

Assignment: Final paper

Choice: The Maryland State Performance Assessment Program

INTRODUCTION

The Maryland School Performance Assessment Program, or MSPAP, has been a topic of conversation both in Maryland and nationally. This test is different from many other standardized tests currently in use in that the MSPAP, though administered to individual students, is primarily used to measure school or district performance. The test is supposed to assess how well the students in individual schools are meeting the statewide standards for satisfactory and excellent performance, and provide data to help schools diagnose weaknesses and plan for the resolution of these problems. The state as a whole also avers that it uses the data as an “impetus for many state sponsored programs and legislation to counter discouraging performance trends” (MSPAP, 2001).

A second interesting point about the MSPAP is that the state of Maryland seems to be using this test as a way to change instruction in this state. In fact, the MSPAP web-site states, “Teaching to the test is good instruction, the kind of instruction that results in understanding, not in the mere rote recall of isolated facts”. Maryland seems to agree with Frederickson and Collins (1990) who state that “a systemically valid test is one that induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the

test is designed to measure”. A question that could be asked is if there is any evidence that one consequence of these assessments is that the students’ cognitive skills are stronger rather than that the students have just become better test takers. Inspection of the test scores in most counties does show that the scores have generally increased over the last several years but making the leap from higher test scores to higher levels of cognitive ability will require quite a bit more evidence.

A third point of interest is that the MSPAP utilizes performance assessments as opposed to the dichotomously scored items found on most standardized tests. In order to respond to the problems put forth on the MSPAP students may use short verbal responses, essays, drawings, tables, or graphs rather than simply bubbling in a multiple choice option. It is hoped that the nature of the questions used on the MSPAP exams will allow students to apply their skills and knowledge to solve real world problems. In other words, the MSPAP is supposed to measure higher order thinking skills. It certainly seems that the structure of the MSPAP items lend face validity to the assessments which in turn helps the students to make the connection between their academic studies and the real world. But whether or not the test measures higher-order thinking is open for debate.

In order to make sense of the MSPAP the evidence-centered assessment design strategy (Mislevy, et. al., 1999) will be applied to the test. This strategy offers a framework for designing assessments that can, in this case, be applied to evaluating a test. Three models, the student model, the evidence model and the task model, are encompassed within this framework. Essentially, the student model identifies the relevant characteristics of the individual being tested such as knowledge, skills and abilities. The evidence model looks at the “work products” that will be supplied by these individuals. This model establishes the argument for reasoning from a

work product to “(1) what’s important about it and (2) how it revises beliefs” about students. The task model specifies the types of tasks that are necessary to get the evidence needed about students. The success of an assessment rests in the strength of the links between each of these models. In order to determine how successful the MSPAP is at meeting its objectives the evidence-centered assessment strategy will be employed and these links will be examined and evaluated.

STUDENT MODEL

In order to examine the relevant characteristics of the individuals being assessed by the MSPAP, both cognitive psychology (the rationalist perspective) and situative psychology (the pragmatic-sociohistoric perspective) can be applied to the test. It should be noted that these perspectives are not mutually exclusive, but rather they frame issues in distinctive but complementary ways. By applying both of these perspectives on learning and knowing a more complete picture of the student will emerge. Hopefully, this will lead to a more complete picture of the Maryland State Performance Assessment Program as a whole.

Perhaps the most obvious perspective to apply to the MSPAP is that provided by cognitive psychology. The fact that the questions on the MSPAP exams ask students to do things such as furnish a solution to a problem, or offer an explanation or a rationale for a response, provides us with some evidence of the nature of this assessment. Further evidence is found when one considers the design of the tasks on the MSPAP exams. According to MSPAP, the performance tasks on the exam “are designed to elicit the thinking processes and skills described in the ‘Dimensions of Thinking’, developed by the Association for Supervision and Curriculum Development”. The ASCD have identified five dimensions of thinking that they feel

reflect the common threads running through most current research and theory in the field of cognition. Their five dimensions of thinking are:

- Metacognition
- Critical and creative thinking
- Thinking processes
- Core thinking skills
- The relationship of content-area knowledge to thinking

Clearly these terms come directly from cognitive psychology in which the concern is what goes in the students' minds during instruction and assessment. So, from the cognitive perspective we can say that the relevant characteristic of the student to be measured is the ability to apply what has been learned to solve problems.

The cognitive perspective is not the only one that can be applied to the MSPAP. The pragmatist-sociohistoric perspective taken from situative psychology can also be applied to this exam. From this point of view, thinking and learning are seen in the context of the community. Students are expected to use their knowledge and skills to participate in the work of the community and in doing so learn the habits of mind and standards of that community. According to the MSPAP web-site, one of the characteristics of the tasks is that they “engage students in real-world activities”. Examination of some of the tasks on the science portion of the exam provides examples of this premise in action. One task asks students to imagine that they are scientists working at the Goddard Space Flight Center in Maryland. Another has them solving problems by examining real salinity maps of the Chesapeake Bay. By structuring the tasks in this manner, the MSPAP encourages students to see themselves as part of a community of scientists who work together to solve problems in the real world. By making this connection between the classroom and beyond, the MSPAP is helping to motivate students to acquire the

knowledge and skills that will make them productive members of the community. Clearly this is not a characteristic of the individuals being assessed that will actually be measured by the test. Still it is hoped that after the students take the tests they will appreciate their place in the community of scientists and learners. For that reason, this must be included as part of the student model.

