

**THE MARYLAND SCHOOL PERFORMANCE ASSESSMENT PROGRAM (MSPAP) 1991-2002:
POLITICAL CONSIDERATIONS**

Paper to support a presentation in the National Research Council Workshop *Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards*, December 10-11, 2009

Steve Ferrara
CTB/McGraw-Hill
January 24, 2010

THE MARYLAND SCHOOL PERFORMANCE ASSESSMENT PROGRAM (MSPAP) 1991-2002:

POLITICAL CONSIDERATIONS

OVERVIEW AND INTRODUCTION

The Maryland School Performance Assessment Program (MSPAP) grew out of evolving beliefs in Maryland about the policy role and influence of educational testing on curriculum, teaching, student learning, and school quality. Those beliefs arose with the school accountability movement of the 1970s. The initial policy goal was to hold schools publicly accountable for their responsibilities to children and for their use of public tax dollars. In the 1980s, high school students and their teachers were required to ensure that high school graduates had achieved basic academic knowledge and skills. In the 1990s, schools were held accountable for improving the quality of public education by demonstrating that students were progressing toward meeting newly rigorous standards for achievement and test performance. Test designs in each decade corresponded to the educational beliefs of those decades.

In this paper, I describe MSPAP, the role MSPAP played in educational reform in Maryland, and the role of politics in the demise of MSPAP. Each section of the paper addresses issues requested by the workshop Steering Committee. I rely on my recollections and interpretations of what I saw, heard, and experienced directly as a designer of MSPAP and the Maryland State Assessment Director from 1992 to 1997. I have documented my observations and assertions as much as possible. I have checked facts and tested my observations with others as much as possible. Some details may be slightly inaccurate, but most details and conclusions are essentially correct and endorsable by other observers.

HISTORY AND BACKGROUND

More than a decade of public reporting on school performance—that is, reporting student achievement test scores—and using state tests to drive changes in instructional content and approaches set the stage for MSPAP and its role in school reform. Accountability testing, followed by minimum competency tests as high school graduation requirements, set the stage for relatively calm adoption of MSPAP. Michaels and Ferrara (1999) and Chamblin and Herndon (n.d.) trace the role of testing prior to MSPAP. I reiterate that history below.

School and School System Accountability: 1975-1989

Beginning in 1975 or thereabouts, the Maryland State Department of Education (MSDE) required each school system to administer the same nationally normed test to all students in all schools in the state. We know now that “all students” was a loosely defined term that did not include students with disabilities, (small numbers of) English language learners, and other students that schools may have exempted from the state requirement. The testing requirement included reading, language, and mathematics subtests in grades 3, 5, and 8. Each year, MSDE published average grade equivalent scores in these content areas and grades for each school, all 24 school systems, and the state. No consequences beyond public disclosure were attached to the accountability reports. Low poverty systems like Montgomery County posted grade equivalent means at and above the required grade levels. High poverty systems like Baltimore City, Prince George’s County, and St. Mary’s County posted below grade means. Over time, all means increased, at different rates, but rank ordering did not change much.

Project Basic: 1977-2001

In 1977, State Superintendent David Hornbeck launched Project Basic, which included basic skills instructional objectives for high school students in reading, language arts, and mathematics. The program also included minimum competency tests that targeted those objectives. Multiple choice

tests in reading, mathematics, and citizenship skills and an essay test were implemented statewide between 1979 and 1983. These tests were intended as additional high school graduation requirements. After some implementation errors and much resistance, the three multiple choice tests were included as graduation requirements beginning in the early 1980s. Grade 9 pass rates on the reading test exceeded 90 percent as early as 1985. MSDE conducted an unpublished study in 1987 of the “compound failure rate” (i.e., percentages of 12th graders who failed one, two, three, or all of the functional tests) that indicated that fewer than 500 of the students who had not dropped out of high school by grade 12 (approximately 60,000 students) would have failed to receive a diploma, if all four tests had been diploma requirements that year. (The study did not adjust the results for students who failed to meet course credits and attendance requirements.)

The Maryland Writing Test was scheduled as a diploma requirement for the graduating class of 1986. However, grade 9 pass rates were low (e.g., 54.4% in 1985) and local school system resistance was high. In response to pressure from local school systems and legislators, MSDE hired ETS to evaluate the writing test and recommend revisions to the test and scoring procedures. MSDE implemented recommendations for tightening scoring benchmark selection and related procedures. Students in ninth grade in 1988-1989 (the graduating class of 1992) were required to pass the Maryland Writing to receive a high school diploma. The grade 9 pass rate in 1990 was 88.2 percent. Within three or so years, grade 9 pass rates exceeded 90 percent, and many local school systems implemented a new option to administer the writing test in grades 7 and 8. Numerous informal reports from around the state documented that school systems and schools had initiated fundamental changes in writing instruction in the middle 1980s. Changes included teaching the composing process in grade 9 (i.e., planning, outlining, drafting, revising), rather than sentence construction exercises from *Warriner’s English Grammar and Composition: First Course* (see http://openlibrary.org/b/OL7364035M/Warriner's_English_Grammar_and_Composition), initiating writing instruction in middle schools, and adapting the composing processing as early as kindergarten. (I observed this process in elementary schools in Howard County and elsewhere, where Maryland Writing Test performance increased the most dramatically.)

