

Introduction

A major policy of the U.S. Department of Education (USED) during the past two years contained in the Race to the Top (RTTT) competition for Federal funding and elsewhere is that states agree to use their statewide assessments designated for school accountability purposes under the No Child Left Behind (NCLB) act as part of the evaluation of classroom teachers. Although Michigan has twice applied for such funding (and each time has not been awarded such a grant), this provision is still in place in Michigan, due to the state's adoption of a package of "reform" measures designed to enhance the state's chances of receiving RTTT funding.

Legislative Requirements

The state reform law (Public Act 205 of 2009) calls for, among other things, local districts to use the achievement of students, especially their *growth* on such measures, to evaluate the effectiveness of teachers. Specifically, Sections 1249 and 1250 of P.A. 205 state:

Sec.1249. With the involvement of teachers and school administrators, the board of a school district or intermediate school district or board of directors of a public school academy shall adopt and implement for all teachers and school administrators a rigorous, transparent, and fair performance evaluation system that does all of the following:

- a) Evaluates the teacher's or school administrator's job performance at least annually while providing timely and constructive feedback.
- b) Establishes clear approaches to measuring student growth and provides teachers and school administrators with relevant data on student growth.
- c) Evaluates a teacher's or school administrator's job performance, using multiple rating categories that take into account data on student growth as a significant factor. For these purposes, student growth shall be measured by national, state, or local assessments and other objective criteria.
- d) Uses the evaluations, at a minimum, to inform decisions regarding all of the following:
 - (i) The effectiveness of teachers and school administrators, ensuring that they are given ample opportunities for improvement.

"The state reform law calls for, among other things, local districts to use the achievement of students, especially their growth on such measures, to evaluate the effectiveness of teachers."

- (ii) Promotion, retention, and development of teachers and school administrators, including providing relevant coaching, instruction support, or professional development.
- (iii) Whether to grant tenure or full certification, or both, to teachers and school administrators using rigorous standards and streamlined, transparent, and fair procedures.
- (iv) Removing ineffective tenured and untenured teachers and school administrators after they have had ample opportunities to improve, and ensuring that these decisions are made using rigorous standards and streamlined, transparent, and fair procedures.

Sec. 1250.

- 1) A school district, public school academy, or intermediate school district shall implement and maintain a method of compensation for its teachers and school administrators that includes job performance and job accomplishments as a significant factor in determining compensation and additional compensation. The assessment of job performance shall incorporate a rigorous, transparent, and fair evaluation system that evaluates a teacher's or school administrator's performance at least in part based upon data on student growth as measured by assessments and other objective criteria.
- 2) If a collective bargaining agreement is in effect for teachers or school administrators of a school district, public school academy, or intermediate school district as of the effective date of the amendatory act that added this subsection, and if that collective bargaining agreement prevents compliance with subsection (1), then subsection (1) does not apply to that school district, public school academy, or intermediate school district until after the expiration of that collective bargaining agreement. (Public Act 205 of 2009)

The purpose of evaluating teachers in this manner is to attempt to improve the effectiveness of teachers (and school leaders), retain effective educators, and serve as a mechanism to remove those educators, non-tenured and tenured who are viewed as ineffective, even given ample opportunities to improve. In addition, unless prohibited by current contracts, P.A. 205, Section 1250, also sets in place a means of compensating educators based on their "job performance and job accomplishments," based in part "upon data on student growth."

"The purpose of evaluating teachers in this manner is to attempt to improve the effectiveness of teachers (and school leaders)."

Using Assessment to Evaluate Educators

Therefore, a key issue before the Michigan Department of Education, Michigan teachers and school leaders, and the professional organizations that represent them, is how the Michigan Department of Education can most effectively develop and implement a program that uses the state's large-scale assessments and other measures to evaluate the state's classroom teachers and school leaders. A secondary issue is how educators can become involved in helping to assure that this program is implemented in the most constructive manner possible.

“For groups of students, this can lead to cross-sectional comparison - that is, how did fourth graders do this year in comparison to fourth graders last year?”

Two types of achievement information might be available for use in teacher evaluation. The first is status (S) data. This sort of information shows what level of performance students have achieved in a content area at the time of the assessment. Thus, when a fourth grader is tested on MEAP, or an eleventh grader on MME, we know what their level of accomplishment was on the day of testing. We do not know, of course, what led up to that level of achievement, whether good or not so good.

For groups of students, this can lead to cross-sectional comparisons - that is, how did fourth graders do this year in comparison to fourth graders last year? Have more eleventh graders done well on the MME this year than the eleventh graders did last year? In each case, *different sets of students* - in fourth or in eleventh grade - are being compared. To imply that increases in performance from one year to another in this situation is due to teacher or school leader effectiveness could be misleading, especially if the type(s) of students tested change from year to year. This is something that can easily occur in small and not-so-small schools. Thus, cross-sectional comparisons to measure educator effectiveness may likely be inconclusive and potentially inaccurate.

The tracking of change in performance of the *same* students across grade levels became feasible as a result of the NCLB requirement for grade 3-8 testing. This means that it is possible to track how a student does in third grade, and connect to their performance in fourth grade and so forth through eighth grade (using the state's Unique Identification Codes or UICs). This presumes that the state assesses students in a roughly comparable manner each year. If the same assessment is given twice, once at the outset of instruction and the second time at the conclusion of it, we might be able to attribute changes (or lack thereof) to the actions of the teacher(s). In this case, each student is serving as his or her own control. In this type of measurement model, we are looking at the growth (G) in student learning. For groups of students, this can lead to longitudinal comparisons - that is, how did last year's third graders do in comparison to their performance this year as fourth graders?

