistical flags" too lenient, and therefore allow test items to remain in the assessments that may contribute to less than optimal test score equating and score reporting?

3. Continue to monitor the design for the effects of scorers from one year to the next and their impact on scoring. Also, a "double link" should be considered.

4. More information on "repeated tasks" needs to be reported, especially as it relates to linking of test forms. For example, would conclusions about growth be the same based on old versus new tasks?

Score Reporting

Conclusions

1. We commend the MSDE for its openness in reporting and for the clarity in its reports of the MSPAP program. There are, we believe, some areas that could benefit from closer scrutiny and change.

2. In general, the score reporting in the MSPAP program is in compliance with the Standards (AERA, APA, NCME, 1999). However, there are some areas that are questionable in this regard. Specifically, providing individual scores to schools that might be reported to parents is not consistent with the stated policies of MSPAP, it is not consistent with the admonitions in the Technical Report (that indicate these individual scores are not valid), or with the Standards. (See Standard 5.12, which indicates that "when group-level information is obtained by aggregating the results of partial tests taken by individuals" that individual's scores should not be reported without evidence of the validity, comparability, and reliability of such scores.)

3. Reporting gain scores without adequately describing the limitations of the interpretation of those scores is not consistent with the recommendations found in the Standards. (See Standard 13.17, which indicates that the technical qualities of such scores should be defined and reported if such scores are to be used.)

4. The availability of a methodology to estimate a standard error of the percentage of students above the cut score is a positive advance in the measurement literature. However, the use of this statistic to make between-school comparisons is inconsistent with the stated purposes of the program. (See Standard 1.2, which indicates that the "test developer should set forth clearly how test scores are intended to be interpreted and used." Also relevant is Standard 13.15, which suggests that "reports of group differences should be accompanied by relevant contextual information.")

5. Our impressions are that the approach to score reporting is well done. The materials appear to be clear, comprehensive, and useful. We were impressed with the clarity and simplicity of the public reports, which do not contain complex tables, or extraneous caveats that detract the reader from making reasonable interpretations of the results.

Recommendations

1. The Technical Manual states explicitly that individual scale scores are not interpretable. To the extent that the program continues as it is presently designed (a matrix sampling such that students take only a small portion of the total assessment), these scores should not be reported to the schools. Thus, either the program should be modified in such a way that individual scale scores can be interpreted, or stop reporting these scores to schools.

2. More caution should be exercised in reporting gain scores. Any reported gain scores should not be likely to have occurred by chance.

3. The report describing the methodology for computing a standard error above the cut score should be incorporated into the Technical Report. The method for estimating the standard error that permits comparing two schools should be relegated to an appendix and its use discouraged.

4. Exploration of ways to keep the public reports easily read and understood by the various constituencies for whom the reports are prepared should be continued.

13. References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). Standards for Educational and Psychological Tests, Washington, DC: American Educational Research Association.

Atash, N. (1994). Establishing proficiency levels and descriptions for the 1993 Maryland School Performance Assessment Program (MSPAP). Rockville, MD: Westat, Inc.

Council of Chief State School Officers. (1997). Annual Survey of State Student Assessment Programs Volume 1: Fall 1997 Data on 1996-97 Statewide Student Assessment Programs. Washington, DC: Author.

Creech, J. D., et al. (2000). Educational benchmarks 2000. Atlanta, GA: Southern Regional Education Board.

CTB/Macmillan/McGraw Hill. (1992, June). Final technical report Maryland School Performance Assessment Program 1991. Monterey, CA: Author.

Firestone, W.A., Mayrowitz, D., & Fairman, J. (1997, April). Performance-based assessment and procedural reform: The limited effects of testing in Maine and Maryland. Paper presented at the meeting of AERA, Chicago, IL.

Green, B.F. (1993). The structural validity and generalizability of the 1992 Maryland School Performance Assessment Program. Baltimore: Maryland State Department of Education.

Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. (1995). Review of the measurement quality of the Kentucky Instructional Results Information System, 1991-1994. Frankfort, KY: Office of Educational Accountability, Kentucky General Assembly.

Koretz, D. (1997, October). Reactions to correspondence with MSDE concerning MSPAP. Memorandum to Catherine Walsh dated October 9, 1997.