Educational Reform and the Maryland School Performance Assessment Program:

1991-2002

In 1987, newly elected Governor William Donald Schaefer appointed the Governor's Commission on School Performance. The final report, published in 1989, recommended sweeping changes in state content standards, statewide assessment, and school accountability. The recommendations were intended to improve the effectiveness of schools in preparing students for the emerging information based economy and international economic competition. Maryland's State Superintendent later noted that this was the first time that a call for school reform and improvement came from outside of the educational policy community (Grasmick, 1997). Although Maryland was a leader among states in school accountability and education reform, the vision was shared nationally. In 1989, President George H. W. Bush met with the nation's governors and set six educational goals for the U.S. (Chamblin & Herndon, n.d.) and which President Bill Clinton incorporated in the Goals 2000 act in 1994. MSPAP and options for turning around persistently low performing schools (called reconstitution) were hallmarks of the educational reform effort, the Maryland School Performance Program.

The educational policy and political context in Maryland enabled these significant changes in aspirations for school improvement. Governor Schaefer, who captured 82 percent of the vote in his first 1986 gubernatorial campaign, had emphasized education as a priority. (Schaefer was governor from 1987 to 1995.) Schaefer had completed highly successful urban renewal projects as Mayor of Baltimore City (1971-1987), and led development of Baltimore's popular Inner Harbor as a center of commerce and tourism. Among his circle of Baltimore City reformers were Walter Sondheim, who chaired the Governor's Commission on School Reform and later was President of the State Board of Education; and Robert Embry, President of the Abell Foundation in Baltimore, who was President of the State Board of Education that adopted the Maryland School Performance Program and MSPAP. It is not insignificant that Maryland's governor appoints State Board members and that the Board appoints the State Superintendent. The current State Superintendent, Dr. Nancy Grasmick, was first appointed State Superintendent in September 1991, immediately after the first administration of

MSPAP. Grasmick was well known and highly respected in educational and political circles prior to her appointment.

The premises for the Governor's Commission recommendations echoed the theme for education reform in the 1990s: all students can learn, should be able to attend effective schools, and have access to rigorous academic content. In response to the recommendations, the State Board of Education adopted new content standards, the Maryland Learning Outcomes, in 1990. These academic content standards aspired to so-called higher order thinking skills and conceptual understanding. In addition, the board required annual school report cards and reconstitution of schools that persistently failed to improve toward rigorous school performance goals in academics, drop-out rates, and other areas. MSDE assessment and curriculum staff designed the Maryland School Performance Assessment Program (MSPAP, or "mizpap") to align with and target the learning outcomes.

MSPAP DESIGN AND CONTENT AREAS

The Maryland Learning Outcomes and MSPAP address six content areas in grades 3, 5, and 8: reading, writing, language usage, mathematics, science, social studies. MSPAP evolved from the role of the Maryland Functional Testing Program's role in guiding and goading instruction and, perhaps especially, the Maryland Writing Test's essay prompts and rubric scoring. However, MSPAP represented a significant break from traditional achievement testing for Maryland that occurred simultaneously with innovations in California, Kentucky, Vermont, and a small number of other states.

Design Innovations

MSPAP's innovations seem rather mundane today. In 1991, they represented advances, calculated risks, and considerable financial investment.

- All MSPAP items required examinees to construct responses rather than select responses. All examinee responses were scored by trained raters using generic or item-specific rubrics.

With at least eight responses per content area, six content areas, three grades, and approximately 60,000 examinees per grade, more than 9 million score decisions were required after each administration.

- All items were contained in coherent assessment tasks that were organized around a theme (e.g., acid rain) and purpose (i.e., conduct an experiment using acids and chalk and other to examine the effects of acid rain).
- Most tasks integrated several content areas. For example, the eight language measures in each test form were captured by scoring responses to prompts in reading, writing, mathematics, science, and social studies. In other cases, a single scoring decision was counted twice, for example, once in reading and once in social studies. “Mega-tasks” integrated items for as many as several content areas. For example, the task “Archimedes,” which requires examinees to investigate properties of buoyancy, contains items that assess reading (a passage Archimedes’ discovery of the principle of buoyancy), writing (about a logo for an Archimedes t-shirt and a speech about his contributions to science), language usage, mathematics (estimating the constant π), and science.
- Content standards were matrix sampled across three forms so that a larger number of outcomes in each content area could be assessed across the three forms.
- No individual scores were reported. Individual student scores were aggregated to report the percentages of students in each school, system, and the state that reached each of the MSPAP performance levels: Levels 1 (highest) to 5 (lowest).
- The Maryland Learning Outcomes did not specify facts (e.g., specific authors, arithmetic facts, historical dates) to identify what students should learn in each content area. The outcomes described concepts (e.g., inquiry as a habit of mind), content area skills (e.g., inferring an author’s intent, estimating in mathematics, hypothesizing and predicting in science), and broader skills and processes (e.g., interpretation, persuasion) that students were expected to develop and be able to apply as a result of studying each academic content area.