The implications of this are that in designing a system to use assessments to evaluate teachers, we should strive to look for how growth models could be used. This is because with status models, too many things can be different between last year's third graders and those enrolled this year, and thus may make comparisons too weak to be useful (much less legally-defensible). The enrollment boundaries may have changed, school consolidation or school openings may have changed the demographics of those attending, program changes might have occurred, and, of course, students are different from one another. These year-to-year differences are the greatest in small group sizes that especially typify many elementary schools.

As desirable as it is to use *growth* data, however, there are some significant challenges in using tests for this purpose. First, the assessment needs to remain the same from grade-to-grade. Some tests are specifically built with enough overlapping content from one grade to another to make sure that the tests used in adjacent grades

are about the same. Others may not be constructed in this manner, for good reasons. This latter type of test design will facilitate growth comparisons, but may not be approvable under NCLB Standards and Assessments Peer Review requirements (since those requirements dictate that the tests used at each grade must be fully aligned with state's academic content standards for those grades).

Second, to measure growth, the assessment would need to be administered at least twice - either within the same year or at the same time each year. This could increase both testing burden and costs, and may take valuable instructional time as well.

Third, while there are a variety of statistical models to calculate growth, most of these are inaccessible to educators and anyone else without advanced psychometric training. Thus, educators and the public have little or no idea about the factors that result in good growth scores from the use of such models. This could be frustrating to educators hoping to understand what they need to do to help improve the growth in performance of their students.

Yet it is the growth measure that measurement specialists look to use to measure the effectiveness of educators. There is rich and growing literature about different methods and statistical models for estimating growth. These include the use of value tables, value-added testing models, and other ways of indicating how much growth occurred in a classroom or a school. It is not the purpose of this paper to describe or critique such methods, but suffice it to say that these models require two data points to work, whether these are from annual testing used at adjacent grades or pre-post testing done in the same school year. Without these two data points, growth modeling is not possible.

Therefore, this paper will review the current assessment options available for teacher evaluation, pointing out the strengths and challenges in using each type of measure for this purpose. In addition, it will present a broader, somewhat different approach to teacher evaluation, still based in part on student achievement results. The goal of this part of the paper is to suggest more constructive ways in which this state mandate and Federal policy can be carried out more productively.

Assessment Options

There are several options available to be used for the achievement test component of this reform initiative. These include the following:

- Statewide assessment programs required by state or Federal law
- Standardized achievement tests that local districts might choose to use
- Common assessments from commercially-available or built locally from item banks or locally-developed assessments
- Interim benchmark assessments, whether commercially-available or locally-developed
- Classroom assessments developed and used by individual classroom teacher

“..to measure growth, the assessment would need to be administered twice - either within the same year or at the same time each year.”

Each of these assessment options is described more fully below, along with how they might be used in teacher evaluation, and the advantages and challenges of each in the context of educator evaluation.

Statewide Assessments - Michigan currently has several statewide assessment programs intended for its students. These are the:

- *Michigan Educational Assessment Program (MEAP)* - This statewide assessment program assesses all third through eighth graders in Mathematics and Reading, fourth and seventh graders in Writing, fifth and eighth graders in Science, and sixth and ninth graders in Social Studies.
- *Michigan Merit Examination (MME)* - This assessment program is based on the ACT college-entrance test, the WorkKeys work assessment, and various Michigan-developed components. Students receive ACT, WorkKeys (and are eligible for national skills certification), and MME score reports.
- *MI-Access* - This assessment is given to all students with disabilities unable to participate in the MEAP and MME, even with accommodations, in all of the grades and content areas assessed by MEAP and MME. This program covers the same content standards as the general education assessment programs, but uses alternate achievement standards to report the achievement of these students with severe disabilities.
- *English Language Proficiency Assessment (ELPA)* - Annually while enrolled considered to be an English learner (EL), as well as for two years after no longer receiving language assistance, each EL student must be assessed on a measure of English proficiency. This is defined as reading, writing, speaking, and listening, as well as overall comprehension. The ELPA is Michigan's assessment system for ELs. It includes a screener assessment given in the fall, and a complete assessment battery administered in the spring. ELPA assessments are given K-12. In addition, these students participate in the regular state assessments (MEAP and MME) in each content area, and starting in their third year enrolled in the U.S., must be assessed in English. Prior to this, students who speak Arabic or Spanish can take the MEAP or MME tests administered in Spanish.

“The MEAP, MI-Access, and ELPA assessment programs involve students at different grade levels participating in the program, using tests that are equated from year to year.”

Uses - The MEAP, MI-Access, and ELPA assessment programs involve students at different grade levels participating in the program, using tests that are equated from year to year. The MME is also a test that is equated from year to year, but is administered in only one grade level. While all four programs can provide *status* information at some grades, only the MEAP, MI-Access, and ELPA can produce *growth* data (as these terms were defined above). However, none of these assessment programs covers all grades or all content areas, as the table given on the next page well illustrates.

Grade	Mathematics	ELA/Rdg	Science	SS	Health/PE	Art	Music
K							
1							
2							
3	S	S					
4	S, G	S, G					
5	S, G	S, G	S				
6	S, G	S, G		S			
7	S, G	S, G					
8	S, G	S, G	S				
9				S			
10							
11	S	S	S	S			
12							

S = Status data G = Growth data

Advantages - The definite advantage of these state-administered programs is that they provide data to local districts at no added cost to the districts. The MEAP, MME, and MI-Access can provide useful data for school improvement, and often serve as a basis for the achievement outcomes referenced within school improvement plans required and developed by schools and districts. The MEAP and MI-Access can provide growth data in two content areas at five grade levels (this is true of MI-Access presuming that students are assessed each year with MI-Access, rather than switching back and forth between MI-Access and MEAP).

