

Position Statement on the use of the Michigan Educational Assessment Program (MEAP) for High Stakes Decisions for Students in Michigan

PART I: INTRODUCTION

Purpose statement

In April 2003, the National Association of School Psychologists (NASP) adopted a position statement on the use of Large Scale Assessment for High Stakes decisions. The Michigan Association of School Psychologists (MASP) subsequently formed a subcommittee to review the NASP position statement and specifically relate it to the Michigan Educational Assessment Program (MEAP) and Education YES! accountability system. The following position statement serves as a guide to critical stakeholders including teachers, parents, administrators, policy makers and legislators involved in high stakes decision-making in Michigan.

Statement of the problem

Federal Legislation. The Elementary and Secondary Educational Act (ESEA) of 2001, known as No Child Left Behind (NCLB), has redefined the federal role in K-12 education. With far reaching implications and unprecedented expediency to implement the mandates, this legislation strongly promotes standards-based accountability and reform in general education. More specifically, NCLB mandates regular large-scale assessment for all third through eighth grade students to ensure progress towards high academic standards in reading and math (science to follow). It is up to individual states to choose the tool to measure academic achievement. These assessments are considered “high-stakes” because consequences, by way of incentives or sanctions, are directly experienced by students, teachers, and administrators (Braden, 2002). Performance on these assessments is supposed to be summarized for groups such as minorities, those with low socioeconomic status, students with disabilities, and students who have limited English proficiency. NCLB intends to close the gap between high and low performing groups of students, especially between these subgroups of students, by sanctioning schools who fail to make Adequate Yearly Progress (AYP), a pillar of the federal legislation. The federal goal is to have 100% reading and math proficiency across the county within the next 12 years (2013-2014 school year).

Prior to the new NCLB legislation, most states including Michigan, were in some way using an accountability system which often included statewide assessment, to rank and report how schools were performing. The results of these accountability systems included publishing the results, which often led to students and teachers experiencing unintended outcomes (Barksdale-Ladd & Thomas, 2000; Smith & Fey, 2000) such as a restricted range and depth of curricula, inequitable educational opportunities due to tracking, and test corruption (Braden, 2002).

In contrast, with the NCLB legislation, far more serious consequences exist for schools that fail to meet a predetermined threshold of proficiency including removal of funding and the power to close or take-over the “low performing” schools. For individual students, poor performance on high stakes tests includes consequences such as grade retention and denying graduation from high school. Across the country, sanctions associated with such requirements as these have too often resulted in misuse of test results and had negative consequences for students, parents, teachers, as well as reactive policy development at the local and state level (Halladyna, 1992).

Public Impact. It must be acknowledged that there is an enormous impact associated with the public's perception of the quality of their schools. Just as children need confidence that they can succeed, so also

do parents need confidence that their school can succeed. The public perception of a “good” school in a neighborhood has a profound effect on home values, the attractiveness of the area to industry, the degree of pride or involvement in the school, willingness to approve school bonds, etc. When the public perceives a poor performing school, effort and resources are often not refocused for the troubled school, but are instead drawn away from the school. Prior to federal legislation, families had the option to enroll their children in public school academies or move away to neighboring districts, which often made passing bonds for school improvement difficult. With NCLB legislation, parent choice is the backbone of the plan, giving parents even more options including the ability to evoke school of choice options, demand that the school pay for public transportation to a new school, or demand supplementary services such as tutoring and summer school at the district’s expense. All of these events are usually based on the perception of the quality of the school from information provided to the public as a result of high-stakes testing.

The Achievement Gap. The new legislation has changed the public’s expectations about equality for children. What originally meant “equal in rights” has become misinterpreted as “equal in capability.” The expectation has come to be that all children, regardless of background or capability, can achieve equally in school. If they do not, then it is assumed that the school is at fault. However, for many years, educators have known that a child’s background can have a profound effect on their educational progress and that children do not develop skills in a “locked-step” fashion. Classroom teachers know that both the child’s readiness as they enter school, and their talents, can affect their progress, even in schools with the best resources. The child’s readiness is directly related to their literary experiences before entering school, which is associated with family resources. Hence, children from disadvantaged homes often enter school already behind, and after several grades the achievement gap is very difficult to correct, even under the best circumstances. This means that schools with greater proportions of disadvantaged children are much more likely to have lower test scores.

This persisting achievement gap based on a child’s background is generally not related to the effectiveness of the teacher. However, when the public is encouraged to choose their schools based on test scores, they erroneously believe that the better teachers must always be in the schools with the highest scores, those with the most advantaged children. An unfortunate additional consequence of the public being told that all children can achieve equally, is the belief that if there are differences between them, it must be due to deficits in character. Such deficits are often unfairly attributed to minorities because they are disproportionately represented among the disadvantaged.

As taxpayers, the public has the right to understand how their neighborhood schools are performing and what steps are being taken to improve achievement for all students. It is critical that a system of measurement of instructional effectiveness be based on valid and reliable methods. When facts are given to the public about the quality of instruction, those facts must not be inaccurate or misleading, and should be useful in pointing to a practical solution to any identified problems in schools.

Appropriate use of assessment in education

MASP unequivocally believes in accountability, program evaluation, and monitoring the progress of students. Large-scale assessments are helpful, as a part of a broader evaluation system, in understanding the educational needs of large groups of students. However, it is not helpful to assess students globally and sanction individually (i.e., retention). Large-scale assessments are only valuable when followed with additional instructional interventions for diverse groups, comprehensive professional development and needed resources so that all students can make appropriate progress.

MASP believes in the following basic assumptions about education:

- All children can learn
- Student performance should increase for all students, including those students who are most vulnerable, such as students with disabilities, those not native speaking, those whose families lack economic resources, or those who come from unstable homes and suffer frequent mobility.
- Educational instruction, assessment, and values should accommodate diversity in children.
- Expectations of student progress should be based on high but realistic expectations. We should respect diversity of skills as much as we applaud intensity of effort.

PART II: REVIEW OF ISSUES

Background information on NAEP, MEAP, ED YES!