- All reading passages, graphs, charts, and other illustrations in all content areas were taken from real, published materials. Most mathematics and science tasks required skilled use of typical manipulatives (e.g., rulers, calculators) and those typically found only in the best equipped classrooms (e.g., microscopes, ice, lard, cocoons). Students worked in groups to conduct science experiments, discuss social studies issues, and to provide comments on draft essays; students responded to items independently after completing group activities.
- Responses in each content area were scale using the two-parameter partial credit model (see Yen & Ferrara, 1997). Overlapping test forms in each content area were calibrated to a single reporting scale based on randomly equivalent groups using Stocking-Lord equating.
- Cut scores to distinguish the five MSPAP proficiency levels were set using an item mapping approach that was a precursor for both the Bookmark and Item-Descriptor (ID) Matching standard setting methods (see chapters 10 and 11 in Cizek & Bunch, 2007). In addition, standards were established to identify schools that were performing at the Satisfactory and Excellent levels, and each school's progress toward reaching the Satisfactory was reported each year.

MOTIVATIONS FOR MSPAP

In some ways, MSPAP represented a revolutionary break from assessment and school improvement traditions in Maryland education. In other ways, MSPAP represented a natural evolutionary advance in those traditions. Grasmick (1997) observed that Maryland confronted several issues related to implementing school reform in the 1980s. Confronting these issues set the stage for the Maryland School Performance Program and MSPAP (Michaels & Ferrara, 1999).

- Educators and the business community came to believe that students need to learn to be information consumers and users, and that facts and figures are means to solving problems, not ends themselves.

- Public expectations for student achievement and school performance increased significantly and education became a high priority on public agendas.
- Maryland's state education funding had grown to 20% of the state budget; only health, including Medicare funding, was higher (22%). In addition, 50% of local education funding came from local budgets. The state legislature and State Board of Education expresses the need to see returns on these investments.
- The diversity of student population increased, the demands of special education services increased significantly, and school enrollments increased.
- Education became increasingly politicized nationally, beginning with state block grant funding in the Reagan administration, and continuing in the first Bush administration. In addition, several other states implemented far reaching and high profile education reform programs (e.g. California, Kentucky).

ADOPTION: OBSTACLES AND SOLUTIONS

After publication of the report of the Governor's Commission on School Performance in 1989, adoption of the Maryland Learning Outcomes, MSPAP, and even the broader Maryland School Performance Program by the State Board of Education was relatively smooth. The Governor's Commission recommendations were supported by the Governor Schaefer, State Superintendent Grasmick, and the State Board President, Robert Embry. And the design and innovations seemed to be a natural next step—or leap—after almost eight years of influence of the Maryland Writing Test and the higher order thinking skills movement of the late 1980s in Maryland and around the country. Schaefer had emphasized public education in his campaign and was highly popular and influential. (The damage to his popularity from, for example, his “outhouse” reference to Maryland's Eastern Shore and endorsement of George H. W. Bush for re-election, came later.)

Despite misgivings, local school system leaders showed some support for the Maryland School Performance Program. The long-standing Bloom Committee, made up of influential

representatives from some of the 24 local school systems, and which had been advising the State Department of Education on local concerns resulting from Project Basic and the Maryland Functional Testing Program, appeared to view it as a natural next step in school reform. Department leadership met with their local school system counterparts—local Superintendents, Assistant Superintendent for Curriculum and Instruction, Accountability Coordinators—almost monthly. Soon after, department staff met regularly with local special education leaders on providing test administration accommodations for students with disabilities and counting them in school report cards. In addition to misgivings, there appeared to be a fair amount of skepticism that MSPAP could survive long and a wait and see attitude about embracing school reform. (Although no one said it out loud, some people appeared to be waiting for inevitable collapse.) Staff of the Montgomery County assessment and research office opposed MSPAP vigorously, as they had opposed the Functional Testing Program. Other local school systems (e.g., Frederick, Harford, Howard) appeared to embrace MSPAP and school reform, capitalized on the state’s drive, led the way targeting curriculum, instruction, assessment, and professional development to the Maryland Learning Outcomes and MSPAP. They also were among the highest performing school systems on MSPAP and other school report card indicators.

The recommendations of the Governor’s Commission and the assessment vision that MSPAP represented were funded by the state legislature because of strong support by the governor and key legislative leaders. The State Board adopted the recommendations, the Maryland Learning Outcomes, and MSPAP in part because of the influence and connections among the governor, State Board President, and the State Superintendent.

DEVELOPMENT: OBSTACLES AND SOLUTIONS

Design and development of MSPAP—as an assessment system and as individual content area assessments—presented significant intellectual, operational, and practical obstacles. Those obstacles also represented opportunities to develop solutions and force institutional growth in MSDE.