Challenges - The most substantial issue with these assessments, especially if they were to be used to gauge student growth in learning, is that they do not cover all grades or subject areas. The table above shows the grades for which growth data would be available. The table shows that there are many more grades and content areas for which growth data are not available. Even when “status” data is added, there are still a number of areas where no state-developed information is provided. Only a small number of educators could be held accountable via state tests, and only in the elementary and middle school grades. The percentage may well be below 50% of the educators in an elementary school, and virtually no educators at the high school level. Only the middle school staff is relatively well covered, but only in mathematics and English, not the several other content areas.

Thus, any state-developed system that relies on state-developed and administered assessments would have many more areas to somehow “fill in” than could rely in the state information.

In addition, at the current time, MEAP and MI-Access results are not tied by the state to the teacher that the student had last year or this year for mathematics and English. Thus, while growth data could be available for some classrooms for the MEAP and MI-Access, the implementation of the coding necessary to carry this out is at least a year away.

“The definite advantage of these state-administered programs is that they provide data to local districts at no added cost to the districts.”

“Another drawback with these state-run assessment programs is that the data they provide is not particularly useful for individual student assistance because districts receive the data much too late to assist teachers.”

Another drawback with these state-run assessment programs is that the data they provide is not particularly useful for individual student assistance because districts receive the data much too late to assist teachers. Thus, teachers may find it difficult to help students learn the materials on which they are being held accountable.

A final serious but far more subtle drawback of using the MEAP or MI-Access for showing growth is that the instruments are built to measure well what is taught and should be learned in only a *single* grade level. Thus, the use of them to make grade- to-grade comparisons is problematic because of the differences (sometimes significant in nature) in what the state chooses to assess in the same content area at adjacent grades. Thus, the MEAP tests, which appear to be the most fruitful of the existing state exams to use for growth purposes suffers this issue of assessing different skills from grade to grade, so much so that its use for growth calculations may be challenging. While this is not as bad as “an apples-to-oranges” comparison, it may more akin to comparing “McIntosh versus delicious apples.” This is a violation of the assumptions behind some statistical growth models, the result of which led the state to adopt in 2007 a “progress” model for reporting grade-to-grade changes in student performance at the student, school, and district levels.

Standardized Achievement Tests - There are a number of products that are available on the market that could be used by local school districts to provide information that could be used for measuring growth in student achievement. The most comprehensive of these assessment batteries measure students from kindergarten through grade 12, in several content areas (including mathematics, reading, English, science, and social studies). Although high school tests are available, the quality of the norms is not as good at the high school level than at the elementary and middle school levels. All of these are norm-referenced - that is, they report the status of students in comparison to the performance of a norm group. Thus, they are useful for reporting the *relative* performance of students, using metrics such as percentile rank, grade-equivalency, stanines, normal curve equivalents (NCEs), and so forth. Most can also be reported in criterion-referenced manner as well, relative to the standards assessed by the tests.

The primary reason why these tests can show growth is that tests are often based on a multi-grade test blueprint. Thus, a fourth grade norm-referenced achievement test will have a substantial part of it based on material typically taught in fourth grade, but will also have a little material typically taught in second and sixth grades, and a bit more taught in third and fifth grades. The design of the fifth grade test will center on fifth grade, but include material typically taught in third through seventh grades. Because of the overlap in content, grade-to-grade comparisons are more statistically sound.

There are several basic types of norm-referenced tests available for purchase by school systems. These include:

- *Comprehensive achievement test batteries* - These instruments include individual sub-tests in multiple grades and content areas.

As mentioned above, these typically cover every grade from pre-K through grade 12 in all or most all of the content areas taught at the elementary, middle school, and high school levels. The scores from these batteries can be compared with one or more types of norm groups - e.g., a national norm group, public school or private school norm groups, and an urban norms group. There are several commercial batteries available:

- Iowa Test of Basic Skills/Iowa Tests of Educational Development (Riverside Publishing)
 - Comprehensive Test of Basic Skills (CTB/McGraw-Hill)
 - California Achievement Test (CTB/McGraw-Hill)
 - Terra Nova (CTB/McGraw-Hill)
 - Metropolitan Achievement Test (Pearson Assessment)
 - Stanford Achievement Test (Pearson Assessment)
- *Single-subject achievement examinations*, typically available in reading or mathematics - Single subject exams are used most often at the elementary level for diagnostic purposes, when additional information is needed to determine potential causes of learning issues on the part of some students. Although these are group administered tests and could be administered two or more times to all students in a grade level to gauge growth, they are not the sort of screening assessments that would most efficiently assess student achievement since achievement in one content area is not sufficient to judge successful performance in elementary school. Hence, it is recommended that such assessments not be used on a wide-scale basis. They might be appropriate for use by teachers who wish to demonstrate the effectiveness of remedial instruction with a handful of students.
 - *End-of-course tests* for various high school courses in subjects such as English, mathematics, science, and social studies - Increasingly, accountability for the performance of high school students, whether for student accountability (i.e., graduation) or school accountability, has turned to the use of end-of-course examinations. Fewer states are using graduation tests or general assessments of all students at a single grade level, and are substituting these exams for them. The advantage of them is that only the students who take the course take the exam, so that there should not be much differences in opportunity to learn (of course, in reality, there are vast differences in the actual course content delivered to students, which is why the exams are so useful for school accountability purposes - to assure that students are taught the content that goes with the course title.)

