Accountability Movement

"Accountability" is a theme in public education that seems to be growing in importance. In fact, it seems as though everyone wants schools to be held accountable. Legislators want schools held accountable so that they can measure whether or not public schools are failing. Taxpayers want schools to be held accountable so that they can see that schools are effectively using limited resources. Parents want schools held accountable so that they will know where to send their children for the best education. Administrators and teachers want accountability in order to improve their schools. Every public educator now feels as though he or she will be held accountable for school performance.

Accountability? I Thought This Was About Assessment!

Don't worry – I'm getting there. You see, while most everyone agrees that schools need to be held accountable, we don't agree on how we should measure school performance. We don't even agree on what makes a school successful.

Back in the 1980s, school success was judged by the amount of resources a school spent on education. The belief was that schools with a large amount of resources were more successful than schools that had fewer resources. This led to the belief that the best way to improve schools was to give more money to the poorer schools. When research showed that increased spending on education did not correlate with increased academic performance, legislators quickly decided that schools should not be measured simply by the amount of money they spent.

It was back to the drawing board for legislators in the 1990s. Since money didn't seem to have a direct relationship with school success, we needed a new way to hold schools accountable. It was decided that schools should be judged by measuring student academic performance in key content areas (mainly reading, math, and science). This turn away from measuring money towards measuring academic achievement may have led to the current theory that successful schools should receive more funding while poorly performing schools should be punished with less funding.

And that's where we are now. We believe the best way to measure school performance is to hold schools accountable for their students' academic achievement. We believe that this type of accountability could lead to an accurate measure of school performance that can be compared across districts, states, nations, and time. This information could then be used to allocate resources more efficiently, improve student academic performance, and improve the image of the public school system.

You Still Haven't Gotten To Assessment…

This is where assessments step in. We acknowledge that measuring student academic achievement is important, but we need to know how to measure that achievement. Ideally, we would want an unbiased measure of student performance that would tell us exactly how successful our public schools are in educating our students. Assessments provide us with evidence indicating that level of success.

"Assessment" is the classification of someone or something with respect to its worth. We find the worth of an object and give it a label (such as "great," "A+," or "100"). Whether we realize it or not, we assess things all the time through formal and informal assessments. Informal assessments are little judgments we make everyday about the worth of a person or object. When we go shopping, we are constantly using informal assessments to judge the value (or worth) of products. We also use informal assessments when we talk with others. During a conversation, we look for nods or questionable looks to see if the other person understands the topic. Informal assessments are judgments we make without thinking too much about it.

Formal assessments, on the other hand, do not occur so frequently. A formal assessment usually involves something being written or recorded. Let's use that supermarket example. Suppose you wanted to buy the best box of cereal in the store. Before driving out to the grocery store, you would probably think of all the factors that make a cereal great – taste, cost, toy prizes inside, etc. You would then come up with a way to record those factors – like a checklist. Then you would go to the store, look carefully at each cereal, record the information, and make a judgment about which cereal is the best. This is a formal assessment

Now we see the difference between formal and informal assessments, but we still don't know what they would look like in an educational environment. How does a classroom teacher use informal and formal assessments? Examples of informal assessments used by teachers include observing students during a lesson and asking review questions. Examples of formal assessments include classroom tests, homework assignments, and standardized tests.

So formal assessments are better than informal assessments? Well, no - they're just different. Informal assessments are extremely useful as formative assessments. Formative assessments are little judgments we make during a lesson or a unit in order to see if students are learning the material. If a teacher makes an observation (informal assessment) that the students are having trouble with fractions, that teacher can go back and re-teach fractions. The observation (formative assessment) allowed the teacher to see the problem early enough to correct it.

A formal assessment, on the other hand, is more useful for summative purposes. Summative assessments tell us what the students learned after the lesson or unit has finished. A teacher who finishes a lesson on fractions can give a classroom test (formal assessment) to see how much students learned. The test (summative assessment) allowed the teacher to see the results of instruction.