Conceptualizing MSPAP, Assessment Tasks, and Items

The step from the eight years of operating the Maryland Writing Test—two essay prompts scored on 4-point scales—to all constructed responses items in reading, writing, language usage, mathematics, science, and social studies was a conceptual leap. The idea of several constructed response items organized coherently around a theme or purpose was not new in assessment, but was a conceptual breakthrough in high stakes state testing. Each item in a task had to be aligned to one or more of the Maryland Learning Outcomes. MSDE staff had little experience with developing items that appeared to align with these outcomes. Both the Maryland Learning Outcomes and the assessment tasks were intended to reflect the aspirations to the higher-order thinking skills (i.e., HOTS) movements and counteract the narrowing and “dumbing down” of curriculum and instruction (e.g., Michaels & Ferrara, 1999, p. 105). One of the most daunting obstacles was time. The contract kick-off meeting with CTB/McGraw-Hill (CTB) was in June 1990. The first administration of MSPAP was May 1991.

MSDE assessment and curriculum staff addressed these obstacles using a number of strategies. First, because the report of the Governor’s Commission on School Performance gave the department authority to conceptualize and develop MSPAP tasks and target the Maryland Learning Outcomes. This authority enabled department managers and intellectual leaders to take design risks like using no multiple choice items and requiring schools to provide manipulatives (e.g., calculators) for test administrations because they were expected to provide them as part of instruction. Staff called on their contacts in other state programs with experience with performance assessments (i.e., constructed response items in content area tests) in Connecticut and elsewhere for design ideas. Further, Maryland staff worked collaboratively with CTB staff and local school system staff to design tasks and scoring rubrics and align items and tasks with the Learning Outcomes. This collaboration with a contractor and local school system staff had far-reaching positive influence on subsequent MSPAP development and operations and on efforts at improving local curriculum, instruction, and teacher professional development.

Designing Tasks that Could be Administered Feasibly, Scored Reliably and Efficiently, and Scale Successfully

Designing tasks that could meet these three crucial requirements seems simple enough today. In the early 1990s, concepts, methods, and procedures to meet them did not exist. Consider the grade 5 science task, *Salinity*, for example. In this task, teachers had to introduce the purpose of the task (i.e., assemble a hydrometer using a drinking straw and clay to measure the salinity levels of two different water samples; decide which water sample was most appropriate for aquariums for different kinds of animal and plant life). Teachers had to follow instructions to create the water samples, assemble task manipulatives (e.g., straws and clay) for students, organize the students into small groups, ensure that students followed instructions for examining the salinity levels of the water samples, and re-assemble students at their desks so that they could respond individually and independently to assessment items about the investigation design, procedures, results, and implications for the aquariums. This sort of activity may have been familiar to elementary school teachers who taught hands-on science lessons, but most teachers in Maryland taught little science of any type. And designing and writing a complicated task like this that could meet the standardization requirements for a psychometrically sound, large scale assessment that was scaled using an IRT model was revolutionary for Maryland.

Further, each item in this task had to meet psychometric requirements so that examinee responses could be scored and item scores could be summed, could support IRT scaling using the two-parameter Partial Credit Model (Yen & Ferrara, 1997), and enable valid interpretations of what students know and can do. Items had to be designed to elicit student responses that were locally independent of one another (i.e., examinee ability to respond to an item does not require that the examinee can respond to any other item, and that independence is statistically demonstrable) even though the items were related to one another through the task's theme or purpose. Further, the items had to elicit examinee responses that could be scored by trained scorers (i.e., Maryland teachers) with adequately high reliability and accuracy and quickly enough to manage the cost of making more than nine million score decisions. Collaboration among MSDE staff, contractor staff (CTB, and in

1993, Measurement Incorporated), and local school system staff was crucial to meeting these requirements. CTB's psychometric expertise—which rivaled the expertise of academic institutions and other testing organizations—was the key to enabling the use of the two-parameter Partial Credit model to scale responses and items, equate test forms within and across years, adjust the equatings for year-to-year rater drift (see Fitzpatrick, Ercikan, Yen, & Ferrara, 1998), and conduct validation studies.

Getting Content Staff on Board and Up to Speed

Staff of the Division of Curriculum and Instruction had limited involvement in the state assessment program, except in providing extensive professional development related to the Maryland Writing Test. These staff had dedicated their time, effort, and professional identities to training teachers and developing curriculum and instruction materials. Staff in some of the content areas were enthusiastic from the beginning about dedicating much of their time to MSPAP test development. Others resisted at first. Once they recognized, and then accepted, the potential of MSPAP's design for influencing daily classroom instruction, content staff accepted the time commitment required to develop MSPAP tasks. Assessment and content staff worked as a team to plan, design, and develop MSPAP assessment tasks each year. The team met weekly from August through March and conducted weekend task development workshops during the winter each year.

Content staff had little time to plan and conduct teacher professional development in the early years of MSPAP because test development was so labor intensive. As the program began to mature and become routinized after the 1993 administration, content and assessment staff were able to devote more time to developing and delivering training teachers, principals, and other local school system staff.