EVIDENCE MODEL

Unlike other standardized tests in which the work products are dichotomously scored, the work products associated with the MSPAP may be either dichotomously or polytomously scored. Some of the questions on the exam require brief responses that are graded as either correct or incorrect (i.e. 0,1). Other questions however require longer responses that may receive up to four points for a perfect response, and partial credit for responses that are less than ideal. Questions that fit this profile include drawings, tables, graphs and essays. In the case of the MSPAP the task of determining what in the work product is meaningful falls on specially trained teachers using scoring keys for the brief responses and rubrics for the extended responses. Interestingly, most student papers are read and graded only one time by each subject matter reader. While scoring accuracy is maintained through a variety of methods such as spot checks, the fairness of the scoring procedure is ensured by randomly sorting the test answer books to ensure that each rater scores books from a variety of schools and districts (MSPAP, 2000). Though a scorer within a subject reads the answers only once, they may be read a second time by a scorer in a different subject area. For example, an answer book may be scored one time for the science content and then a second time on language usage.

The varied nature of the students' responses on the MSPAP requires that an adaptation of traditional item response (IRT) models be used. In order to accommodate the fact that items may be awarded a varying numbers of points the Master's Partial Credit model was chosen (MSPP,1998). Though this model takes care of the issue of variety of scoring points, another adaptation needs to be made. The Master's model is essentially a modified Rasch model that forces all items to have the same discrimination. This is clearly not suitable for the MSPAP, which contains items that vary a great deal in terms of their discriminations. This is particularly true of the items that require the use of scoring rubrics. To solve this problem a generalization of the Master's model was developed called the two-parameter partial credit model. This model states that the probability of a respondent with ability θ having a score at the k-th level of the j-th item is:

$$P_{jk}(\mathbf{q}) = P(x_j = k - 1 / \mathbf{q}) = \frac{\exp Z_{jk}}{\sum_{i=1}^{m_j} \exp Z_{ji}}, k = 1, \dots, m$$

In this equation m_j is the number of score levels and $Z_{jk} = \alpha_j(k-1) - \sum \gamma_{ji}$. The variables α_j and γ_{ji} are the parameters to be estimated from the data. This is implemented using PARDUX software that estimates the item parameters simultaneously for the dichotomous and polytomous items. This program utilizes marginal maximum likelihood estimation procedures implemented via the Expectation-Maximization (EM) algorithm. One can think of this as essentially a Bayesian approach in that the posterior distribution for each item parameter is obtained using Bayes theorem (Roberts, 2001).

After the item parameters are found they are used in the determination of the student scale scores. In this procedure the score that a student receives on an individual item is weighted

by that item's discrimination coefficient. For each student, summing the weighted scores on each item provides a raw score. These are converted into student scale scores that are set to have an approximate mean of 500 and a standard deviation of 50. However, since a variety of forms are used within each grade on the MSPAP this is not the end of the procedure. To allow for scores on the different forms to be compared an equipercentile equating procedure is carried out. This sort of procedure is utilized rather than linear equating using an IRT approach as one might suspect because there are few items on the MSPAP and those items often have unusual score distributions. This makes linear equating, which typically involves comparisons of means and standard deviations, a poor choice (MSPAP, 1998). Finally, student outcome scores can be determined. These are a percentage of the maximum score that a student would be expected to obtain if the student had taken all items on all forms in a content area. This EPM (expected percentage of maximum) is determined by procedure which essentially adjusts an individual student's score for the difficulty of the questions administered to that student (Yen, 1997). These outcome scores, which are reported in four ranges (0-25,26-50,51-75 and 76-100), are determined for each student.

Since MSPAP is a test designed to measure school performance it is necessary to somehow take this data about students and turn it into data about schools. This is done in several ways. One report that schools receive contains the average outcome score and the percentages of students in each range for each content area. The standard error of the school means for these outcomes is generally in the range of 2 to 7 points (Yen, 1997). Schools also receive reports detailing mean scale scores for students. These school means are quite accurate with standard errors of 5,6, and 7. Finally, schools receive reports of the proficiency levels (1 through 5) of

the students. It is the percentage of students at each proficiency level that provides the actual basis on which schools are judged by the state.

There are five scale score ranges of proficiency on the MSPAP. In determining the cut scores for these ranges it was necessary to once again look at the item level. Every MSPAP item was placed on the scale score range at the position at which it provides the most measurement information (Yen, 1997). As it turns out large numbers of items were located at four scale score levels, 490, 530, 580 and 620. These values were established as the proficiency level cut-points in 1991 and in most content areas they remain the cut points today. By examining the items within each proficiency level, committees of content experts have developed descriptions of the knowledge, skills and processes that students at each level typically display. For a school to receive a “satisfactory” rating (or a 3) in a content area, 70% of the students must have scores above 530. For an “excellent” rating the school must also have 25% of its students above 580. These percentages seem to have been arbitrarily determined.