So What Assessments Are We Using?

Teachers and administrators in Clinton are using all types of assessments. Teachers use observations, classroom tests, assignments, journals, performance assessments, standardized tests, and other types of assessment devices.

Performance Assessments? Standardized Tests?

It's time to step back and explain the terminology some more. First, I said there were two types of assessments: formal and informal. I then said there were two types of formal and informal assessments: formative and summative. In reality, there are lots of types of assessments.

Standardized tests are tests like the Iowa Tests of Basic Skills (ITBS) or the ACT tests we all dread. "Standardized" refers to the conditions under which the tests are administered. A test is standardized if every student takes the same test under the same conditions. That's why people take the administration of these tests so seriously. The directions and conditions must be the same for every student taking the test in order for the scores to have any reliable or valid meaning (reliability and validity will be explained much later). A student in Florida takes the ITBS in the same way and with the same instructions as a student in Iowa. Since the conditions have been standardized, we are able to compare the ITBS scores from the student in Iowa and the one in Florida.

Performance assessments, also called authentic assessments, are another classification of assessments. You see, most people aren't very fond of paper-and-pencil multiple-choice tests. One argument against such tests is that they do not measure "real world" abilities. People see these tests as measuring trivial knowledge, like vocabulary words or math problems, instead of valuable skills that would be used outside of the classroom. Performance assessments are an answer to the perceived limitations of paper-and-pencil tests. Performance assessments usually involve the student performing some skill and an instructor judging that skill. For example, if you were teaching students to become barbers, you would probably want to use a performance assessment to judge their skills. While a paper-and-pencil test may tell you how knowledgeable they are about cutting hair, a performance assessment would show you if they are able cut a head of hair without cutting the head.

So a performance assessment and a standardized test are opposite things? No, not exactly. A performance assessment can be a standardized test, if every student takes it under the same conditions. A standardized test can be paper-and-pencil (like the ITBS) or a performance assessment. Remember, "standardized" refers to the conditions under which a test is administered.

So What Makes A Standardized Test Better Than Other Assessments?

Nothing. Well, almost nothing. There is room in education for all types of assessments. Standardized tests are valuable, because they provide an unbiased look at student performance. With a standardized test, a student's gender, race, economic status, personality shouldn't influence the test scores (unfortunately, it seems as though some of these factors do influence the test scores). Since every student takes the test under the same conditions and is scored through the same method, standardized tests give us the best way to compare students from different states or different times.

A nonstandardized test, such as an informal observation or classroom test, is valuable for a different reason. Standardized tests are useful, but they are also slow and costly. Informal assessments, on the other hand, are usually quick and relatively inexpensive. Teacher-made tests also let teachers use their judgments (their subjectivity) to make decisions and to judge student performance. Another way of seeing the value of informal, nonstandardized assessments is by imagining what school would be like with only standardized tests. Suppose you were teaching a science lesson on evolution. You would present some information about Charles Darwin, survival of the fittest, etc. You would probably want to know if the students understood what you had taught them so far. If you only had large-scale standardized assessments, you would have each student clear off his or her desk, take out a #2 pencil, and start filling out those little ovals. You would then wait a couple weeks to get the results back and then see that Suzy, a student in the class, didn't understand what was being taught. With informal, classroom assessments, a teacher could quickly ask a review question to see if Suzy understood what was being taught. This example is a bit ridiculous, but it makes the point.

Finally, another way in which large-scale standardized tests may be better than classroom-developed tests is that they are usually more technically sound.

Technically Sound?

Remember earlier when I brought up reliability and validity? Here's where they play a part…

Assessment experts develop standardized tests like the ITBS and ACT over the course of several years. Each item is written, edited, rewritten, scrutinized and selected to perform a function. Thousands of students are given early forms of the test so that the development experts can make sure that their test is as reliable and valid as possible.

Reliability and validity are two extremely important concepts in assessment development. In fact, one could write entire books about each term. Lucky for you, I'm not one of those people. I'll try to explain what reliability and validity are in a few paragraphs.