Designing, Developing, and Implementing the First Administration of MSPAP in 11 Months

MSDE and CTB/McGraw-Hill conducted a project kickoff meeting in June 1990 during the CCSSO large-scale assessment conference. The first administration was scheduled for May 1991. Local school system staff, State Department Staff, and contractor staff worked intensively through the summer, fall, and early winter of 1990 to develop assessment tasks for three test forms. State Department assessment staff and contractor staff worked intensively throughout winter and spring 1991 to format tasks for publication in test booklets, to refine test administration directions, and work out the logistics for delivering test materials (including a long list of manipulatives not usually used in statewide assessments) and retrieving them for inventorying and scoring. To ease the burden, only reading, writing, and mathematics tasks and language usage measures were developed for the 1991 administration. Science and social studies were added in the 1992 administration. The intensive work continued throughout summer and fall 1991 and winter 1992, when CTB scored all student responses and conducted scaling and validation studies. The State Department and CTB conducted a standard setting in winter 1992. School performance was reported publicly in late winter 1992. Intensive, year-round work continued in preparation for the 1992 and 1993 administrations, until procedures became more routinized after the 1993 administration.

IMPLEMENTATION: OBSTACLES AND SOLUTIONS

Two primary obstacles made implementing MSPAP challenging: resistance to high stakes school reform represented by the Maryland School Performance Program and MSPAP and teacher, principal, and school preparedness for MSPAP. Prior to MSPAP, the stakes in the Functional Testing Program were applied to students who struggled to meeting graduation requirements. In MSPAP, the stakes—including published school report cards and the threat of intervention and reconstitution in poorly performing schools—were applied to schools and their staffs. MSPAP, by design, became the instructional driver in schools, not just a curriculum guide or a no-teeth accountability measure as in the past. Community opposition arose in conservative enclaves around the states, many spurred on

by the charismatic conservative education activist Peg Luksik¹ and Maryland parents, who reviled what they saw as social engineering of outcomes based learning (see, for example, <http://www.ncrel.org/sdrs/areas/issues/envrnmnt/go/go4outcm.htm>) and the Maryland Learning Outcomes. Local superintendents and school principals also expressed strong misgivings about the fairness and feasibility of achieving Satisfactory and Excellent school performance standards.

On the practical side, teacher preparedness to administer MSPAP tasks and deliver instruction that was represented in the tasks and explicated in the Maryland Learning Outcomes were considerable obstacles. The complexities of organizing students to work in small groups to conduct hands-on investigations and focused small group discussions, and then to respond to items independently, is illustrated in the description of the *Salinity* task. Teachers did not work with their intact classrooms. All students in a grade were randomly assigned to testing groups. Teachers were required to manage all of the complexity with unfamiliar students from other classrooms—and to maintain order, focus, and motivation for nine hours over five days of testing. In addition, many teachers were unfamiliar with the broad expectations for conceptual understanding, reasoning, explanations, and other so-called higher order skills represented by the Maryland Learning Outcomes. Their students were tested on these outcomes. Teachers were expected to teach the outcomes. For the first couple of years of MSPAP, most resources were dedicated to test development and implementation rather than professional development.

The State Department's strategy for counter-acting opposition to MSPAP and the school reform program included involvement of local school system staff, open communication, and maintaining a consistent message. Teachers were directly involved in task development, scoring, of examinee responses standard setting, and a range of other tasks in developing and operating MSPAP. Teachers regularly volunteered that these experiences were the "best professional development" of their careers. The state superintendent and assistant superintendents for instruction and for special education met monthly almost year round with their counterparts to address assessment and reform policy issues. The state assessment director conducted day-long meetings with the local assessment eight times a year to address MSPAP logistical, operational, and policy issues. The state superintendent and other management staff proactively sought meetings with state

legislators who expressed concerns about MSPAP. As a result, these staff often were requested to testify before legislative education subcommittees. MSDE staff also met with local board of education local politicians, PTAs, political action groups, and even individuals who represented opposition movements or who were particularly vociferous. All State Department staff delivered consistent messages about the importance of school improvement for student achievement and success. An effective strategy involved demonstrating an MSPAP task (often the *Salinity* task) and posing the question to irate parents and teachers, “Would you want your child (or student) to be able to do these tasks and answer these questions?”

HOW WELL DID MSPAP WORK? SUCCESSES AND ITS DEMISE

Despite many bumps in the road, MSPAP operated successfully from 1991 through 2002. In this section, I answer the question How well did MSPAP work: psychometrically, as a force for change in schools, and politically.

Psychometric Quality

Literally scores of conference presentations, conference papers, and published papers by MSDE staff and independent researchers document the strong technical quality of MSPAP content area assessments. Yen and Ferrara (1997) summarize MSPAP’s design and highlights its psychometric characteristics with respect to scaling, equating, standard setting, score accuracy, and validity.