TASK MODEL

As its name indicates, the MSPAP use performance assessments rather than the objectively scored tasks traditionally found on large-scale tests. The question is why? Clearly these tasks are much more difficult to administer and grade, so what rationale is there for using them? One reason is that the MSPAP is an exam designed to elicit thinking processes rather than simply recall of knowledge. By applying skills and knowledge to solve problems and make decisions students utilize higher-order thinking skills. Their thought processes become visible because they are asked to explain the process they went through to reach an answer. A cynic could argue that well written multiple-choice exams can also measure higher-order thinking so

why use performance assessments? While multiple-choice tasks can, in fact, provoke complex thinking, they cannot capture very much evidence about it. Hence the need for a performance assessment. A second rationale for using these types of tasks is that by doing so the state of Maryland will be able to change instruction by classroom teachers. In fact, the MSPAP web-site states, “Teaching to the test is good instruction, the kind of instruction that results in understanding, not in the mere rote recall of isolated facts”. It is important to remember that in the final analysis the MSPAP is a test of schools, and as such the items on the test should directly impact those schools.

Saying that the items on the MSPAP are performance assessments only begins to describe what is presented and how, as well as what is presented and evaluated. Some characteristics of the MSPAP items are detailed below with an example of how each characteristic is shown in the Grade 8 task titled Planetary Patterns.

- *The tasks consist of multiple outcomes* – For example both Science and Language Usage are tested within the same task, although using different rubrics.
- *The tasks are set in real-world contexts* – The students are supposed to act like scientists working at the Goddard Space Flight Center in Maryland.
- *The tasks are appropriate to the discipline* – For example, students are required to read and interpret charts and communicate their beliefs; two attributes that are very important in science.
- *The tasks are theme-based* – Students are graded on 10 questions that all relate to the same central idea of planetary patterns.
- *The tasks elicit numerous student performances and varied modes of response* – Students need to read a chart, interpret data, draw diagrams and write paragraphs to successfully respond.

One final characteristic of all of the tasks on the MSPAP is that they reflect the Learning Outcomes specified by the state. For example, in this case one of the science skills expected of students in Grades 6-8 is that they are able to “explain that the motion of the objects in the solar

system is regular and predictable and explains phenomenon”. The degree to which students understand this concept is clearly being measured in the Planetary Patterns tasks.

As is the case with most large-scale tests, students are administered the MSPAP tasks in a group setting under standardized conditions. That is, the students are all read the same instructions by the exam proctors and they are under the same time constraints. The rules for individual tasks are dependent on the nature of that task however. Most tasks require the students to work individually but there are some tasks on the MSPAP that allow for group interaction. While most of the tasks require only paper and pencil, some do allow for the use of manipulatives to either aid in the student’s thinking or as an essential part of the task. For example, in the case of the Planetary Patterns task, students were encouraged to move pennies or poker chips to simulate the motion of the planets. The activities are generally interpretive in that they require that the students respond to information that is contained in introductory paragraph or presented to them in a graph or a chart.

The work product that is captured after the administration of the MSPAP is an *Answer Book* for each student which contains hand-written answers to each of the activities within the task. Like the activities that produced them, the individual answers to the activities may be in a variety of forms. Some activities require short answers or visual representations that can be dichotomously graded. Other activities require more detailed answers that will be graded using a scoring rubric (with 0 to 3 points awarded). Based on the limited number of tasks that I have seen it appears that most of the detailed answers are structured in a way that allows scorers to utilize analytic, as opposed to holistic, scoring rubrics. From the situative perspective it is very important that the work products captured in some way mimic the real world of science. Unlike multiple-choice answers, which are clearly an artifact of the educational system, using a variety

of response techniques does seem more realistic. However, one does need to question why computers are not being used in this endeavor to add another layer of authenticity.

The Link between Performance Tasks and Instruction

Now that the student, evidence and task model have been applied to the MSPAP it is appropriate to use this framework as a means of evaluating the MSPAP. Are the claims being made by the state of Maryland backed up by the data that has been gathered on the MSPAP? Are there other plausible explanations for the data? Using Toulmin diagrams, such as the one shown below, this paper will now attempt to evaluate the claims made by the state of Maryland and determine whether or not MSPAP is meeting its educational goals.