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996, March). Project 3.2, state accountability models in action, perceived effect of the Maryland School Performance Assessment Program (Final Report, U.S. Department of Education Office of Education and Research and Improvement Grant No. R117G10027 CFDA Catalog No. 84.117G). Los Angeles, CA: National Center for Research on Evaluation Standards, and Student Testing (CRESST), Graduate School of Education and Information Studies, University of California-Los Angeles

Lane, S., Parke, C. S., & Stone, C. A. (1998, April). Consequences of the Maryland School Performance Assessment Program. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Lane, S., Ventrice, J., Cerrillo, T. L., Parke, C. S., & Stone, C. A. (April, 1999). Impact of the Maryland Performance Assessment Program (MSPAP): Evidence from the Principal, Teacher, and Student Questionnaires (Reading, Writing, and Science). Paper presented at the meeting of the National Council on Measurement in Education, Montreal.

Linn, R.L. & Harnisch, D. (1981). Interactions between item content and group membership in achievement test items. Journal of Educational Measurement, 18, 109-118.

Maryland State Department of Education, CTB/McGraw-Hill (1992). Final Technical Report, Maryland School Performance Assessment Program, 1991. Baltimore: MSDE/CTB/McGraw Hill.

Maryland State Department of Education, CTB/McGraw-Hill, Measurement Inc. (1993, May). Pre-Release Edition, 1992 Technical Report Maryland School Performance Assessment Program, Baltimore, MD.

Maryland State Department of Education (1994, February). 1993 Maryland school performance assessment program Score Interpretation Guide. Baltimore, MD: Author.

Maryland State Department of Education, CTB/McGraw-Hill, Measurement Inc., University of Maryland Baltimore County Center for Educational Research and Development, Westat, Inc. (1994, March). 1993 Technical Report Maryland School Performance Assessment Program, Baltimore, MD.

Maryland State Department of Education, CTB/McGraw-Hill, Measurement Inc., University of Maryland Baltimore County Center for Educational Research and Development, Westat, Inc. (1995, March). 1994 Technical Report Maryland School Performance Assessment Program, Baltimore, MD.

Maryland State Department of Education, CTB/McGraw-Hill, Measurement Inc. (1996, January). 1995 Technical Report Maryland School Performance Assessment Program, Baltimore, MD**.**

Maryland State Teacher Association MSPAP Survey (1997, June).

Maryland State Department of Education. (1997a). Score Interpretation Guide, Maryland School Performance Assessment Program – 1997 MSPAP and Beyond. Baltimore: Author.

Maryland State Department of Education. (1997b). MSPAP Communications Packet for Principals. March 1997 Supplement. Baltimore: Author.

Maryland State Department of Education (1998). Maryland school performance report 1998 state and school systems. Baltimore, MD: Author.

Maryland State Department of Education, CTB/McGraw-Hill, Measurement Inc. (1999, May). 1998 Technical Report Maryland School Performance Assessment Program, Baltimore, MD**.**

Maryland State Department of Education (1999, July). Maryland School Performance Assessment Program, Specifications and Procedures Manual for Test and Task Design and Development.

Maryland State Department of Education. (1999). Specifications and procedures manual for test and task design and development. Baltimore: Author.

Maryland State Department of Education, CTB/McGraw Hill, Measurement Incorporated, University of Maryland Baltimore County, Center for Educational Research and Development, & Westat, Inc. (1999, May). 1998 Maryland School Performance Assessment Program Technical Report. Baltimore: Maryland State Department of Education.

Maryland State Teacher Association MSPAP Survey. (1997, June).

Measurement Incorporated. (1998, November). 1998 Maryland School Performance Assessment Program Scoring Report. Durham, NC: Author

National Evaluation Systems, Inc. (1990). Bias concerns in test development. Amherst, MA: National Evaluation Systems, Inc.

No author attribution, undated. Use of confidence intervals to interpret standard errors of percent above the cut (PAC) for the 1998 Maryland School Performance Assessment Program Results. Available from MSDE, Baltimore.

Task Force on the MSPAP Testing Program. (1998). Report to the 1998 MSTA Representative Assembly.

Thorn, P., Moody, M., McTighe, J., Kelly, N., & Peiffer, R. (1990). Establishing standards for Maryland's school systems: A systematic approach. Baltimore, MD: Maryland State Department of Education.

Use of Confidence Intervals to Interpret Standard Errors of Percent Above Cut (PAC) For the 1998 Maryland School Performance Assessment Program Results (no date). Xerox.

Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity issues in collaborative performance assessment: Group composition and performance. American Educational Research Journal, 35, 607-651.