NAEP. Since 1969, a test known as the National Assessment of Educational Progress (NAEP) has been administered across the nation as a nationally representative assessment of student performance in reading, mathematics, social studies, science, and other fields. Referred to as the Nation's Report Card, this assessment measures student proficiency in math, reading, writing and science. (Since NCLB was signed into law, all states receiving Title One funds must participate in the reading and math sections of the NAEP every two years). The NAEP assesses a representative sample of students and reports the percentage that perform at various levels: *Basic, Proficient, and Advanced*. While there remains considerable argument about the rigor required to obtain a proficient level, 32% of Michigan's fourth grade students and 33% of eighth grade students scored at or above proficient in reading (NAEP, 2003). In math, 34% of fourth grade students and 28% of eighth grade students scored at or above proficient (NAEP, 2003). The NAEP also serves as a broad comparison to other states; however, if the purpose of large-scale assessment is to inform curriculum adjustments and increase performance of all students, the NAEP is too far removed to be helpful in understanding local district curriculum and programmatic needs.

MEAP. Assessment of all students in Michigan began with the use of normative-referenced, commercially available tests (1969-1973) to measure student achievement. This information was not measured against any local standard. The Michigan Revised School Code (1977) and the State Aid Act (1979) required the establishment of educational standards and a corresponding assessment program. The MEAP assessment program was subsequently designed to provide locally based information regarding the educational progress of students on the state curriculum at vital points during their educational careers. The MEAP is a criterion referenced test designed to align with the Michigan's Curriculum Framework that defines content standards and grade level benchmarks in the four core subject areas: language arts (includes reading and writing), math, social studies and science.

While the MEAP program began in the early 1970's to align curriculum across the state, the present purpose of MEAP is multiple-faceted, not just to inform instruction. This radical change is reflected in changes across the country related to intense focus on accountability for results. Issues of accountability at all levels (i.e., taxpayer's need to know how their money is improving education, districts want to compare their scores to other districts to secure taxpayer base, and legislators making comparisons to other states) have influenced changes in the intended purpose of testing. The performance on MEAP is linked with earning a scholarship for college bound seniors, state funding, and even possible state takeover.

Education YES! Program. Accountability systems are required under No Child Left Behind (NCLB) legislation. Prior to the passage of NCLB and similar to other states focused on standards-based accountability, Michigan implemented Education YES!- A Yardstick for Excellent Schools (2003) as an accreditation system that issues a letter grade to all schools. Grades are "A", "B", "C", "D-Alert", "F-Unaccredited". According to the State Board of Education, a school's rating serves to assist schools in identifying instructional needs and to provide parents with easy to understand information about their

home school. Education YES! includes multiple indicators at different points in time to determine a school's grade. The result is that each school in Michigan receives four grades: MEAP Achievement *Status* (a report of the school's three-year history of MEAP achievement), MEAP Achievement *Change* (rates the degree of change in MEAP performance), School Performance Indicators (self-assessment based on student engagement, instructional quality, and learning opportunities), and a *composite score*. The State Board of Education does consider if the school makes Adequate Yearly Progress as part of the determination of accreditation.

While Michigan's Education, YES! school accreditation program *alone* does not have any corrective actions aside from public praise or public criticism, failing to make AYP according to NCLB legislation has serious and swift sanctions.. As required by federal NCLB legislation, all states must use a performance standards and a testing program to determine AYP status. Although many states do not have such systems in place, Michigan has had performance based standards associated with the MEAP for decades. Michigan is using data from as early as 1997 to determine the baseline for AYP status.

School either make AYP or do not make AYP based on meeting all three of the following criteria: 95% of the school must participate in Michigan's assessment program, 2) All groups of kids (i.e. entire grade level and associated subgroups) need to meet a pre-specified proficiency threshold 3) elementary schools must meet a attendance rate threshold, while for high school, graduation rates are the final criteria. Failure in any one of the three criteria will result in the school being labeled as not making AYP. Sanctions are associated with the number of years a school fails to make AYP. The target starting points to determine status include 47% proficient in elementary math, and 38% proficient in elementary reading. As of December 2003, Michigan's criteria resulted in 216 schools being identified for continuous improvement or corrective action in reading and/or math.

Reliability and Validity of the MEAP

MEAP Reliability. Test reliability can be thought of as a consistency in measurement (Gregory, 2000) and is an estimate of the measurement error within a given score (Anastasi & Urbina, 1997). It is usually expressed as a correlation coefficient between two sets of scores, which results in a potential range of – 1.0 (perfect inverse relationship) to +1.0 (perfect linear relationship). Salvia and Ysseldyke (2004) warn against using tests with poor reliability and suggest minimum standards of .60 for data used for group decisions, .80 for tests used for screening decisions about individuals, and .90 for tests used for important decisions about individuals. However, it is important to note that reliability is not a universal concept. In other words, tests are only reliable for specific populations (Alessi, Clarizio, Bradley-Johnson, Lesiak, & Lesiak, 1981) and purposes (Gregory, 2000). Therefore, there are different approaches to reliability. Test-retest measures score consistency over time, internal consistency measures the homogeneity of the test items, alternate form examines the consistency across forms, and interrater reliability is an indication of consistency between test administrators/scorers.

The Michigan Department of Treasury (MDT, 2000) published the internal consistency and interrater reliability data for MEAP tests given during the 1998-1999 school year. Table 1 lists the correlation coefficients of the internal consistency estimates. Three of the coefficients exceeded the .90 minimum needed for important decisions, and all but one of the remaining tests exceeded the .80 screening minimum. The reliability coefficient of the eleventh grade writing test exceeded the minimum standard for group tests (.60), but did not meet the higher levels of the other tests.