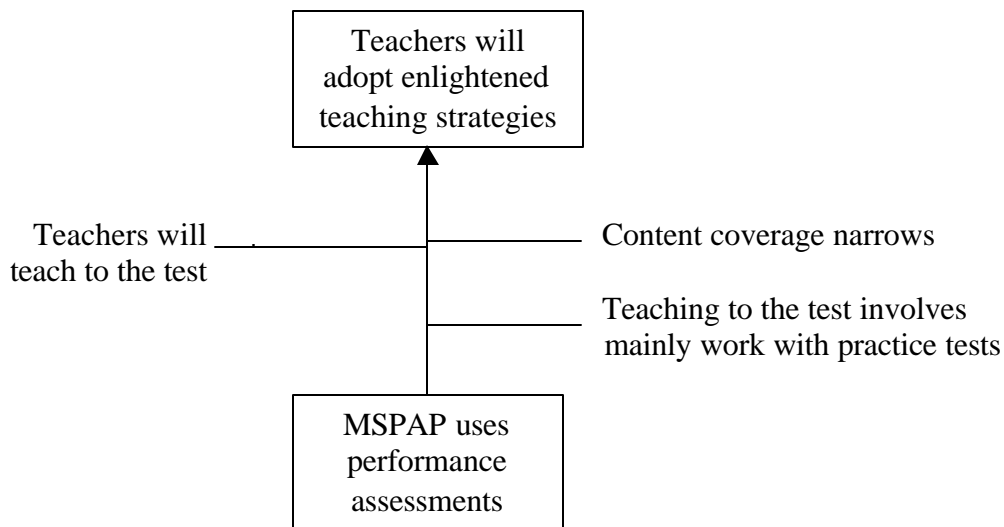


Figure 1 – Toulmin diagram linking MSPAP tasks and one intended outcome.

Over the last decade policy makers have decided that performance assessments should be embraced as a tool for educational reform. Their thinking is summed up well by Haertel (1999) in his discussion of performance assessments. He says the policy maker’s argument is that “if multiple-choice tests could drive instruction in the wrong direction, then why not use some better

sort of test to drive instruction in the right direction?” Following this line of thinking one can reason that using high-stakes tests that call for high-order thinking and engage students in problem solving will drive instruction that will be more in line with these goals. The state of Maryland certainly believes that this is the case. Their web-site makes the following statement; “Teaching to the test is good instruction, the kind of instruction that results in understanding, not in the mere rote recall of isolated facts.” The question is whether or not this goal of enlightened instruction is being met due to the use of the MSPAP.

In order to back up the claim that teachers are using enlightened teaching strategies in response to the MSPAP, an obvious first step is to ask the teachers if this is the case. Science teachers in Maryland were asked to respond to questionnaires regarding how their instruction has changed since 1992 (Lane, et.al, 1999). When asked to indicate the extent to which MSPAP influenced them to make positive changes in their instruction, 76% of science teachers said that MSPAP had a moderate or great amount of influence. Over 70% of the science teachers also indicated that their emphasis on the science learning outcomes had increased since 1992. Principals also indicate that the MSPAP is having an impact on classroom instruction.

Researchers from CRESST/RAND (Koretz, 1996) also reported large majorities of teachers and principals said they had been successful in meeting MSPAP’s goal of improved instruction. However, other findings of theirs seem to contradict this to some extent. They note that as a consequence of MSPAP some teachers had narrowed their instruction to focus on the test. They also reported the use of practice tests to help students gain familiarity with the assessment format. Clearly using practice tests in that way is not a problem, but about half of the teachers responded that this increased familiarity was responsible for a significant portion of the gains made by their school. Only 15 to 20% of the teachers made the same sort of statement

with regard to improvements in knowledge and skills. Anecdotal evidence from some Maryland teachers also indicates that teachers are, in some instances, training students how to get partial credit on some questions by meeting the general requirements set forth in the scoring rubrics without actually knowing the subject matter. The question is then, are the teachers offering better instruction, as state of Maryland had hoped, or simply better preparation for the MSPAP?

Looking at the general trends in student achievement on the MSPAP science sub-test may offer some evidence in this regard. Certainly some of the jump from 39.7% of the students state wide performing at the satisfactory level in 1994 (the first year for the science exam) to 46.1% the following year must be attributed to an increased familiarity with the exam and its format. Increases since then have been smaller but the trend it still positive. It seems unlikely to me that these increases could be due solely to preparation for the test. If that were the case, I believe

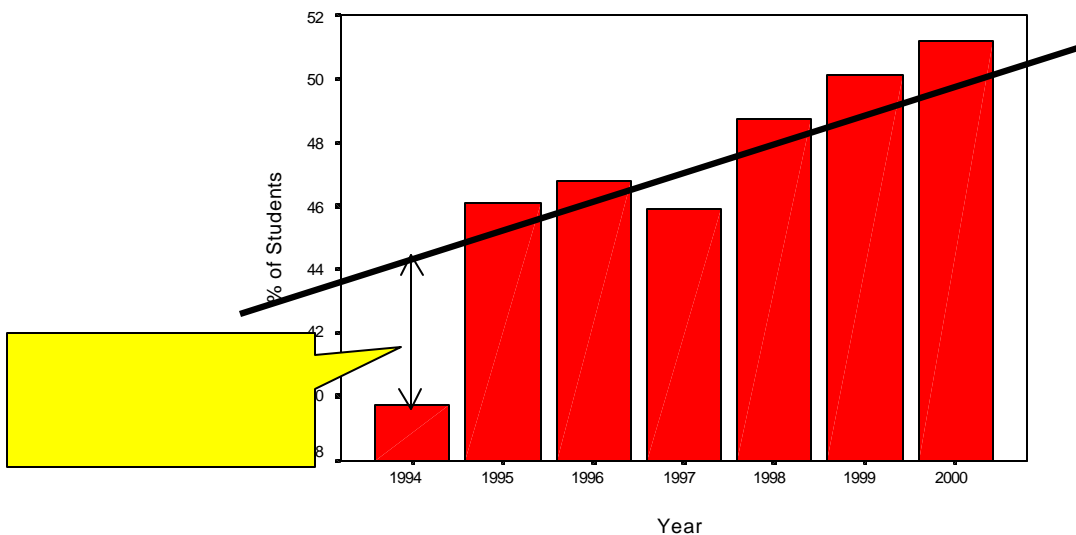


Figure 2: Percent of 8th grade students at “Satisfactory” Level on Science portion of MSPAP (MSPAP, 2000)