These assessments could be used at the high school level (where the comprehensive achievement test batteries are the weakest). To do so would require a different design than how typically used at the current time. Presently, these exams are administered only at the end of the two semesters of the course; to be used for teacher evaluation purposes, an alternate form of the assessment would need to be administered early in the fall, before the course begins, so that the *gains* in achievement can be determined from the end-of-course

“These assessment could be used at the high school level (where the comprehensive achievement test batteries are the weakest).”

assessment that would serve as the post-test. Currently, these exams are not set up to do this, although since multiple forms are typically available, it would not be difficult to set up such a system.

Providers of these sorts of tests include the College Board, ACT, Achieve, and others.

- *College entrance tests* are administered in the middle school and high school to help colleges to determine which students are likely to succeed in college and to aid students in preparing for post-secondary educational opportunities. Each of the two major publishers (ACT and the College Board) has instruments that are not only used specifically for college admissions (the ACT and the SAT), but also publish “preliminary” versions of these tests for use in earlier grades. For ACT, these earlier tests are the EXPLORE and PLAN, while for the SAT, it is the PSAT. Because these earlier tests are statistically linked to their respective college entrance test so performance on the earlier test and the college entrance test could be used as a measure of growth if these tests were used in an appropriate fashion (i.e., at the same time such as spring one grade level apart). For example, many school districts administer the PLAN assessment one year in advance of the MME, which contains the ACT. A growth score could be computed for students based on the change on the ACT score scale from tenth to eleventh grades.
- *Specialized high school examinations* - There are also several types of specialized high school tests that schools can elect to use. These include:
 - Advanced Placement (AP) Tests
 - Armed Services Vocational Aptitude Battery (ASVAP)
 - WorkKeys Tests - these are work readiness assessments provided by ACT, three of which are included within the Michigan Merit Examination. Other tests are also available and these may be used at other times that eleventh grade
 - Career Tech Education required assessments of program completers - These assessments are required of all high school students who finish a CTE program in one of 16 career pathways. Only these students are to be assessed. At the current time, not all of these assessments are available. They are administered only once - at the end of the high school program.

Each of these specialized high school examinations have limited use in the evaluation of high school teachers, since they are single-use assessments, most often given at the completion of a high school course of study.

Advantages - A primary advantage of the comprehensive achievement test batteries is their complete coverage of all grades and almost all content areas. More than one form of each is available. Because these tests are designed to measure student performance from grade to grade, the results of the previous year can serve as the “pretest” and this year’s results, from testing in late spring, can serve as the “posttest.” These batteries are also somewhat aligned to state standards, so that the need for augmentation to assure complete coverage is moderate (a consideration only if the state is

“College entrance tests are administered in the middle school and high school to help colleges to determine which students are likely to succeed in college and to aid students in preparing for post-secondary educational opportunities.”

using the achievement test battery as its NCLB-compliant testing program). By adopting a single battery or related batteries, the school system could measure the growth of all students in almost every content area.

The use of end-of-course exams at the high school level would help to focus teachers on providing a common set of learning opportunities to students. If commercially available end-of-course exams are used, these would help assure that students were exposed to challenging content and that students who did well on these exams were fully prepared to succeed in subsequent course work in the content area, whether in high school or in college (two- or four-year college). Not only could these tests serve to evaluate classroom teachers, they could also be used to assure equal learning opportunities and that student grades signify true accomplishment.

The primary advantage of the college entrance tests used to compute growth is that they are already being administered in many school systems. Since all schools already give the MME with the ACT in it, many also provide the PLAN assessment for their students, as a warm-up for the ACT. Hence, the comparative data is already available in many schools. A number of students also take the AP exams.

Challenges - One of the largest challenges in using these commercially available instruments for teacher evaluation is that they would cost local districts \$20 per student or more to use them. Because so much state-mandated assessment is already occurring, many local school systems have eliminated the use of standardized testing as a redundant and unnecessary expense.

Besides the costs, there is also the issue of redundancy of testing, since these tests would likely be used in addition to the statewide assessments. Time would need to be provided to test students twice in grades 3-8 and perhaps high school as well. Thus, reinstating this testing would reintroduce added testing, added costs, and less time for instruction.

For end-of-course tests, the greatest challenge is that there is no one system of assessment in place. The year after the high school graduation legislation was adopted; the Michigan Senate initiated a revision in the legislation that eliminated the requirement that the Michigan Department of Education develop end-of-course exams in the seventeen courses required for graduation. As a result, there is no set of course exams that have been developed by the state to measure the High School Content Expectations (HSCEs). Commercial products are built to measure common expectations among a number of users, yet the HSCEs for some courses are quite rigorous. For example, in 2007, Michigan examined the end-of-course exam produced by a neighboring state in the area of Algebra I to see if it could be used in Michigan. The result was that this instrument was found to measure only about 30% of the Michigan Algebra I content expectations. When combined with that state's Algebra II exam, only about 60% of the Michigan Algebra I HSCEs were measured. Thus, one cannot safely conclude that all end-of-course exams are created equal.

A second challenge with using end-of-course exams to evaluate teachers is that the exams would have to be administered at the start of the first semester and at the conclusion of the second semester (or, if

“One of the largest challenges in using these commercially available instruments for teacher evaluation is that they would cost local districts \$20 per student or more to use them.”

“A deeper, subtler issue in the use of any of these tests is whether or not these test measure the state’s academic content standards.”

students are likely to change teachers for the second semester of a course, they would need to be assessed at the start and the end of each semester). This implies more forms of the assessment and more testing than simply an end-of-course test.

For college entrance tests, the greatest issues are that the tests measure only two or three subject areas (mathematics and English/reading and writing) and only at one grade (grade 10 or grade 11, depending on how the PLAN assessment is used).