Reliability is the easier of the two terms to understand. It refers to the consistency of a test score. There are several types of reliability, including test-retest reliability, parallel forms reliability, internal consistency, and inter-rater reliability.

Test-retest reliability refers to the consistency of scores across time. I'll use this space to introduce you to a fictional student named Timmy. Let's suppose that Timmy took a vocabulary test yesterday and that he earned a score of 15 on that test (we'll get into score types later). If that vocabulary test has a high level of test-retest reliability, we would expect Timmy to get a score close to 15 if he took the test again. So if Timmy took the same test today and earned the same score, we would have evidence supporting a high level of test-retest reliability for this vocabulary test. If, however, Timmy took the test again and earned a score of 42, we could say that they test scores were probably not very reliable.

Parallel-forms reliability refers to the consistency of scores across different (but equivalent) forms of the test. Suppose you gave Timmy two vocabulary tests back-to-back. Let's also assume that the tests were equal in length and difficulty. If Timmy earned similar scores on both tests, we would say that they tests have a high level of parallel-forms reliability. If Timmy did really well on one test and poorly on the other test, we would have evidence showing the tests did not have high levels of parallel forms reliability.

Internal consistency is more or less the same thing as parallel-forms reliability. Hopefully you saw the problem with parallel-forms reliability – you have to have 2 forms of the test that are parallel (equal in almost every way). Internal consistency pretends that one test is actually 2 parallel tests.

What?

Ok, I lost myself on that last sentence. Suppose you gave Timmy a vocabulary test with 30 questions on it. If you wanted to establish the parallel-forms reliability of the test, you would need to develop another 30-question vocabulary test with exactly the same difficulty level as the first test. Internal consistency takes your first 30-item test and pretends that you actually have two tests with 15-items each (see how that works? It just split the test in half and pretended that you then had 2 tests!). These two "half-tests" can then be compared statistically to see if the scores on each half are similar to one another. It involves some statistical "magic" but you do get a measure of reliability without having to create another vocabulary test.

Inter-rater reliability is another type of reliability used primarily with performance assessments. This type of reliability refers to the consistency of scores across judges or raters. A good example of inter-rater reliability can be found in the Olympics. When a group of judges all give similar scores to an athlete, we have evidence of a high level of inter-rater reliability. If one or more judges give unusually high or low scores, we have evidence of a low level of inter-rater reliability (and possibly some collusion).

So which type of reliability is the most important to have? We would hope to have a test that is reliable in every way possible. The most important type of reliability depends on the purpose of your assessment. If you are assessing a skill that should remain stable across time, then test-retest reliability should be the most important form to look for. If you are interested in stability across different forms of a test, then you should look into parallel-forms or internal consistency reliability.

Validity

Validity is a broader term than reliability. While reliability refers to the consistency of scores, validity refers to the meaning of scores. The definition of validity is: the meaningfulness, usefulness, and appropriateness of inferences made from test scores. Suppose you wanted to measure a student's mathematics problem solving ability using one of two methods:

1) Giving the student a math test developed by his math teacher
or
2) Measuring the student's shoe size

It's pretty obvious that the first method would give you more meaningful scores than the second method. In that sense, a math test is a more valid measure of problem solving ability than a measure of shoe size. The scores from the math test probably tell us something about problem solving ability, while shoe size probably tells us simply how big the student's feet are.