that a ceiling effect would be encountered. After all, there are only so many test-taking strategies that a student can be shown. After those have been exhausted the only way to increase scores is

to actually teach the students the subject matter. If the scores indicate an increased level of mastery it is plausible to infer that better teaching strategies have lead to these increases.

Based on the information provided above regarding teachers' sentiments and the increases in scores, it appears that MSPAP is changing instruction in the state of Maryland. There does appear to be sufficient evidence to support the notion that teachers are using more problem-solving methods in their classrooms. Of course, not all teachers are preparing students for the test in that manner. Others use more behaviorist approaches, which will be addressed in the next section. Still, we can say that many teachers are teaching their students to analyze what they read and apply their knowledge to solve problems. In the final analysis, it appears that teaching to the test has, in fact, been good instruction just as the state promised when they started the Maryland School Performance Assessment Program.

The Link Between Student and Task Models

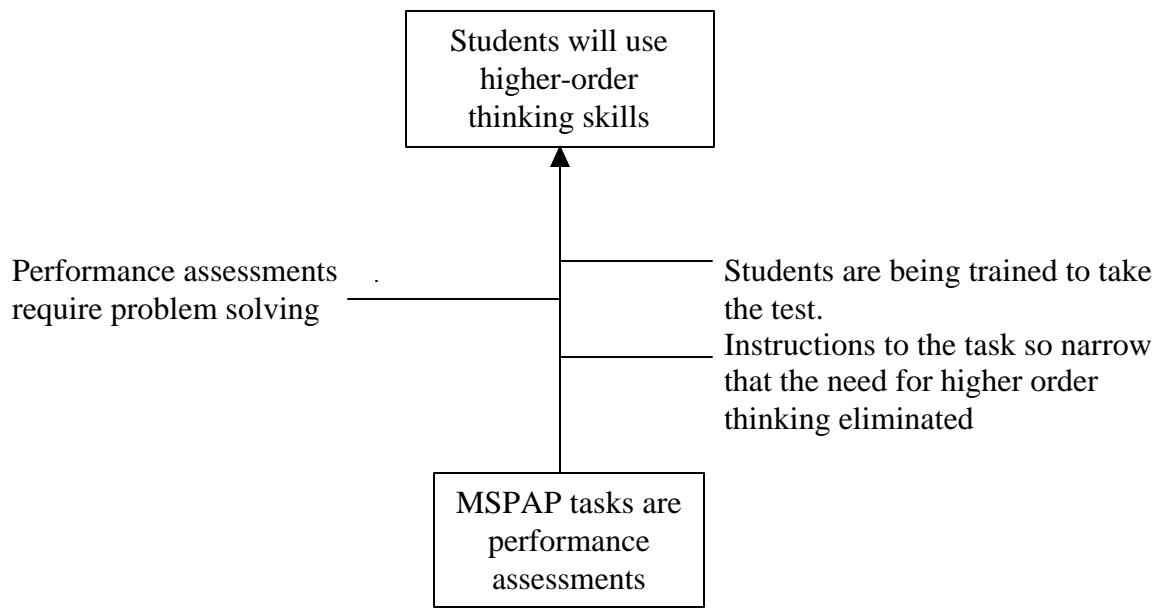


Figure 3 – Toulmin diagram linking MSPAP tasks and student outcomes.

According to Wiley and Haertel (1996), “testing is an activity that is intended to reveal skills, conceptions, abilities, or knowledge of the individuals being tested”. In science, performance assessments are sometimes chosen because they are situationally appropriate. That is, these tasks mirror what scientists actually think and do. Students are expected to approach these tasks as scientists approach real problems, by using their knowledge to solve problems. In doing so the expectation is that they will utilize higher-order thinking skills rather than the mere recall of information. That is the theory behind the use of performance assessments on the MSPAP. The question is whether or not these tasks actually reveal what they intend to reveal.

It is possible that it is the quality of the performance assessment that determines whether or not it taps into anything more than simple procedural knowledge. Glaser and Baxter (2000) state that “the realization or activation of the components of cognitive performance stems in part from the content and process demands of the tasks involved”. They describe “process constrained” tasks that do not provide students with opportunities to plan, select or implement

appropriate strategies. An example of this would be assessments that provide step-by-step instructions that actually prohibit students from being able to demonstrate problem solving. The use of structured tasks is the result of a trade-off that must take place in a large-scale assessment like the MSPAP. Providing step-by-step instructions for tasks allow more students to interact meaningfully with each task and allows scoring rubrics to be more specific. The pitfall of doing this is that the evidence about problem-solving ability is weaker when the structured tasks are used. Clearly any student that cannot solve the structured task will not be able to do so if the task is unstructured. This same sort of generalization cannot be made for students who solved the structured task correctly.