Since the MME is administered in March, if the PLAN is given at the start of tenth grade, the combination of the PLAN and ACT could be used to evaluate tenth grade teachers, but if the PLAN was given in the spring, it would be more useful for eleventh grade teacher evaluation. Of course, in either case, the time span will be either a grade and a half or two-thirds of one grade, neither ideal for teacher evaluation purposes. Hence, this approach to teacher evaluation is definitely not recommended.

A deeper, subtler issue in the use of any of these tests is whether or not these tests measure the state’s academic content standards. To the extent that they do not, or do not do so fully, teachers will be faced with the dilemma that they need to focus their instruction on the state standards so students do well on the statewide assessments, and then turn their attention to the skills assessed by these commercially available tests. What may result is an incoherent approach to instruction that ends up hurting students’ learning. This has been the case in the past, which was one of the reasons why the more comprehensive approach to school accountability called for in the No Child Left Behind law (with assessment annually in grades 3-8) was implemented. To use another set of tests, measuring somewhat different sets of skills at the same time, is a significant step backwards.

Common Assessments - A number of local school districts, alone and with the assistance of intermediate school districts/regional service agencies (ISDs/RESAs) have built or are building common assessments. In other instances, individual ISDs or RESAs have also created “common assessments.” This has occurred primarily as a result of the high school graduation legislation that requires school districts to use assessments as part of the determination of whether students receive credit for the courses that they have taken in high school. Because the requirement that the Michigan Department of Education create end-of-course assessments was removed from the graduation legislation, the resources for the state to create such assessments were never appropriated. However, the requirement that local school districts ‘use assessment as part of the determination of whether students receive credit’ in required courses was not removed. This means that the state is unable to meet the need for end-of-course examinations, and because of the lack of commercial products in the past, local school systems and ISDs/RESAs began building their own instruments.

There are two primary types of common assessments that local school systems have built:

- *End-of-course examinations* used for secondary courses at the middle school or high school levels.

- *Grade-level examinations* used in elementary or middle school levels to measure the annual performance of students in core subjects such as mathematics, reading, science and social studies.

Advantages - The advantages of these instruments is that they focus on the accomplishment of students in a particular grade or content area. If administered twice - at the start of a class and at the conclusion of it - this pre-post data can focus on what students have learned in the interim. Using this sort of design would make it clearer in which class the learning occurred, and thus would make it easier to determine which teacher could be held accountable (given the shortcomings of using a single measure of teacher performance for accountability purposes described later).

While end-of-grade assessments in the elementary or middle school are less common than end-of-course assessments at the high school level, they do exist. Often, these are used as part of decision making about grading students or whether students have learned enough to be promoted to the next grade level. Thus, these tests tend to have lower stakes for students than the end-of-course tests used at the high school level.

Challenges - As mentioned in the previous section above, it may be necessary to use pre- and post-testing in each semester of a two-semester course, thus increasing the amount of testing four-fold over simple end-of-course testing. This will require extra forms to be developed and will greatly increase testing costs. Few districts have built multiple forms of their common end-of-course assessments.

In addition, because many of the common assessments written in Michigan at the current time are “home grown,” many to most of them have not been developed in accordance with generally acceptable standards in the testing industry. For example, the APA/AERA/NCME Standards for Educational and Psychological Testing (1999) indicate that tests should be field-tested at least once before they are used, should be carefully reviewed for bias, and any high stakes decisions to be made should be based on multiple sources of information.

Since the initial reason for developing these assessments is to use them as part of the determination of whether students receive credit in the course (arguably, a high stakes decision for the student), these tests should meet at least the three standards noted above. Unfortunately, most common assessments built by local school districts or ISDs/RESAs in Michigan have *not* been field tested (either before or even after their initial use with students), the items have not been scrutinized for bias, using either review panels, statistical bias detection methods or both, and in some cases, are used to determine whether or not students get credit for a high school class regardless of the grade that they earned in the class. Each of these practices is inappropriate and could lead to litigation against school systems.

If these tests suffer from the problems noted above, their use in evaluating teachers would be even more problematic. Evaluating educators based on faulty instruments or basing their compensation on such assessments would be challenged legally. If challenged legally, it would be hard to

“White end-of-grade assessments in the elementary or middle school are less common than end-of-course assessments at the high school level, they do exist.”

defend the use of an instrument that has not been properly developed, much less the information such an instrument as the sole measure of a teacher's performance.

Interim Benchmark Assessments - These are assessments, based on school or state curricula, and that are used for several purposes, depending on the design of the interim assessments. As originally developed in some of the nation's urban districts, these were quarterly tests, based on local curricula or pacing guides that were used to predict which students would do well and poorly on the state assessment tests, so that educators could intervene with students predicted to do poorly. These tests tend to be administered quarterly.

More recently, a different version of the interim assessment design has emerged. This is to determine the instructional units that will occur in a course, such as Biology, and then build a unit exam for this instructional unit. Thus, rather than testing students every marking period, these exams are given at the conclusion of an instructional unit. Because there may be ten, twelve or more instructional units in a two-semester course (such as Biology); there will be a corresponding number of interim assessments.