This example is a bit absurd, but validity is an important concept in everyday assessment. We usually accept assessments as being valid without looking for any evidence to support the validity of the inferences we make. Teachers assign grades based on classroom tests without asking if the tests they create are valid measures of student achievement. Also, the next time you are in line at the supermarket, take a glance at the headlines of magazines they have displayed. You'll undoubtedly see headlines like the following:

1) "Loch Ness Monster Found. Creature Says: 'I'm just shy'."
2) "Lose 10 Pounds in 10 Days With Our Diet Secrets!"
3) "Is Your Marriage Working? Take Our Relationship Test To Find Out"

Focus in on that 3rd headline. Should we decide who to marry (or who to divorce) based on a relationship test in a magazine? Hopefully you're thinking, "No." Any decision or inference we make from that relationship test wouldn't be very meaningful, useful, or appropriate. Making a relationship decision based on this test would be no better than making a problem-solving inference based on a student's shoe size. This is because these tests have no evidence to support their validity. If a magazine were to develop a relationship test and demonstrate that inferences made from the test are valid, we could make our relationship decisions based on that test. And you thought the SAT was a high-stakes test…

So How Do We Demonstrate The Validity Of Test Score Inferences?

The validity of test scores (depending on the purpose of the test) is established through a great deal of research. This research has to show that inferences made from the test are useful, meaningful, and appropriate. I'll go through several factors of validity using the example of a mathematics problem-solving test.

Factor #1: Content & Process Validity

This first factor of validity requires us to show that every item on our problem-solving test actually involves problem solving. On the surface this seems obvious, but it can get a little tricky. Suppose we have a problem-solving item that reads: "The recipe calls for 2 tablespoons of sugar, but you only have a teaspoon. How many teaspoons of sugar will you need to use in the recipe." On the surface, this appears to be a problem-solving item involving division. If you look a little deeper, you will see that this item is actually testing a student's knowledge of how many teaspoons are in a tablespoon. The best problem-solver in the world might miss this item if he doesn't know how many teaspoons are in a tablespoon. Looking even deeper, this item involves reading comprehension. A world-class problem-solver who cannot read will not answer this question correctly. If we dig deeper still, we can see that this item might require students to know something about cooking or recipes. So this simple problem-solving item actually involves cooking knowledge, reading comprehension, teaspoon/tablespoon conversions, and (finally) problem solving.

Some questions to answer when demonstrating content & process validity:

- Do the items measure what they are supposed to measure?
- Are the items comprehensive?
- Do they sample the entire domain of interest?
- Are there any irrelevant items?
- Are the processes needed to answer each item relevant?

Factor #2: Fairness

Remember that one important part of the definition of validity refers to the appropriateness of test score inferences. You cannot make appropriate inferences and decisions from a test that is not fair to all students. Assume that every item in our problem-solving test involves cooking, baking, or recipes. This test would be unfair to students who spend very little time in the kitchen. They would receive lower problem-solving scores than students (with an equal amount of problem-solving ability) who have experience in the kitchen. The impact of an outside factor on the fairness of score inferences destroys the validity of our inferences. The same logic applies to tests that are unfair to a particular gender, ethnicity, or socioeconomic class of students.

There are many ways to demonstrate the fairness of a test across student subgroups. The main questions you should ask yourself are:

- Do students of various backgrounds have an equal chance of succeeding on this test?
- Do any items contain stereotypes are any content that would be offensive to any student?
- Will students with equal ability (but different backgrounds) earn similar scores on this test?

Factor #3: Construct-related Validity

Construct-related reliability refers to the degree to which a test matches the underlying construct we are trying to measure. In our example, we would need to demonstrate that our problem-solving test actually measures problem solving. We do this in several ways.

First, we look at the reliability of the test. If a test is unreliable, the scores it yields are filled with error. Thus an unreliable problem-solving test actually measures random error (and not problem-solving). A highly reliable test may actually measure problem-solving ability, depending on how it stacks up on the other factors of validity. This leads to an important fact: a test must be reliable in order to be valid. However; a test may be reliable, but not valid.

The second way to demonstrate construct-related validity is to look at a test's internal relationships. If we have 20 items on our problem-solving test, we would expect that every item would be related. Students who are good problem solvers should answer most of the items correctly, while poorer problem-solvers should answer most items incorrectly. Every item should also be related in that they should all measure some aspect of problem solving.