MSPAP as a Force for Change in Schools and School Systems

Similarly, considerable numbers of studies demonstrated MSPAP’s effectiveness as a force for change in what teachers taught, how they taught academic content, and their expectations for what students could learn and do. For example, one study of changes in teachers’ and principals’ expectations for student learning and performance indicated that principals encouraged teachers to

raise expectations for students, though teachers tended raise their expectations primarily for higher achieving students (Koretz, Mitchell, Barron, & Keith, 1996, p. ix). A series of studies, supported by a US Department of Education grant to develop and demonstrate methods for investigating so-called consequential validity, indicated MSPAP's influence related to overall school performance gains (Stone & Lane, 2003), gains in reading and writing (Parke, Lane, & Stone, 2006), and gains in mathematics (Lane, Parke, & Stone, 2002). Another study demonstrated local school system and schools' use of MSPAP and other data to guide instructional planning (Michaels & Ferrara, 1999).

Political Viability and Opposition

Opposition to MSPAP often was formidable. Individual state legislators raised challenging questions about the fairness and appropriateness of MSPAP, most often in response to concerns expressed by activist constituents. The questions often raised legitimate concerns—for example, how can scoring of constructed responses be fair and accurate? and How can group science investigations work in a state test? In the absence of real, accurate information, people apparently speculated to create opposition positions and statements. As a result, other questions were based on myths: MSPAP was “social experimentation” and “social engineering” and MSPAP contains a “secret sex survey.”² Conservative and religious right activist groups raised objections to the Maryland Learning Outcomes. They associated the reasoning, problem solving, and other instructional objectives contained there with the broader “outcome based education” movement, advocated by Bill Spady, and which aimed to prepare students for life, not just academically (see, for example, Brandt, 1992). Further, they believed that MSPAP disregarded content knowledge (this objection used the Burger King commercial slogan at the time, “Where’s the beef?”) and traditional basic skills. Getting out real information and combating myths, which often went viral, required addressing irate parents, PTA groups, and teachers in meetings in school cafeterias. The 1992 administration of MSPAP was disastrous in numerous schools. Teachers in these schools did not have many of the manipulatives they needed nor adequate training to manage group activities in the new science and social studies tasks. Media around the state covered the outcry for the most of the two weeks of statewide test

administration. The incident became known as the “Mizpap Mishap.” In 1996 (or thereabouts), parents in dispersed neighborhoods in several school systems held their children out of school for an entire week to boycott the MSPAP administration. Simultaneously, a group of approximately 10 parents and school age children picketed the State Education Building, carrying signs that urged parents to “Call the state Testing Czar” and displayed my name and phone number.³ Educators and parents regularly complained that MSDE would not release individual student scores or publish MSPAP tasks. MSDE’s need to withhold tasks for future administrations raised suspicions about the appropriateness of the content for school age children.

Despite this opposition, mishaps, other errors, and problems that attend newly developed testing programs, MSPAP thrived throughout the 1990s. And growing public support for MSPAP and Maryland’s school improvement efforts were indicated in a considerable number of surveys, many conducted independently of MSDE (see, for example, Maryland State Department of Education, 2000). Sometime around 1999, new, more formidable and well funded sources of opposition grew in prominence and influence.

WHY WAS MSPAP DISCONTINUED?

I believe that MSPAP would have survived further into the 2000s and would have undergone fundamental design revisions to reduce scoring time and costs, reduce administration complexities, and enable faster turnaround of school score reports (just under six months). Before a redesign process could be implemented, three forces defeated MSPAP.

1. Robert Embry, President of the State Board of Education when the Maryland Learning Outcomes and MSPAP were adopted, later repudiated the Maryland Learning Outcomes. Embry has been president of the Abell Foundation in Baltimore since 1987. The Abell Foundation funded external reviews of MSPAP psychometrics and content. The psychometric review of a panel of nationally known psychometricians was favorable (see Hambleton et al., 2000). The separate reviews of the reading, writing and language usage, mathematics, science, and social studies assessments by individual academics were unfavorable. (The

content review reports are not available on the Abell Foundation at the time of writing this report.)

2. New Superintendent Jerry Weast of the Montgomery County school system, one of the largest systems in the US and a successful and highly influential system in Maryland, publicly resisted MSPAP and the state's school reform efforts. When 2001 school performance scores declined—in some cases by 10 percent or more—Weast and others argued that the decline must have been caused by errors. The outcry spread quite quickly (see <http://pqasb.pqarchiver.com/washingtonpost/access/110344428.html?FMT=ABS&FMTS=ABS:FT&date=Mar+11%2C+2002&author=Nurith+C.+Aizenman&pub=The+Washington+Post&edition=&startpage=B.01&desc=Once-Lauded+MSPAP+Undermined+by+Format>).

I believe that MSPAP would have survived these assaults, primarily because of its soundness and the political will, skill, and influence of State Superintendent Nancy Grasmick.

3. In the end, two requirements of No Child Left Behind defeated MSPAP: testing in grades 3-8 and high school, which made MSPAP unaffordable, and reporting individual student scores to families by August, which was not feasible for MSPAP.