In Michigan, such interim assessments are available in a couple of ways. These are:

- *Michigan-developed interim assessments* built as common assessments by local school districts or ISDs/RESAs. The Northern Michigan Assessment Consortium has used selected items from the State of Michigan, the Math- Science Partnership, and items created locally, to build interim benchmark assessments in high school courses in mathematics, science, and other content areas. These are available (through the Data Director software) for use by districts within northern Michigan and perhaps elsewhere. Thus, at the conclusion of each unit of study, a teacher could have students take the appropriate end-of-unit interim assessment (either on paper or online) and determine how much they have learned in that unit of study. Such a system would clearly show what students have or have not learned.
- *Acuity*, a commercially available interim benchmark testing system from CTB/McGraw Hill. This is a product that not only provides interim benchmark assessments when school systems want to administer the tests, it also has information on instructional resources that can be used by teachers (and by parents) to instruct students in the first place. This system is available in mathematics, reading and other areas and the testing is available both electronically and on paper. Testing can occur when scheduled by the district, which could occur at the conclusion of major units of study or periodically throughout the school year. Acuity is marketed widely as a "formative assessment," even though such instruments are actually interim benchmark assessments.
- *Learnia*, a commercially available interim benchmark testing

"While end-of-grade assessments in the elementary or middle school are less common than end-of-course assessments at the high school level, they do exist."

system from Pearson Assessment. Learnia is similar to Acuity and provides similar resources. In Michigan, Learnia has not been marketed as widely as Acuity, partly because it was a Harcourt Assessment product acquired by Pearson when it purchased the assets of Harcourt Assessment.

Advantages - For students, the advantages of interim assessments is that they provide “early warnings” about challenges in learning. Instead of waiting until the conclusion of two semesters of a course, failing a test, and being forced into a “credit recovery” situation, the interim tests provide unit by unit information on achievement that students can use to make sure that they are learning as they go along, or if not, can be re-taught the material missed before it becomes a serious achievement issue. For teachers, these assessments could provide a ready source of information on their effectiveness, since there could be ten, twelve or more occasions on which student achievement is assessed. These mini-summative assessments act like an overall end of course examination, except that they cover far less material. Thus, they are most sensitive to the instruction provided by the teacher.

Challenges - There are several challenges in using this sort of assessment. First, are the assessments actually aligned to classroom instruction? As with every assessment, one key to usefulness is whether the test actually measures well the appropriate instructional targets, the content standards. Especially with commercially available test systems, alignment may be an issue. With homegrown interim assessments, do they measure rigorous standards in a manner that adequately represents the rigor inherent in the standards?

Second, these interim benchmark assessments do not provide a pre-instruction and a post-instruction test, so that the *gains* in student learning are typically not calculated nor displayed. The presumption is that if all students do well on the unit exam that they learned a lot in the learning situation (or from the teacher). However, perhaps all of the students knew the material before being placed in the learning situation. There is no way to tell this from the post-instruction data provided by the measures. Finally, many of the homegrown interim assessments suffer from the same lack of adequate validation and review as the common assessments given above. Because these tend to be shorter tests than the typical common assessments, it is even more critical to field test and review the interim assessments since the impact of one or two poor items is much greater in a shorter test.

Classroom Assessments - These are assessments developed and used by individual classroom teacher. The two types of classroom assessments used by teachers include:

- *Formative assessment strategies and tools*, used by teachers as they teach, to determine in an on-going manner whether every student is learning what they are teaching. Strategies such as “question out the door” can help teachers determine whether students are learning the concepts and skills they are teaching, and if not, what misconceptions students have. This is especially helpful if the teacher has given thought to what will be done next if all of the students have learned what they have been teaching, if only about half of the students have learned the material or

“For students, the advantages of interim assessments is that they provide “early warnings” about challenges in learning.”

virtually none have. It is this thoughtful use of the information that determines the effectiveness of this approach.

- *Summative assessments*, which in the classroom context means that teachers are using traditional tests or non-traditional approaches to assessment to determine student learning at the end of units of instruction. The difference between these tests and those mentioned above under common assessment or interim benchmark assessments is that in this case, these are tests individually developed by teachers and used typically in just one classroom.

Advantages - One of the obvious advantages of teacher-made tests is that they are constructed by teachers based on their lesson plans for the units of study they are teaching. Thus, the alignment between the test and the instruction should be high. In addition, since teachers have planned the lesson and the test(s) to go with it, they should be available when needed and return results to students and to the teacher quickly.

Challenges - There are several challenges in teacher-made tests. First, such tests are notoriously of poor quality. Educators who have not learned about assessment or how to build a reliable and valid instrument construct them. Second, the items are rarely examined before they are used, nor afterwards before scores are used for grading. Third, each test is unique to the individual teacher, which makes the determination of the performance of a teacher very problematic. This is because the evaluator will not be able to judge the rigor of the assessment separately from its use in a single classroom. Finally, classroom teachers rarely use a pre-test/post-test model for their classroom tests, so that *change data* on the students in a classroom is rarely available. For all of these reasons, these tests developed by individual teachers would serve poorly in a system to evaluate them.

The use of formative assessment, as defined above, would also not serve as a useful tool for evaluation purposes, since the data is so idiosyncratic to the teacher, the lesson being taught, and the students in the classroom. To use this information for high stakes purposes (e.g., personnel evaluation) would undoubtedly lead to the distortion of the information.

Effective Personnel Evaluation

While the purpose of this paper is primarily to identify the issues in using tests for teacher evaluation, it would be incomplete if it ignored the subject of *how* the tests and other information would be used. Thus, this final section of the paper presents a review of some of the issues in using tests for teacher evaluation and concludes with a suggested model by which tests along with other information about instruction and learning could be used to evaluate and to compensate teachers.

“One of the obvious advantages of teacher-made tests is that they are constructed by teachers based on their lesson plans for the units of study they are teaching.”

