The Crooked Road to Academic Assessment
Are We Going Anywhere - and How Are We Getting There?

Charles J. Clock, Ed.D.

Introduction

Testing, for determining aptitude or general mental ability, got its start in the U.S. in the very late 1800's and early 1900's. This type of assessment led to the development of achievement testing in the early 1920's. The primary use of these tests was for selection purposes in both the military and academic areas. The key here is the word "selection", which remains dominant in the testing field today. The major question remains, "what are we really trying to do?" With many of our tests, we end up measuring what the child brings into the classroom rather that what is learned in the classroom. Whatever we use is only going to measure a relatively small aspect of any human behavior, and we have not been very accurate in that process. In education, the best we can do is estimate performance on a very limited number of skills in very specific areas. As long as we continue to "test", the need is to find the best way to establish the least amount of estimation. The long term goal should be to develop a more instructionally based assessment system. But that is another chapter.

The new Federal law (NCLB) states that academic assessments are to be valid and reliable for the purpose for which they are used and "involve multiple, up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills." Very few multiple-choice tests, if any, measure higher-order thinking skills. Assessments must also "produce individual student interpretative, descriptive, and diagnostic reports…that allow parents, teachers and principals to understand and address the specific needs of students". To date, reviews of the current testing situation across the United States find that **no** state or commercial tests now used fully meet these requirements. The almost total misuse or misinterpretation of testing information by federal and state agencies, as well as the educational community and the media, demonstrate we are far short of proper knowledge about how to accurately understand or interpret individual or group test data. This law does **not** require that any particular test or assessment process should be used, as long as it meets the above requirements.

It is important for the public to understand that this Federal "accountability" mandate is a major and costly undertaking, and one that could have undesirable results, both socially and academically. The consequences are serious in that it sounds logical and sensible, but the necessary direction and financial resources are not there to support the desired implementation or results. Just designing the tests alone would be a major task, one that many locations may not have sufficient expertise to accommodate. As Grant P. Wiggins stated in 1993 (Assessing Student Performance) - *we have the tests we deserve*.

The purpose of this document is to define some of the theories and words that are often used by test publishers or test specialists in dealing with the public. Many of these are terms that are not even understood by educators, since few receive training in testing and measurement. This document will also outline some of the problems in designing, developing, using, and interpreting the products of the current academic testing generation. This is a brief culmination of articles, research, and my own personal experiences over the past 45 years. The focus is to take a short

look at where we were, a longer look at where we are, and some thoughts on where, perhaps, we should be going.

Brief History

The early testing movement brought about the development of Classical Test Theory (CTT), and the primary authors of this theory were Francis Galton and Charles Spearman. Much of Galton's work was focused on the development of statistical techniques that would support the theories of "natural selection" originally proposed by Charles Darwin. Carl Brigham, a professor at Princeton University, conducted a study in 1923 of members of the military using a test he designed to determine the soldier's "mental capacity". From the results he concluded that foreign born individuals were intellectually inferior to native born Americans - but there was no question about the fact that these tests were loaded with item bias - favoring what Brigham referred to as the "Nordic" types. Brigham went on to create the original Scholastic Aptitude Test (SAT) which was designed to select "appropriate" students for Princeton University. (Note: The original SAT was called the **Scholastic Aptitude Test**). In defense of the SAT, it has undergone considerable change in both content and psychometrics since 1923; however, the main key is still "selection". To this day, two of the variables that correlate most significantly with **high** SAT scores are: family socioeconomic status and parent educational backgrounds. I single out the SAT now briefly to make this point, for it was really one of the foundations of the current U.S. testing movement. The following are the factors that have the highest correlation and most influence on SAT performance - listed in order of importance:

Family has high income	YES!!!
Family lives in a high socioeconomic community	YES!!!
Mother and/or Father is a college graduate	YES!!
Students attend wealthy suburban schools	YES!
Students are male	Yes (females do better in college)
Students use superficial thinking strategies on items	Yes
Students are multiple-choice item test-wise	Yes
Students have a high GPA in high school	Maybe
Students have a high rank in class	Maybe
Students have a high GPA in college	Sometimes

To put it another way, a major purpose for the SAT and ACT is to predict college (particularly freshmen year) grades. Numerous studies over the years have shown that these tests are poor predictors of college or life success. Even when added to high school GPA, their contribution is very low.

Test Theory

To date, the most common test theories, particularly in the use of multiple-choice test items, are: Classical Test Theory (CTT), and, more recently, Item Response Theory (IRT). There are considerable differences between these two theories - in terms of item construction, scoring, and the interpretation of test results. Both of these two theories involve considerable psychometric considerations which will not be dealt with in this document. However, an attempt will be made to identify some of the more important features.