The third way to demonstrate this factor of validity is to look at a test's external relationships (relationships to other measurement devices). In order to do this, we need to define our construct (which, in this case, is "problem-solving"). This definition should tell us what problem solving is and (more importantly) what problem solving isn't. In our example, we would expect our problem-solving test to be related to other problem-solving tests (if you earn a high score on our test, you should earn a high score on any other problem-solving test). We might also expect our test to be related to other tests (a high scorer on our test probably should earn high scores on intelligence tests, math computation tests, math achievement tests). On the other hand, we should expect little or no relationship between our problem-solving test and other tests (reading tests,

In order to demonstrate the validity of inferences made from a problem-solving test, we would need to show that students who earn high scores on the test are, in reality, good problem-solvers. We would also have to demonstrate that students who score low on the test are poorer problem-solvers.

Several other factors of validity include (we'll stick with the problem-solving example):

1) Practicality: Is the test worth the cost and time it takes to administer?
2) Reliability: Are the scores consistent over time?
3) Generalizability: Do the scores tell us about problem solving in general or do they only tell us how the student performed on this test?
4) Internal Factors: Do the items relate to each other like we would expect them to?
5) External Factors: Will this test give us scores equal to a similar math test (like we would expect them to)?

You can appreciate how difficult the validity of a test score is to prove. In fact, you cannot prove the validity of a test score. Validity depends on purpose. While we found that the math test was more valid than the shoe size measurement for our purpose, we can think of purposes in which a shoe size measurement would be more valid than a math test.

I'll leave this section with one more note. Did you notice the 5th factor of validity is "reliability?" This shows you that a test can be reliable and not valid, but it cannot be valid and unreliable.

Where Were We?

We were discussing accountability and the types of assessments used in Clinton. Even if you believe large-scale standardized tests are great, you might wonder why we use them in Clinton. Well…

In order to hold its public schools more accountable, the Iowa Department of Education has provided several guidelines for measuring student performance. One guideline states that schools need to measure the achievement of their 4th, 8th, and 11th grade students in reading, math, and science.

But We Already Have Informal Assessments & Classroom Tests!

That's true, but the state guidelines continue…

This measure of academic achievement must come from at least 2 assessment devices that align with the district's curriculum. These devices also must hold up to the most rigorous measures of reliability and validity. The devices must also label students as being advanced, proficient, or below proficient in each of the content areas. So this is why we use the ITBS – they provide us with a highly reliable and valid measure of student academic achievement.

In the past, scores from the ITBS have been reported in percentile ranks, raw scores, grade equivalents, stanines, normal curve equivalents, and standard scores at both the state and national levels. These scores are usually collected, analyzed, and reported back to schools as quartiles. While quartiles show how students compare to a national norming population, they do not indicate which students are advanced, proficient, or below proficient in each content area. In order to fulfill the state requirements, we need a way to measure which students are proficient in reading, math, and science.

Percentiles, Raw Scores, Grade Equivalents, Stanines, Normal Curve Equivalents, Standard Scores???

Ok, I need to take a step back here. All these score types may seem overwhelming at first, but they really aren't that bad. In order to understand ITBS scores, we need to know what a score type means and how it can be used. So let's get started…

Another Step Back…

Actually, I better give you the story of the ITBS before I explain the various score types. Every 8 years or so, the folks down in Iowa City develop a new form of the ITBS. I won't go into detail on how they develop the tests; it is a long and involved process.

After they think they've got a pretty good test developed, they take the test and give it to a sample of 100,000 students across the nation. This sample of students is known as the "norming population." The norming population is a representative sample of students all over the USA. It includes males and females; students from wealthy and poorer families; majority and minority students; special education and non-special education students; and students from the north, south, east, and west.

This norming population becomes the standard to which every student's score is compared. The last norming population was given the test in the year 2000. Students taking the test in 2002 will have their scores compared to the norming population who took the test two years earlier. Having a stable set of scores to use as a comparison makes the ITBS so useful for comparing students across different years.

Step Forward Again

Now we'll go back to explaining the various score types. Remember that each score type has its own purpose.