Table 1
Internal Consistency Coefficients for 1998-1999 MEAP Testing

Test	Reliability Coefficients				
	Grade 4	Grade 5	Grade 7	Grade 8	Grade 11
Reading-Story	.814	NA	.891	NA	.830*
Reading-Informational	.809	NA	.902	NA	NA
Mathematics	.931	NA	.962	NA	.892
Science	NA	.886	NA	.892	.878
Social Studies	NA	.882	NA	.883	.888
Writing	NA	NA	NA	NA	.610

*11th grade test reading does not differentiate Story from Informational

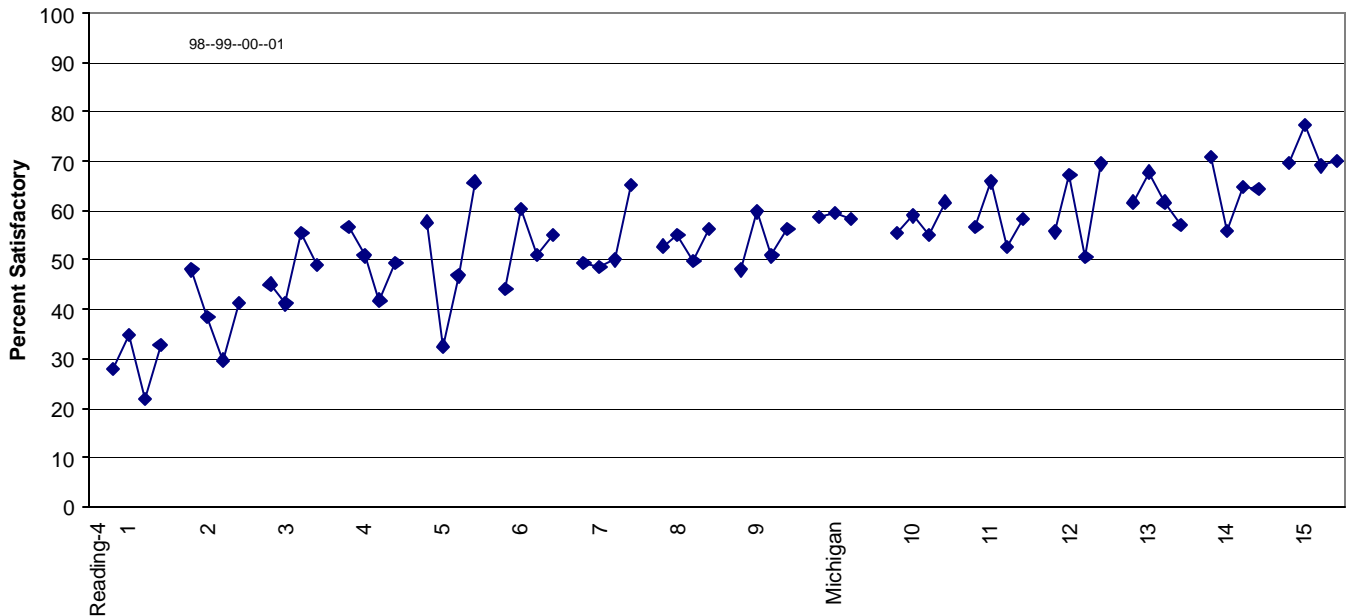
Data provided by the MEAP tests appear to demonstrate sufficient internal consistency to be used in group decision making, and with the exception of eleventh grade writing, can be used to make screening decisions about individual students. This presents a concern about potential uses of the data for eleventh grade students. The MDT currently uses this data to decide whether or not the state will endorse the local diploma, and to make decisions about Merit Award Program scholarship recipients. Reliability coefficients for the eleventh grade tests ranged from .61 (writing) to .89 (mathematics), with four of the five tests falling between .83 and .89. Therefore, based exclusively on internal consistency coefficients, data from the 1998-1999 MEAP testing is not sufficiently reliable to make important decisions, and the writing test is not appropriate for any decisions about individual students. While some might call the MEAP a group test, which therefore requires a lower reliability standard, it is important to note that the use of the data determines the type of test, not how it is administered.

Two other points of concern should be examined when discussing MEAP reliability. First, only internal consistency and interrater reliability was released to the public. While both forms offer important data, interrater reliability only supplements other estimates of reliability and internal consistency has been criticized for a lack of precision (Anastasi & Urbina, 1997). Test-retest reliability assesses consistency between scores from different administrations and estimates measurement error associated with administration schedule. This form of reliability has been described as the most straightforward and as essential to evaluate a test's reliability (Gregory, 2000). No data regarding this approach are offered from the State of Michigan.

Second, the actual scores on high stakes tests such as the MEAP will vary according to sampling error and measurement error (Hill & DePascale, 2003). Sampling error refers to the reliability of testing a different group of students year to year. Measurement error refers to variation in scores associated with testing students on a particular occasion. Contrary to what might be expected, fluctuating scores on high-stakes tests are more likely due to sampling error. This is consistent with many perceptions from teachers and principals that the range of student skills is not fixed, but can fluctuate in unpredictable ways from year to year, despite having similar curriculum and veteran teachers. Even though the score of a current group of students may not be representative of typical children across the years in the same building, a school could still be seen as failing.

Figure 1 is a chart of 15 schools from a single district (including the Michigan average) of fourth grade MEAP reading results over a four-year period. The results are arranged by the average percent proficient. By examining the variation in the four scores attributed to each school, one can see the large variation in the percent of students proficient in reading. These fluctuations occur both in more-proficient and less-proficient schools, and are likely due to sampling differences and not any real differences in curriculum or teacher style. It is important when interpreting differences in reading scores, that they be above the threshold of what may be expected due to sampling error.

Figure 1
1997-98 - 2000-01 MEAP Results by School
4th Grade Reading
(E. Bauer, Oakland Schools 2003)
(Arranged by average over 4 years.)



Another issue is that for individual elementary schools, the number of students at each grade level is small. Because most of the analysis is done with small subgroups of students, small fluctuations in the percent proficient can cause large changes in assignment of the school's change grade in Education YES! In fact, the number of students in a grade can have more impact on the mean than the reliability of the individual student scores. Finally, a school will only get "credit" for moving students from one threshold or category to the next (apprentice, meets expectations, exceeds expectations) as opposed to relative increases in mean performance. For schools that have been implementing systematic changes in curriculum, these "all or nothing" categories may not accurately credit their obtained changes in student performance across the curriculum.