If, in fact, performance assessments are sometimes “process constrained”, then the concept of construct underrepresentation and its impact on validity must be considered. Messick (1994) states that “the level and sources of task complexity should match those of the construct being measured and be attuned to the level of developing expertise of the students assessed”. So, if problem-solving ability is what MSPAP seeks to measure about students, then the tasks used must be sufficiently complex to evoke this. If not, construct underrepresentation will result and the use of MSPAP test scores as a measure of problem-solving ability would not be valid.

One other issue with regard to the link between the student and task model deals with teachers training students to take the test. Whenever any high-stakes assessments are administered the possibility exists that teachers and administrators use inappropriate strategies to increase performance. Since the MSPAP is a test that stresses thinking, it must be considered inappropriate to offer students strategies allow them to get points without utilizing the problem solving skills intended. Anecdotal evidence obtained by talking to teachers in Maryland reveals that they are preparing students for the MSPAP by helping them to recognize clues on the test

and respond properly. For example, one of the tasks on the English portion of the assessment is to write to persuade. Students are trained that as soon as they see that the task involves that they should begin their paragraphs with the words ‘I am writing to persuade you....’ because this automatically gets them one of the three or four points assigned to the task. In science, it is conceivable that a teacher could use a very behaviorist teaching strategy and tell his or her students to refer to a chart or a graph within an explanation for a “guaranteed point”. Certainly charts and graphs contain data that should be used in formulating a complete response to a question. The problem is that if some students do not understand why they are using something then the inferences that can be drawn are weaker. Did the student get a “2 out of 4” because he had some problem solving ability or did he get that score because he was well trained on how to get points? These conditioned responses are classically behaviorist in nature. A student does something and gets a reward for it. In this case the reward is a point on the test. In the larger sense schools are rewarded with high rankings on the MSPAP if the teachers get the children to respond properly.

The evidence proving that the use of performance assessments causes students to use higher-order thinking skills is not very compelling. The tasks certainly have face validity in this regard. However, when one looks deeper it becomes apparent that the design choices made for the MSPAP have compromised the test in terms of its ability to illicit high-order thinking or measure problem solving ability. Given the difficulty of writing and grading performance tasks and the time spent having the students take the tests, it is ironic that not much more evidence is gained over using a multiple-choice format.

The Link Between Student and Evidence Models

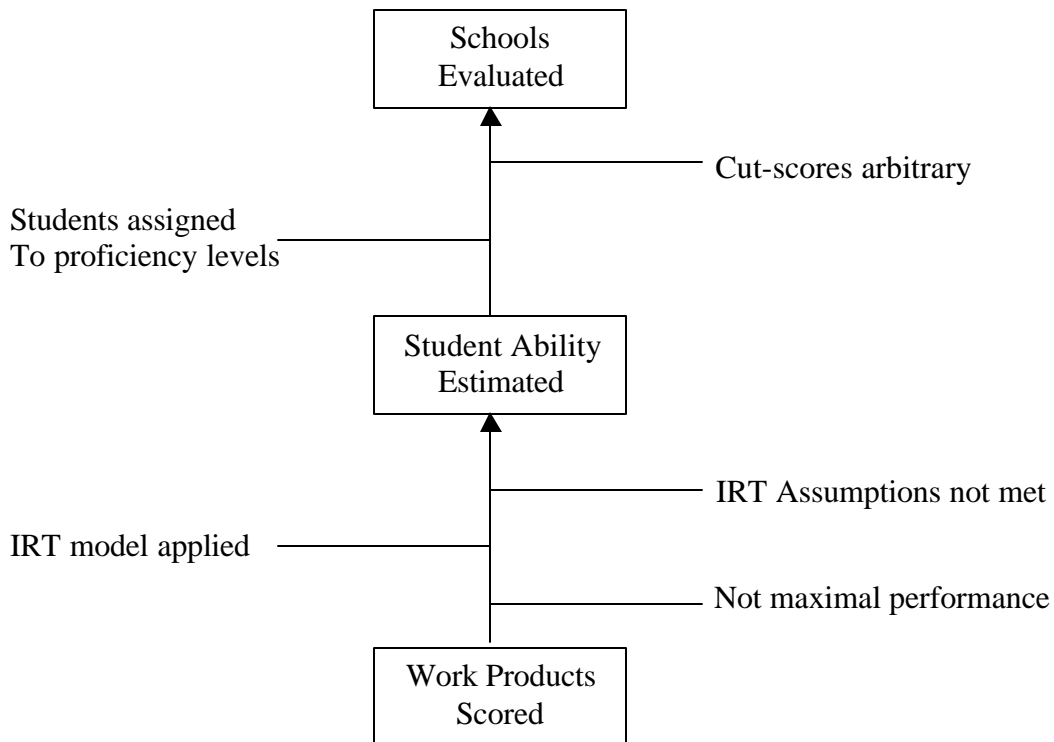


Figure 3: Toulmin diagram linking evidence and student models to the evaluation of schools.