Issues in Personnel Evaluation by Test - There are a number of issues inherent in the use of tests to measure the performance of teachers. Unless these are adequately dealt with, the entire teacher evaluation system could be jeopardized. The most serious issues include:

- *Adverse impact* - the use of a test to determine who is selected (i.e., to remain on the job or how to be compensated) is a high stakes employment decision. One aspect of testing that occurs in employment situations is to determine the adverse impact of any approach to teacher evaluation. In the context of this paper, this means that once instruments are selected for use, the differential impact of the use of these assessments needs to be determined. Specifically, developers of the evaluation system need to determine if there is any adverse impact from the use of the system on protected groups. If there is, the entire process needs to be reviewed to determine if such adverse consequences can be mitigated in some fashion. Such information would weigh negatively in any court cases brought because of the use of this system on teachers who were negatively impacted by it. None of the assessments listed in this paper have been examined for adverse impact here in Michigan
- *Making important decisions based on one source of data* - Because this review of various tests did not indicate one test currently being used (e.g., the state assessment tests) that is suitable for evaluation purposes, there will be a temptation, if a new testing system is installed specifically to measure teacher performance, to use one and only one source of information about students to evaluate teachers. As mentioned earlier, the AERA/APA/NCME professional standards (1999) indicate strongly that important decisions based to be made about students, educators or other should be based on more than one types of information.

In the context of teacher evaluation, it means more than just using two tests (although that would be better than relying on just one instrument). It means using, two (or more) types of information about student achievement, such test data and the observations of someone with adequate training in judging teacher competence. A system that relies on just one source of information be it test or other types of data, will be in jeopardy.

- *Multiple causes for good and poor achievement* - One of the most serious issues in using tests to evaluate teachers or others is that there are so many reasons and so many places that students learn (or don't learn). Even in the most controlled situations, the presumption that the learning or lack of learning shown by students is the sole responsibility of the teacher is not accurate. Thus, in high achievement situations, there are other forces at work. In low achievement situations, some of the same forces may be at work, but perhaps not so positively. Thus, educators whose students do well need to be modest in their claims about success, as educators whose students do poorly need not bear the sole

“A system that relies on just one source of information be it test or other types of data, will be in jeopardy.”

*“As much as we value
excellence in
teaching, we do not
have a uniform
definition of what it is.”*

responsibility for students’ failure to learn. Factors that can spell a difference include class size; instructional resources provided; parental interest and involvement; student attendance; student motivation; support staff presence or absence, among others.

- *Low quality of tests* - As this review has shown, few assessments are without issues and challenges to be used for teacher evaluation. Tests are not currently in place to implement the system. The resources to provide such instruments are not in place, either. Thus, the instruments that will be used are not ideal for evaluation purposes. This may make the implementation of evaluation systems very challenging. It should encourage those that do use the available tests to be especially cautious in what additional information is collected beyond the tests, since the assessment base for meeting these requirements is bound to be weak.
- *Nature of the students being taught* - Teachers are not always assigned students randomly. Nor are students randomly assigned to the classes they teach (although more experienced teachers may be able to select to teach the Advanced Placement and college preparatory classes, over the more basic courses). Thus, when it comes time to determine the performance of one teacher versus another, it may be an advantage for a teacher to be teaching a motivated group of college-bound students in high school than a general education class in a middle school. Part of whether this is an advantage or not is whether the evaluation system is criterion-referenced (every teacher *could be* an excellent teacher) or norm-referenced (the best teacher receives the greatest payoff from the system and the highest compensation. If it is norm-referenced, teachers will fight to be assigned to the “best students,” whether or not this is actually shown to advantage their evaluation.
- *Poor or multiple definitions of “good teaching”* - As much as we value excellence in teaching, we do not have a uniform definition of what it is. Perhaps it would be more accurate to say that we don’t have a single definition of good teaching - we have multiple types of what constitutes “good.” Hence, when a supervisor is evaluating a teacher, it is important that the definitions of good teaching are made explicit. There are times when a directive teacher is judged to be okay and times when non-direction by the teacher is okay, too. Some supervisors may value a quiet, orderly classroom, while others see such an environment as sterile of learning, and certainly, messiness may accompany certain types of learning - a chemistry laboratory, an arts room, or small group work in a social studies classroom. Other times, the messy, noisy classroom is a sign that students are not paying attention, aren’t on task, or are not focused on learning. Thus, it will be essential, as it is in any type of predictive study, to have a clear definition of the criterion - good teaching and good learning.
- *Poorly prepared judges of “good teaching”* - One of the reasons why there is such uncertainty about what is good teaching is that relatively few individuals have been trained to be able to

thoughtfully examine teaching and learning and to judge it adequately. The classroom observation visits required of new probationary teachers are often carried out by individuals who are untrained in classroom evaluation. They also may not observe teachers in a manner likely to yield reliable judgments of the teacher's effectiveness. One could say that these individuals wouldn't recognize good teaching if they saw it (which, unfortunately, they are asked to do on a regular basis). This suggests that no matter what system of teacher evaluation is used (and presuming that some form of observation will be included), training in the use of the system and carrying out tasks such as observation is essential.

- *Criterion-referenced or norm-referenced evaluation* - This issue is whether we are evaluating teachers against a standard of "good teaching" (see above) or with one another. The former is "criterion-referenced evaluation" while the latter is "norm referenced evaluation." In the former system, every teacher in a school or district could achieve excellence, or none of them might be judged in this manner. The downside of this approach is that the public or others, already skeptical about educators, might not believe that all teachers in a school or district are worthy of recognition and compensation for excellence, especially given different definitions of good teaching (see above). In addition, it is often difficult for supervisors to withhold pay increases from subordinates, especially ones that they like and work with on a daily basis.