Yet another point of concern involves the lack of data provided for various subgroups of test takers. As stated earlier, tests are only reliable for the population being tested, and data should be provided for at least age or grade groupings (Salvia & Ysseldyke, 1998). In addition, it is recommended that test developers provide reliability estimates for subgroups of the general population that may be susceptible to test bias due to racial, ethnic, disability, or linguistic differences (Hammill, Bryant, & Pearson, 1998). Although reliability data for the MEAP are presented by grade groupings, no information is offered to support the consistency of the data for various subgroups.

MEAP Validity. Test construction procedures outlined by the MDT suggest that the MEAP tests are well developed tools. Each item is based on curricula approved by the State Board of Education, and is subject to content and bias reviews, and empirical evaluation. These careful efforts to assess the content of each item help assure strong content validity, which is defined as the extent that the test items represent the domain being assessed (Salvia & Ysseldyke, 1998). However, content validity is only one aspect of a sound tool. In order for a test to generate useful data, it must have evidence for validity. In other words, inferences made from the data must demonstrate appropriateness, meaningfulness and usefulness (AERA, APA, & NCME, 1999). The three most common approaches to validity are content,

criterion-referenced, and construct validity (Gregory, 2000). Some (Anastasi & Urbina, 1997) have argued that construct validity, the appropriateness of inferences made from test data about the underlying quality or trait being assessed (Gregory, 2000), is the all-inclusive approach to validity.

While the MDT has provided adequate evidence for content validity, which is essential for an achievement test (Salvia & Ysseldyke, 2004), they seemed to have dismissed the other approaches to validity. Test developers concluded that there was no other test that matched the purpose of the MEAP, and therefore no criterion evidence could be provided. In addition, they offered only a four-paragraph description of construct validity and the dilemmas surrounding the concept with no data addressing construct validity of the MEAP. MDT is correct in not using tests such as the ACT or SAT for a criterion-referenced evaluation. However, to simply dismiss the issue is psychometrically indefensible (Burns, 1998). In fact, it is difficult to determine the technical adequacy of the MEAP because the Michigan Department of Treasury has not published technical manuals for the MEAP tests. However, some independent studies have been conducted.

Saginaw Public Schools (1993) used high school grade point average (GPA) as a criterion measure with which to correlate the MEAP high school test. The results of this study are listed in Table 2. While there may be difficulties with GPA as an assessment tool, teacher judgment of achievement has been described as acceptable criteria with which to evaluate validity (Salvia & Ysseldyke, 1998). The overall correlations between MEAP scores and GPA ranged from .365 for Story Selection to .551 for Mathematics. The resulting coefficients suggest a questionable relationship. Therefore, if grade point average is an acceptable representation of teacher judgment of achievement, then some concern exists regarding the validity of the data provided by the high school MEAP. Additional data provided by racial groups suggested low criterion validity for the Story Selection test and poor validity for Hispanic students in general. It should be noted that these findings are based on now outdated versions of the MEAP, and replication of the study is needed. In a similar study, Bauer (1997) found similar correlations using the 1995-97 High School Proficiency Test. Using career grade point average and HSPT from 18 districts (n = 6,745) found a .62 correlation with math, a .56 correlation with science, a .50 correlation with reading and .53 correlation with writing.

Table 2
Criterion-Related Validity Coefficients for GPA and High School MEAP

Test	White	African-American	Hispanic
Story Selection	.320	.408	.422
Informational Reading	.492	.463	.315
Mathematics	.535	.553	.377

Heller (2001) reviewed data from the MEAP High School Proficiency Test used to assign Michigan Merit Scholarships and concluded that 1) minority students and students in poorer high schools qualify for scholarships at lower rates and 2) the merit scholarships were being awarded disproportionately to students who were most likely to attend college even without the financial assistance. These findings, along with the lower correlations between GPA and MEAP scores for children from various ethnic groups, reveals a minority disadvantage, which further suggests questionable validity for many groups of children.

Another important characteristic affecting test validity is test specificity. In a test made up of several components, the parts of the test should be sufficiently different from one another to permit analysis by subtest score. In a study of 800 students' scores on the MEAP and the Metropolitan Achievement Test, a single factor accounted for 75% of the variance in all five MEAP and MAT subtests, as the factor loadings of all five tests on the single factor were above 0.77, with three above 0.90 (Martineau, Somers, & Dehlin 2000). The results showed that the MAT and MEAP tests, despite being norm-referenced and criterion-referenced, respectively, measure similar constructs, regardless of subject matter. Students doing well in one area tend to do well in another; and if they do poorly, they tend to do poorly across the board. This suggests that the tests measure a common cognitive factor more than a specific curricular one. Are we testing talent or learning?

Non-instructional variables.

The literature contains extensive documentation of a number of non-instructional variables that negatively affect scores on achievement tests (Bezruczko & Reynolds, 1992; Butler, Marsh, Sheppard & Sheppard, 1985; Chall, Jacobs & Baldwin, 1990; Grissmer, Flanagan, Kawat & Williamson, 2000; Klinge & Warren, 1990; Lytton & Pryrt, 1998; McNamara & Deaton, 1985; Vanmoorlehem & Newman, 1998). The variables include family income, parent education level, family literacy activities, family mobility, and ethnicity. More recent studies have documented similar effects on high stakes tests.

Burns, Courtad, Hoffman, and Folger (in press) collected district-level data such as percentage of students receiving free or reduced breakfast and/or lunch (FRL), pupil/teacher ratio (PTR), and foundation allowance per pupil (FA) for Michigan school districts. These variables were compared to MEAP mathematics, reading, and writing test scores for grades 4 through 8 in each Michigan school district. Results of this study suggested that the FRL was the only significant predictor of statewide achievement scores. The conclusion was that MEAP scores are significantly influenced by community socio-economic status, which is a variable that is irrelevant to the measurement of academic achievement. It would seem that more instructionally relevant variables, such as pupil/teacher ratio and foundation allowance, would influence scores more than construct-irrelevant variables such as FRL. Thus, the MEAP was reportedly less of a measure of academic achievement than of the construct-irrelevant variable of community socio-economic status (Burns et al., in press).