Three important assumptions underlie the use of item response theory. They are the correct specification of the dimensionality, local independence and respondent independence (Hambleton & Swaminathan, 1985). In the case of the MSPAP one has to wonder whether any of these assumptions are met. Using a unidimensional model, as MSPAP does, implies that a single latent trait underlies the response process. Given the nature of the science tasks this may or may not be an accurate assumption. Many of the MSPAP tasks involve reading an introductory paragraph or section and then answering short-answer questions based on this material. Given this, one has to question whether or not reading and writing ability are secondary traits. If these secondary traits do exist, means for single examinees could be

different. In the case of reading ability it would seem that the means would be underestimated since students who did not read well would do poorly on items. This would mean that the overall means for schools would also be underestimated. In the case of writing ability the overall means may not be affected, because students who write poorly may lose points but students who write particularly well may be able to “finesse” their way to a higher score.

The concept of local independence assumes that an examinee’s responses to different items on a test are statistically independent. That is, an examinee’s performance on one item cannot affect his or her performance on subsequent items. MSPAP states that “there have been no testlets of dependent items constructed since 1992” (MSPAP, 1998). Still, given that several studies of MSPAP have utilized a dependency statistic, analogous to Yen’s Q3, this issue is clearly on peoples’ minds. Inspection of public release items seems to indicate that local dependence is, in fact, still a problem. An example of items that seems to demonstrate local dependence are questions on the exam that ask students to explain their thinking regarding a previous question. If these items are serially dependent in that one needs to answer question 5 correctly to have a chance on question 6, then means will once again be underestimated for individual students and schools.

IRT models also generally assume that the responses of individuals are independent of one another. Violations of this assumption could occur in cases of cheating or when testing is occurring in a group setting. Recent published reports of suspected cheating on the MSPAP offer some evidence that respondent independence has been violated for that reason.

Examination of the tasks also provides an indication that some of the tasks are administered to groups of students. Though these group tasks are not graded, the tasks based on the work of the group are. In addition, there appears to be nothing to preclude “more able” students from

explaining the tasks in depth to the “less able” students and perhaps preparing them for potential upcoming questions. Though exam administrators are instructed to randomly assign students to groups of four when they enter the room (MSPAP, 1997), this random assignment is assumed but not ensured. Given the high-stakes associated with this assessment, it is possible that teachers assign students to heterogeneous groups to ensure that the “less able” students get some help during the test. If respondent dependence due to cheating or “helping” did occur, means would be overestimated.

Besides the issue of the appropriateness of the IRT assumptions, there is a problem regarding student performance on the MSPAP. How should this performance be categorized? Is it typical of what the student will do in most situations? Is it maximal in that students put forth their best effort on the exam? Or is their performance sub-par for some reason? On most standardized tests that students take an individual score will result. Therefore students in these situations are invested in their own level of achievement on the test. While teachers and administrators are invested in student achievement on the MSPAP, how students view these tests is much more uncertain. Some students will probably try to do their best, while others will put in a more typical effort. It is possible that some students will even seek to underachieve in order to punish their school in some way. How these separate groups of students affect the school means is up for debate. If we assume that the MSPAP is attempting to measure typical achievement then school means would be relatively unaffected based on the scenario offered above. However, if MSPAP seeks to measure maximal performance, then the school means will be underestimated.

One final issue that must be examined is with regard to the cut scores for the proficiency levels on the MSPAP. The question of the legitimacy of the cut-scores is one that many

standardized tests face because of the seemingly subjective manner in which the cut-scores are established. The MSPAP arrived at their cut-scores by looking at where items fell on the scale score continuum. Since many items were located near 490, 530, 580 and 620, these values were used as the cut-scores to establish the proficiency levels. This method, which seems analogous to the distribution gap method of scoring, is objective and it seems reasonable. What seems less reasonable is the arbitrary choice of establishing 70% of students at the satisfactory level as the satisfactory level for schools. Figure 2 showed that state wide less than 52% of 8th grade students are performing at the satisfactory level on the science portion of the MSPAP.

Examination of the data for the 24 school systems in the state of Maryland shows how rare it is for an individual school district to meet that level. In the 9 year history of the MSPAP only two school systems have ever reached the “holy grail” on the science portion of the test, Garrett County with 71.3% satisfactory in 1995 and Calvert County with 71.5% satisfactory in 1998. Since MSPAP has not provided evidence of the soundness of their criterion, we are left to question how reasonable it is. Are schools still under-performing in terms of science or is the criterion simply too high. Examination of the 1996 NAEP state data for science yields the following results:

	% Below Basic	% at Basic Level	% at Proficient	% at Advanced
US Average	40	33	24	3
Maryland	45	30	23	2

Figure 4: National averages vs. Maryland averages on NAEP-science in 1996

This figure shows Maryland to be right around the national averages on the science portion of the NAEP. Of course, the NAEP is not the MSPAP, but this does seem to indicate that there is room for Maryland to do better in terms of science. Had Maryland performed in the top 10% of the

nation, one would have to question the worth of aspiring to the seemingly unreachable 70% satisfactory mark. But since the state's comparative performance is mediocre at best there seems to be reason to set a high standard.