Raw Scores: Raw scores are the easiest score type to understand. Unfortunately, they aren't very useful. A raw score simply tells us how many items a student answered correctly on a given test. Suppose we know a student, Timmy, who answered 7 items correctly on a vocabulary test. We would know that Timmy's raw score is 7. Is this a good score or a bad score? This question shows the inherent weakness with raw scores. We can't tell how well a student performed simply by looking at a raw score. In this example, a score of 7 may be good if the test consisted of 8 difficult items. A score of 7 may be terrible if we find out the test was 100 items long (or if the test had 8 really easy items).

Percent Correct: Percent correct scores are just as useless as raw scores. If we know Timmy got 63% of the vocabulary items correct, we still don't know if he did well or not. If the test was very difficult, 63%

might be a phenomenal score. If the test was extremely easy, 63% might be a lousy score.

Percentile Ranks: Percentile ranks are also easy to understand. A good way of explaining percentile ranks is through an example. Let's look at our pal Timmy again. Suppose his raw score of 7 on the vocabulary test is equivalent to a percentile rank of 17. This means that Timmy outscored 17% of the norming population. See how easy a percentile rank is to understand? It simply represents the percentage of the norming population a student outscores. The highest percentile rank is 99 – meaning a student outscored 99% of all students in the norming population. Why can't it be 100%? This is a good question with a simple answer: A student cannot outscore 100% of all students, because that would imply that the student outscored him/herself. Likewise, no student can earn a percentile rank of zero.

Now that you understand what percentile ranks represent, you probably want to know how to use them. Percentile ranks are most useful in determining a students relative strengths and weaknesses. If you look at a student's scores on all of the ITBS subtests, the one with the highest percentile rank is that student's relative strength while the subtest with the lowest percentile rank is that student's relative weakness

I have two notes before we go on. First, a percentile rank isn't really that useful in seeing a student's growth from year to year. Second, normal curve equivalents (another score type) are kinda like percentile ranks – well, alike enough that I won't explain them any further.

Stanines: Next we move on to stanines (pronounced "stay-nines"). Stanines are like big groups of percentile ranks. While percentile ranks can range from 1-99, stanines range from 1-9. Here is a chart showing how to convert from percentile ranks to stanines (if you would ever actually want to do so)...

Percentile Rank:	1-3	1	: Stanine
	4-10	2	
	11-22	3	
	23-39	4	
	40-59	5	
	60-76	6	
	77-88	7	
	89-95	8	
	96-99	9	

Stanines can be used jut like percentile ranks. While percentile ranks are more precise, stanines are more stable. For example, suppose Timmy took the same vocabulary test twice. We might expect his percentile rank to go up or down a little bit, depending on random error. His scores on the vocabulary test would probably fall in the same stanine, however. So if we were to discuss Timmy's vocabulary score, we would say he earned a percentile rank of about 17 (it might be anywhere between 15-20, depending on when he takes the test). We would more confidently say that Timmy's score fell in the 3rd stanine (knowing that he would stay in the 3rd stanine as long as his percentile rank was between 11-22).

Grade Equivalents: Let's move on to grade equivalents. They are another popular type of score reported from standardized tests. While percentile ranks & stanines are useful for determining areas of relative strength and weakness, grade equivalent scores are useful in determining a student's growth from year to year. Grade equivalents are not difficult to understand, but they are frequently misused. Suppose our friend Timmy is in the 4th grade. Let's also assume that his vocabulary test score is equal to a grade equivalent of 2.3. A grade equivalent score tells us a student's performance in terms of grade level and months. Timmy's score of 2.3 tells us that Timmy's score is equal to the score we would expect from a student who is in the third (.3) month of second (2) grade. Here are a few more examples:

1) Grade Equivalent = 4.7. This is a score we would expect from a student in the 7th month of 4th grade.
2) GE = K.4. This is what we would expect from a student in the 4th month of kindergarten.