Myers and Curtis (2003) found that in 600 elementary schools, poverty was a major predictor of scores on the Florida Comprehensive Assessment Test (FCAT). As noted in that study,

“Because so much of the difference in FCAT performance among schools can be explained by poverty, it appears premature to assume that poor teaching is a chief cause of low school achievement... Poverty is primarily a societal problem of economic deprivation that becomes an educational problem because students arrive at school at age five lacking knowledge and skills, which in turn causes below-average performance on whatever tests educators and legislators decide to give them. The problem of lower expectations, whether it comes from parents, teachers, or peers, can negatively influence achievement, but these effects are clearly smaller in size when compared to the effects of poverty.”

Table 3.

Individual effects of risk factors on initial achievement and yearly growth.

Risk factor	Effect on MAT reading scores*				
	Initial achievement	Year 2 gain	Year 3 gain	Year 4-5 growth	Net
Student is female	0	0	0	-5.4	-5.4
Student receives free/reduced lunch	-8.9	0	0	0	-8.9
Student is Black/African-American	-7.8	-8.6	0	0	-16.4
Student's mother grad from HS vs 4-year college	-10.2	-7.8	0	0	-18.0
Spanish is language spoken in student's home	-19.4	0	0	0	-19.4
Student moves every year versus never	-16.9	-19.7	0	0	-36.6
Student receives special education services	-21.1	-17.0	0	0	-38.1
Total effects of all risk factors	-80.6	-53.1	0	-5.4	-139.1

* The numbers in the first column represent the incoming scale score difference between students with and without the risk factors listed. The numbers in the four right-most columns represent the scale score difference in growth of students with and without the listed risk factors.

Noninstructional influences were studied in more detail by Martineau, Somers and Dehlin (2001). Table 3 illustrates the effects, which are cumulative, of each background variable upon incoming reading scores on the Metropolitan Achievement Test and growth thereafter. The same variables were found to affect MEAP scores, but a growth measure could not be obtained on the MEAP because it was not given annually to all grade levels. By examining how the variables affect the MAT, a widely known achievement test, we can postulate similar effects on growth on the MEAP. The above table illustrates that of the home influences, family mobility has the greatest impact on initial scores in first grade, and it has an additional influence on growth through second grade, after which there is no additional influence of the variable on growth scores. Indeed, further effects of all home background variables on growth seem to disappear after second grade. Since the variable effects are cumulative, one can see that students with one or more of the risk factors (family mobility, non-English-speaking home, low parent education level, low income, and being African-American) tend to start school already behind, experience more decline in growth through second grade, and never recover to the level of their advantaged peers, even though their rate of growth in later grades may be the same. The results emphasize the importance of allowing for these important noninstructional influences when comparing achievement scores.

Unintended Consequences of High-Stakes Tests

There is tremendous pressure upon school district administrators, teachers, and parents to have students participate in MEAP testing. There are reports throughout the country of narrowing of the curriculum to increase test scores, teaching to the test, outright cheating, exclusion of students who are likely to do poorly, chronic teacher dissatisfaction, and poor teacher retention (Haladyna, 1992). Some students, in fact, may experience some anxiety and depression. Teachers and administrators are fearful of losing their jobs and students dread the weeks of testing, despite incentives by teachers to gain their participation. In a study of schools who were sanctioned for poor performance, the sanction only "weakly" influenced teacher's work motivation, but had an overall negative impact on job commitment. (Minrop, 2003).

Instructional time: cost and benefit. The MEAP has evolved over the 30 years of its implementation. One critical change has been a movement from assessing basic skills to measuring higher standards of proficiency in core academic areas. While this change is not harmful by itself, it changes the purpose of testing and the time devoted to the assessment. In fact, the test itself has become so intensive, that

weeks of instructional time are devoted to MEAP preparation, and hours to complete the assessment. The enormous amount of time spent on test preparation is in place of actual instructional remediation for students who are struggling. For failing schools, additional resources are withdrawn as a result of the scores. These events often occur despite the fact that we can predict with great accuracy what schools are more likely to be in need of improvement prior to taking the MEAP.

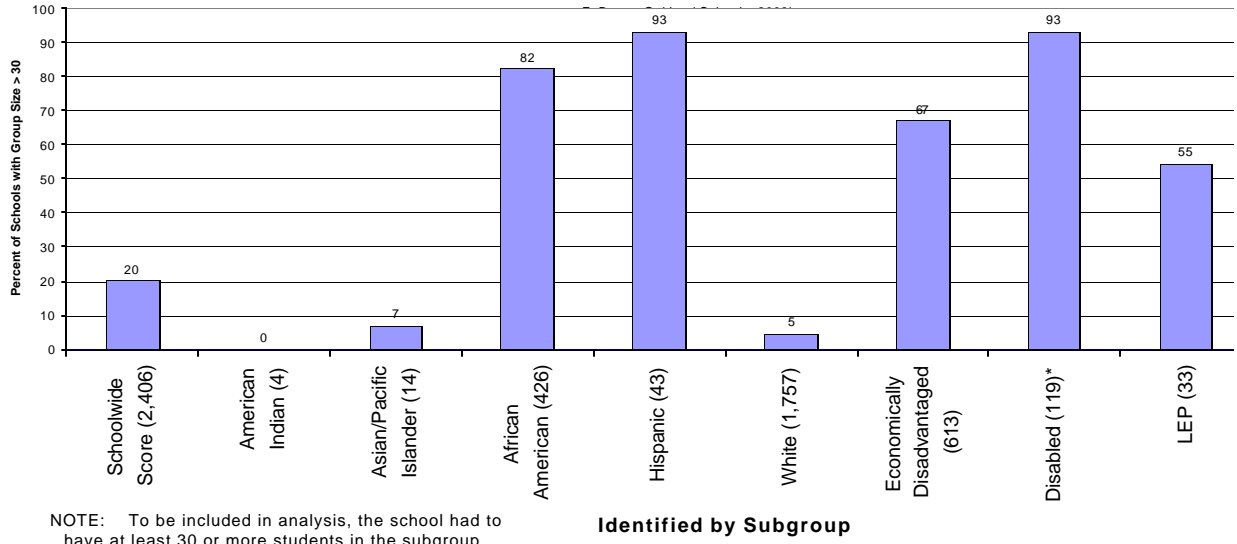
Theory to Practice: Does High Stakes Assessment work? Substantial research is underway to determine the actual effectiveness of high stakes testing. Accountability systems linked to high stakes testing are thought to improve teacher motivation, and increase student performance. Amrein and Berliner (2002) conducted an extensive review of the use of high-stakes testing across the nation. In this study, the authors reviewed the performance of 18 states (Michigan not included) engaged in high stakes testing and compared their results with alternative tests measuring student achievement (ACT, SAT, NAEP or AP) before and after implementation of their high stakes testing program. The results demonstrated that improved student scores on the high-stakes test were not generally reflected on the alternative tests. This suggests that high-stakes scores improved as a result of teaching to the test, and not because of a general increase in the skill area. More studies need to be conducted and replicated to draw conclusions about the adequacy of high stakes testing. This research should be encouraged and should be considered as we balance the cost and benefits of high stakes assessment in Michigan.