The question that remains is in regard to the link between what MSPAP gets from students and what MSPAP says about them. Given the apparent violations of the IRT model it would seem that collecting accurate data about individual students is impossible. But is this data collected with sufficient accuracy and breadth to say something about the schools? With a few exceptions it seems that most of the violations to the IRT model result in underestimation of the means. If the means are underestimated then schools will have a more difficult time reaching the 70% satisfactory level. That will be a problem when and if two things occur. First, Maryland students need to perform better on national tests before people get overly concerned about the cut-scores. Scoring at the mean should not be used as proof that Maryland students know science. Second, student and school achievement needs to reach a ceiling. As it stands now, the number of students at the satisfactory level is still in an up trend. This is true for the state as a whole and for most individual districts. If the percentages were to cease rising and stay at a level significantly under the 70% satisfactory level, then the issue of the arbitrary setting of the cut-scores would become important. Until those two things happen it seems reasonable for the state of Maryland to continue to set high standards, arbitrary though they may be, for students and schools to meet.

CONCLUSION

I went into this project with many negative preconceptions of the MSPAP based solely on conversations with teachers and students in this state. My ideas about the program have changed

largely because I have begun to appreciate what the goals of the program are and the mechanisms that have been put in place to meet these goals. The Maryland School Performance Assessment Program is very well conceived. The creators of the test have spent a tremendous amount of time and effort looking at the state of the art in terms of cognition and performance assessments, and they have applied these ideas to the MSPAP. This test meets many of the goals that the state has set including changes in instruction and improvement in students' skills.

Of course, the test is not perfect by any means. The evidence linking the student and tasks models is weak. Though the tasks are supposed to elicit problem-solving skills from the students, the evidence that they do so is limited. This is partly due to the constraints made necessary by the large scale of this assessment. Tests given to every student in a state must make certain concessions so that all students can interact meaningfully with the tasks and so that these tasks can be graded efficiently and fairly. Unfortunately, these compromises undermine the ability of the test to set up situations where students utilize problem-solving skills.

Overall, however, it seems to me that in terms of cognition and psychometric analysis MSPAP is a good program. It should be noted that there are political issues involving the allocation of funds based on the MSPAP results that have not be examined in this paper. These issues come to the fore when people in the state of Maryland complain about this testing program. Perhaps if people could set those issues aside and see MSPAP as others outside the state do, they would come to appreciate the value of this program.

REFERENCES

Glaser, R.G. & Baxter, G.P., 1997. Assessing Active Knowledge. CSE Technical Report 516. Center for Research on Evaluation, Standards, and Student Testing.

- Haertel, E., 1999. Performance Assessment and Education Reform. *Phi Delta Kappan*, May 99, Vol.80 Issue 9, pg 662.
- Hambleton, R.K. & Swaminathan, H., 1985. *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff Publishing.
- Koretz, D., Mitchell, K., Barron, S. & Keith, S., 1996. Final Report: Perceived Effects of the Maryland School Performance Assessment Program. From the CRESST web-site: www.cse.ucla.edu/CRESST/Reports/TECH409.pdf.
- Lane, S., Ventrice, J., Cerrillo, C. & Stone, C., 1999. Impact of the Maryland School Performance Assessment Program (MSPAP): Evidence from the Principal, Teacher and Student Questionnaires (Reading, Writing and Science). Presented at the annual meeting of the National Council of Measurement in Education, in Montreal, April 1999.
- Messick, S., 1989. Validity. In R.L.Linn (Ed.) *Educational Measurement* (3rd ed., 13-103). New York: McMillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
- Mislevy, R.J., Steinberg, L.S. & Almond, R.G., 1999. Evidence-Centered Assessment Design.
- MSPAP, 1997. Public Release Task: Planetary Patterns, Grade 8. Maryland State Department of Education, July 1994.
- MSPAP, 1998. 1998 Technical Report – How is MSPAP scored? From the web site www.mdk12.org/mspp/mspap/how-scored/98tech_report/scaling_equating.html
- MSPAP, 2000. From the MSPAP web-site: www.mdk12.org/data/worksheets/allgraph1.asp.
- MSPAP, 2001. Information from the web-site: www.mdk12.org/mspp/mspap
- Roberts, J., 2001. EDMS 724 Class notes, Spring 2001. University of Maryland.
- Wiley, D.E. & Haertel, E.H., 1996. Extended Assessment Tasks: Purposes, Definitions, Scoring, and Accuracy. In M.B. Kane & R. Mitchell (Eds.), *Implementing performance assessments: Promises, problems, and challenges*. Mahwah, NJ: Erlbaum.
- Yen, W. & Ferrara, S., 1997. The Maryland School Performance Assessment Program: Performance Assessment with Psychometric Quality Suitable for High Stakes Usage. *Educational & Psychological Measurement*, Feb 1997 Issue 1, page 60.