Classical Test Theory

- The procedure used primarily to develop **Norm Referenced Tests (NRT).**
- First credited to Charles Spearman in 1907.
- Item characteristics (difficulty and discrimination) depend on the particular examinee samples tested. Therefore, items work well only if the sample tested matches the base sample - or the norm group from which the original data were obtained. That is a very big assumption.
- Performance on a parallel test to determine "improvement" is dependent on the ability level of the original examinees. Therefore, most tests are designed for middle ability level students - and individuals are never exactly the same on a second administration. Hence, test reliability (how consistently it measures what it's supposed to measure) is put into question.
- Presumes that the variance of measurement error is the same for all examinees, **and measurement error is based on performance of students on <u>all</u> the test items - the <u>total test score (test statistics)</u> is what is important.**
- Composed, in most cases, of multiple-choice test items (strictly timed) designed to force students into a normal distribution scale for comparative purposes. The key here, is again, the quality of the sample reference group.
- Measurement error applies to all scores in a particular population based on a linear transformation of raw scores into equal interval scaled scores. **Therefore, longer tests are more reliable than shorter tests, and the better will be the <u>test statistics</u>.**

Classical Test Theory is definitely focused on force fitting scores to a normal/bell shaped curve. Bloom, Madaus & Hastings, in their 1981 publication on testing, addressed this problem very well: "There is nothing sacred about the normal curve, it is the distribution most appropriate to chance and random activity**. Education is a purposeful activity, and we seek to have all students learn what we have to teach. If we are effective in our instruction, the distribution of achievement should be very different from the normal curve. In fact, we may even insist that our educational efforts <u>have been unsuccessful</u> to the extent that the distribution of achievement approximates the normal curve."**

In summary, **Classical Test Theory** has value in comparing a student's position to a reference group - giving rise to the name: Norm-referenced tests. That, of course, assumes that the norm group is representative of the skill ability level or population desired. The idea is to "spread the scores out". Items will "work" if they discriminate well enough to provide a score spread that will accommodate rank scores, like percentiles. Consequently, a normal distribution of scores is usually sought and scores are forced to that distribution. Most tests designed under CTT are **<u>not</u>** loaded with items measuring material taught in school. If they did, and the instruction was successful, there would be insufficient score spread. **Test items that most students pass are often deleted from these kinds of tests, and they may be the very ones reflecting skills being effectively addressed in the classroom.** Also, you need lots of items in order to get good reliability indices. This theory is the basis for the PSAT, SAT, ACT, and most current commercial norm-referenced standardized tests. Almost all of these attempts have involved multiple-choice, machine-scored test items, since they carried with them a sense of "objectivity" in the computer scoring process, and were relatively cost-effective to purchase and score. Unfortunately, this "sense of objectivity" was focused on scoring, not on how the stems and answer options of the items were developed. Basically, Norm-Referenced Tests tell us that some students are more or less proficient that others, but they do not tell us how proficient they are in the subject matter being

tested. In contrast, Criterion-Referenced Tests do compare individual student performance to a given standard.

Item Response Theory

- The procedure used primarily to develop **Criterion Referenced Tests (CRT).**
- First credited to Binet and Simon in 1916.
- Originally focused on determining the underlying latent ability, attribute, factor, or dimension being assessed.
- Persons and items are placed on a common scale. **They need not be based on a normal distribution.** Information will be based on **item statistics** that reflect performance on items or tasks related to specific goals and/or standards.
- Key concept: Is there a good fit between item performance and the overall trait or task being measured? If so, then the **difficulty level of the item** is a solid index of where the examinee stands with respect to the underlying trait being measured. In other words, a person scoring higher than another person on a test instrument is assumed to possess more of the construct in question (reading comprehension, algebra, geometry, whatever...). By the same token, an item (or task) which scores higher in difficulty than another item (or task) must be viewed as demanding more of the construct. In short - it is the **individual item statistics** that are important.
- CRT's compare a student's performance with present standards of acceptable performance, and permit the detection of specific strengths and weaknesses in individual achievement. The focus here is on determining specific content strengths and/or weaknesses.
- This test theory is a very useful tool in the design of Computer Adaptive Testing - where the computer will automatically branch the student to easier or more difficult items based on how he/she progresses through the test. **This feature can make shorter tests more reliable than longer tests - and also has the benefit of reducing testing time.**

Item Response Theory places person and items on a common scale. They need not be based on a normal distribution. Your position is how you relate to the overall trait/skill tested and your performance on specific items. **Item "spread" is not as important as item "construct validity" - in short, is the item valid in terms of its content and item statistics.** There has to be a good fit between how well the individual performs on items **in concert** with the overall trait being tested - making reliability inherent in IRT. Consequently, fewer items are required to adequately assess a skill level, and the item content complexity is more a function of how far the student can go. Since item properties are known, directly linked to test behavior, and all results are linked to a uniform scale, the results provide more relevant instructional information as to which skills are successfully accomplished and which will need improvement. The data reflect both indices of standard deviation and standard error - based on specific content areas and item performance within those areas. Measurement error can be used very effectively to determine consistency of item response patterns. This is a very powerful theory for the development of **computer adaptive testing**. It is also the theory that was used in the development of the Idaho State test, the ISAT.