ADVICE, AS REQUESTED

The workshop Steering Committee requested advice, based on the MSPAP experience, on future innovative assessments. I discuss lessons we learned from MSPAP and recommendations for implementing future innovations.

Lessons Learned

We fostered acceptance of MSPAP in much of the educational community in Maryland by involving educators, especially teachers, in meaningful ways. Educators from local school systems worked with MSDE staff to develop the Maryland Learning Outcomes, develop MSPAP tasks, (300 participating teachers were paid for this activity each year), development of performance level

descriptors, setting standards for student and school performance, and as trained, qualified, and paid scorers during the summer (600 each year). We typically asked local school system staff to participate in school report card press conferences. Teachers often volunteered that task development or scoring was the best staff development of their careers. Earlier I commented on the importance of open communication. MSDE staff—from the State Superintendent to program area staff—responded to all requests for information, discussions, and presentations that came from all sources—from state legislative committees to local PTAs and individual families. We learned to be calm and forthcoming in discussing matters of testing program operations, policy differences, and debunking myths and to avoid being defensive in the face of legitimate, and sometimes embarrassing criticisms.

We were ambitious technically and established expectations for changes in school performance and student achievement that were well beyond what the recent past had justified. (Although fundamental changes in writing instruction and the rise in performance on the Maryland Writing Test encouraged us.) The technical soundness of MSPAP vindicated the ambitious technical innovations. And MSPAP performance improvements in several school systems and persistently low performing schools, plus the partially optimistic increases in teachers' expectations for their students documented in a Rand study (Koretz, Mitchell, Barron, & Keith, 1996), illustrate the value in expecting more than the status quo. Achieving those successes required flexibility, persistence, and open communication throughout the conception, design, development, implementation, institutionalization, and validation process. In retrospect, it is surprising to recall how difficult it was for the staff of a state agency to learn to function that way. MSDE did not fund teacher and school leader professional development early enough or adequately. And implementing MSPAP required so much time and energy in the first two years that professional development was a back burner concern. Ideally, we would have planned and provided sufficient funding for professional development on a parallel track with the assessment design.

Implementing Future Innovations

Current, well funded forces enable, even encourage states to incorporate innovative test designs into the next generation of state assessments (e.g., Sawchuk, 2009). Race to the Top funding encourage states to align their assessments to the Common Core standards in English language arts, mathematics, and college and career readiness and to join assessment consortia. The 21st Century Partnership is actively promoting the use of performance tasks and scenarios in state assessment systems (see <http://www.21stcenturyskills.org>). The EdSteps project of the Council of Chief State School Officers plans to provide examples of student work on quality scales in writing, global competence, and other areas (see <http://edsteps.org/CCSSO/Home.aspx>) for use in classroom instruction. Even the political rhetoric promotes innovation (as ironically heedless of the 1990s and 2000s as it may be): “It’s time to stop just talking about education reform and start doing it” (President Obama on November 4, 2009; see <http://www2.ed.gov/programs/racetothetop/executive-summary.pdf>). Proposed assessment consortia focus on balanced systems of summative, interim, and formative assessments and on online testing. The performance assessments of the 1990s like MSPAP illustrate what can work operationally and psychometrically and what innovative assessments can cost, politically and financially. Those states that want to have an opportunity to implement innovative assessment designs that include now-traditional constructed response items and broader, more open-ended performance tasks, projects, and portfolios either in statewide summative assessments or statewide or local interim assessments.

My advice is to capitalize on this opportunity, but in measured fashion. Several of the high profile state assessment programs on the 1990s (e.g., California, Kentucky, Maryland, Vermont) implemented statewide summative assessments that were perhaps overly ambitious, too costly, in some cases precluded returning individual student scores, and made timely turnaround of test scores impossible. Incorporating some performance task designs in statewide summative assessment programs is feasible, especially using innovations like automated scoring and online testing (e.g., West Virginia’s WESTEST 2 Online writing assessment). Broader, less easily standardized, and more time consuming performance assessment designs are likely to be more feasible in interim assessments (e.g., quarterly assessments that school systems may implement, as in Howard County,

Maryland). And we should not lose the opportunity presented by the Race to the Top application scoring system places on LEA involvement to focus on training teachers in infusing classroom formative assessment procedures in the instructional process (e.g., D.4 teacher and principal preparation programs, 14 points of 485 total points; support to teachers and principals, 20 points).

Involvement of local educators in state assessment programs can enhance the quality and relevance of the program and can foster support that can counteract opposition. Involving teachers in test development and scoring was beneficial in the 1990s. It was necessary, as states were inventing large scale performance assessment designs and practices and benefited from the insights that teachers could provide. But involvement in development and scoring of operational programs seems less practical today, especially because of requirements to turn around individual student scores within weeks of testing. Newer innovations that currently are under development—computer based tasks that are not feasible in paper-pencil tests, tasks and items that assess student status on learning progressions—likely would benefit from teacher involvement. Perhaps teachers and principals can make their most significant contributions to efforts to design more informative score reports and to use test results and related information to make school program and instructional decisions and plan highly focused school staff professional development.