Westat, Inc. (1993). Establishing proficiency levels and descriptions for the 1992 Maryland School Performance Assessment Program (MSPAP). Rockville, MD: Author.

Table 1

Grades in which MSPAP-like State Assessment Programs were Administered in 1996-97 (CCSSO, 1997)

State	Reading	Writing/L.A.	Mathematics	Science	Social Studies
AK	4,8,11	4,8,11 (5,7,10)**	4,8,11		
AL	3 THRU 11	5,7 (3,9,11)	3 THRU 11	3 THRU	3 THRU 11
AR	5,7,10 (8)	5,7,10 (8)			
AZ	3 THRU 12	3 THRU ?	? THRU ?		
CA	4,5,8,12	4,5,8,12	4,5,8,12	4,5,8,12	4,5,8,12
CO	4 (3)	4	(5)		
CT	4,6,8	4,6,8	4,6,8		
DE	(3,5,8,10)	3,5,8,10	(3,5,8,10)	(4,6,8,11)	(4,6,8,11)
FL	4,8 (10)	4,8,10	(5,8,10)		
GA	3,5,8 (4,6,8)	3,5,8 (4,6,8)	3,5,8 (4,6,8)		
HI	3,6,8,10	3,6,8,10	3,6,8,10(3,5,7,	3,6,8,10	3,6,8,10
IA	No Mandated State Assessment Program				
ID	3 thru 11	3 thru 11; 4,8,11	3 thru 11; 4,8	(6,10)	(6,10)
IL	3,6,8,10	3,6,8,10	3,6,8,10	4,7,11	4,7,11
IN	3,6,8,10	3,6,8,10	3,6,8,10		
KS	3,7,10	5,8,10	4,7,10	5,8,10	5,8,11
KY	Revised program – New program not described				
LA	3,5,7, (4,8)	3,5,7 (4,8)	3,5,7 (4,8)	(4,8)	(4,8)
MA	4,8,10	(4,8,10)	4,8,10	4,8,10	(4,8,10)
MD	3,5,8	3,5,8	3,5,8	3,5,8	3,5,8
ME	4,8,11	4,8,11	4,8,11	4,8,11	4,8,11
MI	4,7,11	5,8,11	4,7,11	5,8,11	
MN	(3,5)	(3,5)	(3,5)		
MO	(3,7,11)	(3,7,11)	4,8,10	(3,7,10)	(4,8,11)
MS	4 thru 9	4 thru 9	4 thru 9	4 thru 9	4 thru 9
MT	4,8,11	4,8,11	4,8,11	4,8,11	4,8,11

NC	3 thru 8	4, 7	3 thru 8		
ND	3,6,8,11	3,6,8,11	3,6,8,11	3,6,8,11	3,6,8,11
NE	No Mandated State Assessment Program				
NH	3,6,10	3,6,10	3,6,10	3,6,10	3,6,10
NJ	4,8,11	4,8,11	4,8,11	4,8,11	(4)
NM	3,5,8	4,6	3,5,8		
NV	4,8 (10)	4,8 (10)	4,8 (10)	(4,8,10)	
NY	4,8	4,8	4,8		
OH	4,6,9,12	4,6,9,12	4,6,9,12	4,6,9,12	4,6,9,12
OK	3,5,7,8,11	3,5,7,8,11	3,5,7,8,11	3,5,7,8,11	3,5,7,8,11
OR	3,5,8,10	5,8,10	3,5,8,10		
PA	5,8,11	6,9	5,8,11		
RI	4,8,10	4,8,10	4,8,10		
SC	3 THRU 11	3 THRU 11	3 THRU 11		
SD	4,8,11	4,8,11	4,8,11	4,8,11	4,8,11
TN	2 thru 8	4,8,11	2 thru 8		
TX	3 thru 8	4,8	3 thru 8	8	8
UT	5,8,11	5,8,11	5,8,11	?	?
VA	3,5,8	3,5,8	3,5,8	3,5,8	3,5,8
VT	?	5,8	4,8,10		
WA	4,8,11	4,8,11	4,8,11	4,8,11	4,8,11
WI	3,4,8,10	4,8,10	4,8,10	4,8,10	4,8,10
WV	1 thru 11	1 thru 11	1 thru 11	3 thru 11	3 thru 11
WY	Revised program – New program not described				
*Programs change rapidly, thus these data may not be accurate for 1999.					
**Grade levels shown in parentheses are planned, rather than operational.					