Incentives: Equitable or Inequitable? In Michigan, because participation in high school MEAP was so poor several years ago, proficiency on the MEAP was linked to scholarships for students to use in their first year of college (i.e., Michigan Merit Award) funded by a settlement with tobacco companies arising from health care costs of treating smoking-related diseases. The recipients of these awards have tended to be from more affluent neighborhoods. Ineligibility for the scholarships is strongly associated with both racial/ethnic background and poverty (measured by percent of free and reduced lunch). For instance, while 21.9% of white students qualified for the scholarship, only 2.8% of African-American students and 11.3% of Hispanic students qualified (Heller, 1999). In addition, 23.1% of students qualified in schools with low poverty (3% free and reduced lunch), whereas only 7.4% of students qualified from higher-poverty schools (41% free and reduced lunch).

Retention. Some states have used their high-stakes assessment to determine grade level retention as a tough way to increase student success, gain parent support, and increase the performance of teachers. Despite the allure of this strategy, there is consistency in the research community that retention is a failed strategy to improve the academic performance of at-risk students. In fact, while there is an initial gain in academic achievement during the initial retention year, these gains are diminished over subsequent years. In contrast, students who are retained are more likely to drop out of school than non-retained students and are at-risk for numerous mental health related problems. Retention based on test scores is conducted in many states including Louisiana, New Mexico, North Carolina and Florida. (Amrein and Berliner, 2002). For states such as Florida, the cost of retention proves to be fiscally irresponsible. In Florida, based on the FCAT (2000), if all fourth grade students were retained as stated based on low scores, it would be 65,000 students repeating that grade at a cost of 200 million dollars (Florida Association of School Psychologist, 2002). In a separate report, the number of students who must repeat the third grade based on their level-one performance on the FCAT was five times greater than in the previous year (Associated Press, 2003).

Special Populations. Since NCLB and establishing AYP status is centered on achievement for all students, *diversity* can be misinterpreted as a *liability*. While this is against both the spirit of NCLB and special education, schools that serve a diversity of student populations will likely be identified for improvement. These subpopulations are not uniformly distributed across schools and districts. In fact, Michigan ranks second in the country of segregated school populations (MIRS, 2003). To illustrate the disadvantage that schools who engender diversity face, Bauer (2003, Figure 2) used the stringent criteria required by federal law and graphed the number of schools with 30 or more students in each subgroup population by the percent of schools that would be identified for improvement. While only 5% of schools with subpopulations of white students would be identified, 82% of schools with subpopulations of African American students would be identified.

Figure 2
 Percentage of Schools Identified for
 Improvement Using Strict NCLB Definition
 Data computed from Archival Data
 (Bauer, E. , 2002)



Part III: SUMMARY/ CONCLUSIONS

The MEAP was initially created by Michigan teachers and local curriculum specialists in Michigan, to monitor how large groups of students were progressing in regard to Michigan's Benchmarks and Standards. The MEAP test is a criterion-referenced test that provides teachers with the ability to measure proficiency in several academic areas. Within-district comparisons of students of similar background can be helpful to teachers and administrators, as opposed to comparing Michigan's students to a national sample, which engenders little overlap with the state's curriculum expectations. The MEAP has turned into a high stakes assessment instrument, because it is now used to determine Adequate Yearly Progress and to establish a school's report card with both rewards and sanctions. However, research has found MEAP scores to be related to student background such as socioeconomic status, which limits the MEAP's usefulness as a measure of instructional comparison. There are specific conditions under which MEAP scores may be useful, which are noted below:

Part IV: RECOMMENDED POLICIES REGARDING HIGH STAKES ASSESSMENT IN MICHIGAN

Retain the MEAP if applied in the specific manner noted in the items below: While MASP has concerns regarding the developmental appropriateness of the Michigan Curriculum Benchmarks for all students, MASP believes that tests within the MEAP can be better tools than using a nationally normed test that is unrelated to the state's curriculum expectations. However, the MEAP needs further refinement and a technical manual. Our recommendation to keep the MEAP is made with caution and the MEAP should only be applied in the specific manner noted below.

- **Use the MEAP at this time for group, not individual progress:** MASP believes that at this point in its development, the MEAP should be linked only to its original purposes: align district curriculum with the Michigan Curriculum Benchmarks and to assess overall progress of groups of students towards meeting those benchmarks. The MEAP is not yet an appropriate assessment measure to determine instructional decisions for individual students. Separate validation studies must be conducted to support such further uses of the MEAP in Michigan.
- **Add vertical scaling, yearly testing, and use rate of growth as a measure of progress:** MASP opposes cross-sectional comparisons (comparing groups across the same grade) because the groups are often unequal and normal fluctuations in a school's sample do not reflect instructional variables. We endorse the development of vertical scaling (expanding the MEAP so that there is a sufficient item range to reflect the curriculum across multiple grade levels) to enable scores to be compared longitudinally. MASP supports following groups of students to determine instructional needs based on their rates of growth, rather than their progress toward a fixed standard. Research has shown that rate of growth is not affected by background influences after second grade. MASP recommends an eventual system using each individual's rate of growth (value-added method) as a basis for determining Adequate Yearly Progress (AYP).
- **Do not use single scores for *student* sanctions or rewards:** MASP opposes such high-stakes practices as linking performance on a single test score, such as MEAP, to individual student decisions such as grade promotion, retention, instructional placement, graduation, or eligibility for scholarship. Consistent with best practice, MASP supports the use of converging, multiple sources of data to make any individual student decisions. Such variables may include, but are not restricted to grades, curriculum based measurement, teacher evaluations, and parent input.
- **Do not use scores for *school* sanctions or rewards:** MASP opposes using rewards or sanctions for staff, school or district based on their students' performance on a single test, because schools have little control over non-instructional influences such as student background that significantly impact achievement. Further, incentives have tended to occur in affluent communities and are strongly associated with ethnicity and income.