In the norm-referenced situation, the best (and the worst) teacher would be selected in each situation, be that a school or a district. Jack Welch, former CEO of General Electric, was known for dismissing the bottom 10% of his managers, regardless of how well the entire group did or did not do. This is an example of the norm-referenced approach to employee evaluation. The downside of this system is that it pits one employee against another, which would certainly diminish the likelihood that the employees would work well together in a team or work to solve common problems - one employees gain is another's potential loss. Thus, normative evaluation could well be counter-productive in a school setting.

- *Models Used to Judge Growth* - As mentioned earlier, the data that the tests may yield will might be combined and statistically analyzed to produce some sort of growth score. The issue with some such models is that they are so complex that only a few, highly-trained individuals understand them. If such a model is used in Michigan, it could serve to create more frustration than motivation, since educators may be held accountable for improving on an index score that they don't understand well enough to know how to do so.

"The classroom observation visits required of new probationary teachers are often carried out by individuals who are untrained in classroom evaluation."

A Model for Teacher Evaluation

As covered throughout this paper, there are serious issues in the use of

assessments to evaluate teachers and school leaders. The essential issues is how to create a system that serves to provide evaluative data first and foremost to help educators to understand where improvement is necessary, how to go about improving, and only if these efforts are unsuccessful, helping the educator to find other employment.

The following is one model for the outline of a teacher evaluation system that might meet these criteria. It presumes that teacher evaluation will be primarily formative in nature, especially in the early years, so that the emphasis is on helping educators improve their practice, rather than deciding “winners and losers.” Realistically, it could be used, after a period of time, to provide summative judgments as well, but those should be provided in advance (for example: ‘unless I see improvement in ... by next year, I will have no choice but to...’), so that educators have chances to improve before summary judgments are made or carried out.

This model presumes, in response to the issues raised above, that a variety of achievement measures, some to be developed, are used. In addition, it bases the evaluation on the school’s school improvement plan (SIP). Further, it assumes that every educator should be striving to improve and that these areas should be identified both by the educator and the supervisor of the educator.

A Model for Teacher Evaluation

Professional Practices Portfolio

In the Fall, Develop:

1. Goals for the Individual Educator
 - A. Goals from the School Improvement Plan - the educator’s role in achieving one or more the SIP goals
 - B. Goals for the Individual Educator - the educator’s goal(s)
 - Short-term - this school year
 - Long-term - next year and beyond
2. Measures of Performance
 - A. State measures where available and applicable
 - MEAP/MME/MI-Access/ELPA
 - B. School measures
 - School’s comprehensive needs assessment
 - Interim benchmark assessments
 - Common assessments
 - A. Educator-created measures
 - Content organization measures
 - Individually-collected data
 - Summative information
 - Interim benchmark assessments

“This is one model for the outline of a teacher evaluation system ...”

- Formative assessment information
3. Plans for Growth and Improvement
 - A. Plans to help accomplish team/school goals - how will the educator accomplish the school goals within the context of the school improvement team?
 - B. Plans to accomplish individual goals - how will the individual teacher accomplish his or her goals?

In the Spring, Add:

4. Summary of Activities Used to Accomplish the Plans and Goals
 - A. Individual educator achievement of team/school goals - how did the educator help to accomplish the goals of the school improvement team?
 - B. Individual educator goals - how did the educator accomplish the goals he or she set for himself or herself?
5. Evidence of Accomplishment
 - A. Team goals - what evidence is there that the selected goal(s) in the school improvement plan were accomplished?
 - B. Individual goals - what evidence is there that the individual goals were accomplished?
 - Educator-collected information
 - Peer information
 - Supervisor(s) information
6. Reflective Feedback
 - A. Individual educator - looking back on the year, what would the educator have done differently? What does the educator plan for the coming year?
 - B. Peers on the team/school - Do the peers of the educator support the evidence of accomplishment as put forth by the individual educator?
 - C. Supervisor(s) - Does the supervisor support the evidence of accomplishment as put forth by the individual educator?

This process portfolio would provide a means to set goals for improvement that are relevant to the educator and the school, to identify in advance what measures will be used to judge success (hence, no last minute surprises) and agree on a course of action for improvement (thus, giving the educator a say in how improvement will occur and the opportunity to learn). At the end of the year, the educator can indicate the steps taken to improve, whether the indicated opportunities were provided (or not) and other actions taken to seek improvement, the evidence of accomplishment mustered by the educator (and others), and then reflects on the year's accomplishments. In addition, peers and the supervisor can reflect on what the candidate was able to achieve during the year. It is this reflective piece that might be an indication to an educator that improvement has not been adequate and that unless further improvement is forthcoming, some type of action (such as dismissal) might take place. On the other hand, substantial improvement might be compensated

"This process portfolio would provide a means to set goals for improvement that are relevant to the educator and the school..."

This is just one model for how educator evaluation could occur in such a manner that continued employment and compensation systems could be tied to employee evaluation.

Summary

This paper has examined a number of types of achievement measures that could be used, as required by state legislation, to evaluate teachers and school leaders. There are two fundamental types of information on student achievement - status and growth (cross-sectional or longitudinal). Growth data is the most desirable for the purpose of evaluating educators, but as the paper showed, there is not an assessment program currently capable of providing this information for all educators. This means the addition of other forms of assessments and the administration of these one or more times throughout the school year. The paper suggested one possible model for using achievement tests within a process of evaluating educators, but since this was not the primary purpose of the paper, did not touch on this critical aspect in depth. It did raise a number of important issues in the use of assessments in employee evaluation, however. It is hoped that this paper will be helpful to those who set about creating the system to be used for teacher and school leader evaluation.

“There are two fundamental types of information on student achievement - status and growth...”

References

AERA/APA/NCME. Standards for educational and psychological testing. Washington, DC: Author. 1999