Now you can see how grade equivalents can be used to measure student growth from year-to-year. Timmy earned a 2.3 this year on his vocabulary test. Over the course of one year, we would probably expect Timmy's vocabulary performance to grow by one year. We would expect his score to increase by 1.0 (2.3 +1.0 = 3.3 = what we would expect next year).

So what if Timmy, who is in 4th grade, earned a grade equivalent of 7.1? Would that show us that he could skip all the way to 7th grade? This is the common misconception about grade equivalents. Just because a student earns a certain grade equivalent score, that does not mean the child should be moved to that grade. If Timmy earned a 7.1 on vocabulary, we would have to recognize that he was given a fourth grade test. He did as well as a seventh grader would do on that fourth grade test. We do not know how Timmy would do if he was thrown into a seventh grade class with seventh grade vocabulary. Grade equivalents do not tell us to which grade a student belongs.

You Forgot About Proficiency

You're right – I did. Remember that the state requires us to report the number of students who are proficient, advanced, and below proficient in reading, math, and science. So how do we go from those score types to determining if a student is proficient or not?

It just so happens that certain percentile ranks correspond with levels of proficiency. Here is the table showing that relationship:

Achievement Level	Percentile Rank Range
Below Proficient	1-40
Proficient	41-89
Advanced	90-99

Remember that a percentile rank shows the percentage of the norming population a student outscores. Based on this table, we would expect around 40% of our students to be below proficient; 50% of our students would be proficient; and 10% of our students would be advanced.

So why are we bothering to report achievement levels instead of percentile ranks? Percentile ranks show us how well a student did in comparison to other students, but they do not show us what a student can or cannot do. Achievement levels let us know what a student is capable of. In that sense, achievement levels are more of a criterion-referenced measure than a norm-referenced measure.

Criterion-Referenced? Norm-Referenced?

Norm-referenced test scores are simply scores that were compared to a norming population. They allow us to rank-order students to see which students were the highest and lowest achieving. Criterion-referenced test scores, on the other hand, are not compared to a norming population.

Criterion-referenced scores show us how much of a content area a student knows. They allow us to do more than compare students to each other. They show us how well a student is doing in comparison to the domain of interest, be it reading, math, or science. With norm-referenced scores, we always have a student at the top and a student at the bottom. With criterion-referenced tests, it is possible to have every student achieve the highest (or lowest) score.

I should note that the distinction between norm-referenced and criterion-referenced tests is artificial. Scores from any test can be interpreted in a norm- or criterion-referenced manner.

Do We Have A Criterion-Referenced Assessment?

Assuming that we consider the ITBS to be primarily a norm-referenced test, we would probably like to have a criterion-referenced test, too. Remember that the state guidelines tell us that we need to have two measures of student achievement in grades 4, 8, and 11. The ITBS is our first measure; our second measure is…the district Grids.

The Grids will allow us to see proficiency not in a norm-referenced manner, but in a criterion-referenced sense. They will allow us to see the achievement levels of each student individually, instead of comparing each student to a norming population. And since the Grids have been developed internally, we can be assured that the Grids align almost perfectly with our district curriculum.

So you can see how the Grids will help us meet the state accountability requirement. In addition to being a more criterion-reference test than the ITBS, the Grids differ from the ITBS in another way. Remember that the ITBS tries to remain completely objective. It standardizes everything down to one series of tests that represent a student's academic achievement. The Grids will use teacher judgments and multiple sources of information (journals, assignments, observations, classroom tests) in order to show us which students are proficient in reading, math, and science. Since the Grids are also a formative assessment, teachers can use them to help adapt instruction so that all students will have a better chance of becoming proficient.

Wrapping Up

The Grids and the ITBS fulfill our state requirement of accountability, but we do not stop there. We have other assessments and other sources of information that will help us make informed decisions about students and instruction. If we need to make decisions about students and schools, we should at least make informed decisions. The district's assessment plan will help ensure that we will make informed decisions about all students in the Clinton Community School District.