- **Give and score the test in the fall so instruction can be adjusted:** MASP recommends the MEAP be given *and scored* in the fall, which allows teachers to make instructional adjustments during the second semester with the same group of students. Early administration of the test will also limit the inordinate amount of time some schools spend in MEAP preparation.
- **Report Mobility Rates** Although moving once or twice during the public school years may not be particularly detrimental, most research shows that high mobility lowers student achievement, especially when the students are from low-income, less-educated families (Sewell, 1982; Straits, 1987). Since mobility rates are a non-instructional influence on achievement, schools should be required to report the mobility rates for each grade level assessment.
- **MDE should regularly caution the public about misuse of test results:** MASP supports a stronger role by the Michigan Department of Education to distribute information about the appropriate use of large-scale assessments, including cautions to educators, media and parents about the misuse of such test scores in demographic and district-by-district comparisons. If the state wishes to contract these services, the contractor should provide full disclosure regarding appropriate use of test scores.

Michigan Association of School Psychologists

High Stakes Subcommittee

Matt Burns, Ph.D.
Susan M. Petterson, Ph.D.
James G. Somers, M.Ed.

Reviewers

Yolanda Mojica, Ph.D.
Cheryl Somers, Ph.D.

and

Ernie Bauer, Ph.D.
Statistical Consultant
Oakland Schools

References

- Alessi, G., Clarizio, H. F., Bradley-Johnson, S., Lesiak, J., & Lesiak, W. (1991). Non-biased assessment of high incidence special education students. Lansing, MI: MDE.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing (2nd ed.). Washington, DC: American Psychological Association.
- Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Amrein, A. L. and Berliner, D. C. (2002). High stakes testing, uncertainty, and student learning. Education Policy Analysis Archives, 120 (80), 1-67.
- Associated Press (2003). Test has many kids repeating third grade. Detroit Free Press. Available online: <http://www.freep.com>
- Barksdale-Ladd, M. A., & Thomas, K. F. (2000). What's at stake in high-stakes testing? Journal of Teacher Education, 51, 384-397.
- Bauer, E. (1997). High school proficiency test. Unpublished paper.
- Bauer, E. (2003a). Meap fluctuations by schools over a four-year period. Unpublished paper, 11-15-03.
- Bauer, E. (2003b). Education YES and NCLB Adequate Yearly Progress. A paper presented at the Policy Analysis conference on December 6, 2003 at Oakland School, Waterford, Michigan.
- Bezruczko, N, & Reynolds, A.J. (1992). Schools, families, and children: Sixth grade results from the 1992 longitudinal study of children at risk. Chicago Public Schools, IL (ERIC Document Reproduction Service No. ED 354 287).
- Braden, J. L. (2002). Best practices for school psychologists in educational accountability: High stakes testing and educational reform. In A. Thomas & J. Grimes (eds.) Best practices in school psychology (4th ed.; pp. 301-321). Bethesda, MD: National Association of School Psychologists.
- Burns, M. K. (1998). Interpreting the reliability and validity of the Michigan Educational Assessment Program. Testimony presented to the Michigan House of Representatives' Subcommittee on the Michigan Educational Assessment Program (MEAP). Flint, MI. (ERIC Document Reproduction Service No. ED 418 138).
- Burns, M. K., Courtad, C. A., Hoffman, H., & Folger, W. (in press). A comparison of district-level variables and state accountability test results for public elementary and middle schools. Psychology and Education.
- Butler, S.R., Marsh, H.W. & Sheppard, M.J. (1985). Seven-year longitudinal study of the early prediction of reading achievement with kindergarten children. Journal of Educational Psychology, 77, 349 - 361.
- Chall, J.S., Jacobs, V.A., & Baldwin, L.E. (1990). The reading crisis: Why poor children fall behind. Cambridge, MA: Harvard University Press.
- Gregory, R. J. (2000). Psychological testing: History, principles, and applications (3rd ed.) Boston: Allyn and Bacon.

- Grissmer, D.W., Flanagan, A., Kawata, J., & Williamson, S. (2000). Improving student achievement: What NAEP state test scores tell us. *Rand*. Retrieved 11/23/00 from the World Wide Web: <http://www.rand.org/publications/MR/MR924/index.html>
- Haladyna, T. (1991). Generic questioning strategies for linking teaching and testing. *Educational Technology, Research, and Development*, 39, 73-81.
- Haladyna, T. (1992). Test score pollution: Implications for limited English proficient students. Focus on evaluation and measurement. Proceedings of the National Research Symposium on Limited English Proficient Student Issues, Washington, DC.
- Haladyna, T., Haas, N., & Nolen, S. B. (1990). Test score pollution. Paper presented at the annual meeting of the American Educational Research Association, Boston.
- Heller, D. E. (1999). Summary analysis of the Michigan Merit Scholarship Program: Spring 1999 MEAP High School Test Results. www.fairtest.org
- Heller, D. E. (2001, October). Standardized tests and merit scholarships. Paper presented to Michigan Council of Teachers of English Annual Conference, Lansing, MI.
- Hammill, D. D., Bryant, B. R., & Pearson, N. A. (1998). Examiner's manual: Hammill Multiability Intelligence Test. Austin, TX: Pro-ed.
- Hill, R. K. & DePascale, C.A. (2003). Reliability of No Child Left Behind accountability designs. *Educational Measurement: Issues and Practice*, p 12-21.
- Klingele, W.E. & Warrick, B.K. (1990). Influence of cost and demographic Factors on reading achievement. *Journal of Educational Research*, 83 (5), 279 – 282.
- Linn, R.L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, 31, 3-16.
- Lytton, H. and Pyryt, M. (1998). Predictors of achievement in basic skills: A Canadian effective schools study. *Canadian Journal of Education*, 23 (3), 281 – 301.
- McNamara, K.T. & Deaton, B.J. (1985). Educational expenditures and educational measures. Paper presented at the Annual meeting of the Southern Agricultural Economics Association. (ERIC Document Reproduction Service No. ED 255 342).
- Martineau, J.A., Somers, J.G., & Dehlin, J.O. (2001) Predictors of growth on a standardized achievement test: implications for evaluating instructional effectiveness. Paper presented at the annual meeting of the Michigan Association of School Psychologists, Lansing MI, October
- Messick, S. (1989). Validity. In R. L. Linn, *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.
- Michigan Department of Treasury. (2000). State of Michigan Merit Award Program: Design and validity of the MEAP test. Lansing, MI: Author.
- Moore, W. P. (1994). The devaluation of standardized testing: One district's response to a mandated assessment. *Applied Measurement in Education*, 7, 343-367.