Computer based, online testing applications continue to expand. For example, Oregon's statewide assessments are adaptive and online; Maryland's science assessments are online; Delaware, Hawaii, and Minnesota are transitioning to online testing. Even those states that plan innovations in paper-pencil testing would be wise to plan now for transition to computer and online administrations. For example, in case a state board or legislature decided that they could reduce testing program costs by requiring online testing (for example, Oregon reports costs of \$15 per examinee for paper-pencil vs. \$4 for online; see [http://www2.ed.gov/programs/racetothetop-assessment/bios/alpert-presentation.ppt#276,10,Computer Based Testing](http://www2.ed.gov/programs/racetothetop-assessment/bios/alpert-presentation.ppt#276,10,Computer%20Based%20Testing)), state programs would be wise to outline a transition plan (e.g., comparability studies, automated scoring studies, cost-benefit studies, start-up costs). In addition, it would be wise to summarize periodically information on research and applications on item types, automated scoring procedures, and so forth that capitalize on computer and online capabilities.

After content standards are developed, test design should begin with intended inferences, not items. That is the current wisdom, though not the current practice. From a practical point of view, a good place to explicate intended inferences is to write achievement level descriptors before test development begins (e.g., Bejar, Braun, & Tannenbaum, 2007; Ferrara, Svetina, Skucha, & Murphy, 2009).

Finally, this is a moment when state assessment programs can use the pressures of federal incentives to coordinate summative testing for accountability purposes with interim assessment that is more relevant to instructional planning and professional development. In this case, the innovation would be implement more than just accountability testing.

REFERENCES

- Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 1-30). Maple Grove, MN: JAM Press.
- Brandt, R. (1992 December-1993 January). On outcome based education: A conversation with Bill Spady. *Educational Leadership*, 50 (4), 66-70.
- Chamblin, L., & Herndon, C. (Eds.) (n.d.). *Education reform in Maryland 1977-1996*. Baltimore, MD: Maryland State Department of Education.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Ferrara, S., Svetina, D., Skucha, S., & Murphy, A. (2009). Test development with standard setting and growth in mind. In I. Bejar (Moderator), *Standard Setting in an Accountability Growth Context: A Process or One-Time Event?* An invited symposium at the annual meeting of the National Council on Measurement in Education, San Diego.
- Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education*, 11 (2), 195-208.

- Grasmick, N.S. (1997). *The politics of large scale assessment*. Keynote address at the National Conference on Large Scale Assessment, Colorado Springs.
- Hambleton, R. K., Impara, J., Mehrens, W., Plake, B. S., Pitoniak, M. J., Zenisky, A. L., & Smith, L. F. (2000 December). *Psychometric review of the Maryland School Performance Assessment Program (MSPAP)*. Retrieved December 1, 2009 from <http://www.abell.org/publications/detail.asp?ID=65>.
- Koretz, D., Mitchell, K., Barron, S., & Keith, S. Final report: Perceived effects of the Maryland School Performance Assessment Program. (CRESST/RAND Institute on Education and Training CSE Technical Report 409.) National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.
- Lane, S., Parke, C. S., & Stone, C.A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment*, 8 (4), 279-315.
- Maryland State Department of Education. (2000 summer). *The Maryland School Performance Assessment Program: What a Decade of Research Tells Us*. Available from the Maryland State Department of Education, Baltimore (see <http://www.marylandpublicschools.org/MSDE>).
- Michaels, H., & Ferrara, S. (1999). Evolution of educational reform in Maryland: Using data to drive state policy and local reform. In G. J. Cizek (Ed.), *Handbook of Educational Policy*. San Diego: Academic Press.
- Parke, C. S., Lane, S., & Stone, C. A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation*, 12 (3), 239-269.
- Sawchuk, S. (2009, August 11). Stimulus seeks enriched tests. *Education Week*, 28 (37). See <http://www.edweek.org/ew/articles/2009/08/12/37measure-2.h28.html>.

Stone, C. A. & Lane, S. (2003) Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education*, 16 (1), 1-26.

Yen, W. M., & Ferrara, S. (1997). The Maryland School Performance Assessment Program: Performance assessments with psychometric quality suitable for high-stakes usage. *Educational and Psychological Measurement*, 57 (1), 60-84.

END NOTES

¹ Peg Luksik, a Pennsylvania education and political activist, ran in gubernatorial primaries and general elections for the Republican Party (1990) and Constitution Party (1994, 1998) and plans to oppose Senator Arlen Specter's reelection bid in 2010. See http://en.wikipedia.org/wiki/United_States_Senate_election_in_Pennsylvania,_2010 and http://www.tribdem.com/local/local_story_068232130.html.

² A middle school teacher in Montgomery County actually conducted an (ill-advised) survey in his classroom as part of a project in social studies on teenage sexual behaviors around the time of the MSPAP administration.

³ Unfortunately, I was unable to meet with the picketers. I had locked my keys in my car as I rushed into an elementary school in Baltimore to observe test administrations.