Myers, M.A. & Curtiss, D. (2003). Failing the equity test. Principal Leadership, October.

National Association of School Psychologist (1998). Position statement: Student grade retention and social promotion. Bethesda, MD: NASP. www.nasponline.com

National Association of School Psychologist (2003). Position statement on using large scale assessment for high stakes decisions. Bethesda, MD: NASP. www.nasponline.com

Saginaw Public Schools. (1993). A correlational study into the relationship between grade point averages, Michigan Educational Assessment Program scores, and student absences for tenth grade Arthur Hill and Saginaw High School students, 1992-1993. Evaluation Report, Saginaw Public Schools, Saginaw, MI. (ERIC Document Reproduction Service No. 360 432).

Salvia, J., & Ysseldyke, J. E. (2004). Assessment (9th ed.). Boston: Houghton Mifflin Company.

Smith, M. L., & Fey, P. (2000). Validity and accountability in high-stakes testing. Journal of Teacher Education, 51, 334-344.

Van Moorlehem, T., & Newman, H. (1998, January 20) Testing MEAP scores. Detroit Free Press. Available online: http://www.freep.com/news/meap/main_test.htm

Addendum

Here are common arguments for and against High Stakes Tests taken from the references cited. It is used as a tool to organize the reader to the complexity of the issue.

Topic area	Argument supporting High Stakes tests	Argument against High Stakes Tests	Michigan Situation
Goal setting	Needed to know what is important to teach	Goals are already known	Michigan previously adopted the Curriculum Benchmarks
Accountability	Needed to quantify teacher's performance	Without adjusting for student background and school resource differences, measurement of teacher is invalid	Education YES! system adopted as grading system for student progress
Feelings of competence	Students scoring well will feel accomplished	Students scoring poorly feel further defeated	
Curriculum	Represents the curriculum well	Tends to narrow the curriculum, focusing on facts rather than problem-solving	MEAP is aligned with the Michigan Curriculum Benchmarks. Teachers report teaching facts more than reasoning.
Level playing field	Equalize the opportunity to display skills	Equal testing does not mean equal readiness when entering school, nor equal resources in the school	
More valid measure of learning than grades	Not affected by motivation, language, emotionality, social status.	Test IS affected by motivation, language, emotionality, and socioeconomic status.	Michigan research shows test scores have questionable validity because they are related to socioeconomic status

Topic area	Argument supporting High Stakes tests	Argument against High Stakes Tests	Michigan Situation
Instruction	Results will improve instruction, can be used diagnostically	Results will narrow instruction and focus upon test-taking skills rather than content. Scores not frequent or current enough to adjust instruction.	Since MEAP is not annually given in every grade, (not a repeated measure) it has limited instructional usefulness. Further, Michigan test results are not given in time to make instructional changes for the students that actually took the test. .
Professional development	Will be more pertinent to instruction	Will narrow the instructional focus	Although Michigan requires specified hours of professional development; it does not offer the programs. PD is locally controlled.
Parents	Parents will understand results better	Parents will be misled by scores. It should be known how passing scores were determined for the test.	Having only 40% of students initially pass the MEAP resulted in widespread public perception of school failure
Incentives for students (awards, recognition)	Incentives for higher scores will improve performance of students	Incentives for their advantaged peers will discourage lower-scoring students	\$500.00 to \$2000 Michigan Merit Award which increased participation.

Topic area	Argument supporting High Stakes tests	Argument against High Stakes Tests	Michigan Situation
Sanctions for students (retention, not graduating)	Sanctions will make students work harder	Sanctions will discourage students; single test score should never be used without other data in such a decision	State-level individual sanctions for students based on the MEAP have not been established in Michigan to date
Incentives for teachers (bonuses, recognition)	Teacher performance will improve with incentives	Teachers will resent incentives if biased by student background	
Sanctions for teachers (no salary increase, denied tenure, dismissal)	Teacher performance will improve if threatened with sanctions	Teachers will move to suburban schools where sanctions are rare	
Incentives for schools (recognition, grants, better resources)	Will improve school performance	Incentives do little to change the rank-ordering of schools by test score	
Sanctions for schools (students may transfer, school is closed, taken over, administrator replaced,)	Will improve lower performing schools	Low scores subject the school to public ridicule, sanctions do not improve the school for those students remaining there	A criticism of the Education YES! Accountability System
Generalized learning	Learning should transfer (generalize) to other areas	Improved scores on high-stakes tests are not found on similar tests (NAEP, ACT), raising question about whether true improvement has occurred	

Topic area	Argument supporting High Stakes tests	Argument against High Stakes Tests	Michigan Situation
Competitiveness	Will result in better more efficient schools	Assuming equivalent readiness of entering students, schools cannot be competitive with one another until they have equal access to resources such as up to date buildings, equipment, and the ability to procure talented personnel by offering them incentives to join their staff	