Regardless of the "Theory" being used, the power of the product will depend on the quality of the items. It may be possible to get better test quality from IRT, due to the way the items must perform in concert with each other. However, the ISAT, for example, and many like it in measurement today, continue to use multiple-choice test items, and their quality can easily be diminished by poorly designed stems and distractors. Items should be designed to assess content knowledge, not test-wiseness; and complexity should be assessed by content, not by "trick" or inadequate

answer options. **Basic problems with multiple-choice test items - particularly those developed through CTT, is the inability of these items to reflect evidence that the student has actually understood what was taught simply by picking the right answer. It is also conceivable that an incorrect answer may reveal greater insight into the knowledge of the subject if we actually knew why the student made that selection.**

Problems with Percentiles

The type of test used is going to dictate how the results are interpreted. For example, scores from a norm-referenced test will tell you how a student performs in relation to the scores of particular groups of students. These scores are usually in the form of percentiles. Percentiles can be easily misinterpreted. For example, many schools and state governments show graphic displays of percentile ranks because they provide an apparent picture of relative success. You can often see percentile bar graphs comparing a school or district percentiles with "National Norms" or state-wide data. However, local percentiles are usually higher than national percentiles, since the local group will not be as strong a comparison group. This is due to the large difference in the number of students involved in these two groups, and the compression of the relative distributions. The differences, though seemingly great, are often meaningless or insignificant. In addition, percentiles that are called "National Norms" are often misinterpreted as being data from the nation as a whole. These "National Norms" are from a sample of students which may or **may not** be from every state, and may have been tested 10 or more years before. These data may not represent the quality knowledge base, or possibly even the grade level you expect. An even greater problem, Standard Error of Measurement (SEM) is almost never considered - which frequently can make the comparisons totally insignificant. Percentiles are greatly impacted by SEM. Percentiles in the IRT context compare a student against the construct being measured.

Comments from the "The National Center for Fair and Open Testing - FairTest", Cambridge, Massachusetts:

> Norm-referenced tests place test-takers along a "normal" or "bell-shaped" curve, with most in the middle and few at the ends. Half the students must be "below average" by definition. The tests are based on the assumption that whatever the test measures should be distributed along the curve, therefore they only include items which ensure students will be sorted that way. The curve reinforces the view that instruction will be ineffective for many students, and it encourages tracking and low expectations for them. [Some students may feel they "belong" at that low level which could negatively impact motivation to excel.] It also makes it impossible to demonstrate progress except in comparison to other students; if all are improving (or getting worse), the tests cannot show it. Norm-referenced tests are usually composed of multiple-choice questions, and they treat learning as memorizing isolated pieces of information, rules and procedures. This approach assumes one first accumulates the bits, and only later thinks. To the contrary, psychology now understands that humans from infancy learn by actively attempting to make sense and meaning of their world – they construct knowledge and meaning in their minds. Learning is also social and contextual. Because the multiple-choice testing method is incompatible with how people learn, it fails to provide information essential for instruction.

The current work on criterion-referenced tests (CRTs), based on Item Response Theory, is helping to focus more on assessment (in a positive skill feedback process) than simply testing for scores.

CRT's compare a student's performance with preset standards of acceptable performance. This permits the detection of specific strengths and weaknesses in individual achievement. A good example of a CRT would be a performance assessment instrument, where the items are open-ended (not multiple-choice) and they have stated rubrics for scoring. The students must know the answer and be able to explain how they got the answer. This CRT approach lends itself far better as a means of obtaining information for instructional improvement.

Item Samples

Below are examples of tests items - one strictly multiple-choice item (MC), and one is a constructed response item (CR).

A typical Norm-referenced test item question (NRT) in Multiple-Choice format:

Question: In a list of eight integers, 13 is the lowest member, 60 is the highest member, the mean is 32.5, the median is 29, and the mode is 18. If you added the numbers 8 and 65 to the list, what would the new median be?

(A) 26 (B) 27.5 **(C) 29** (D) 32 (E) 37.5

(Note that the item above could also be from a criterion referenced test, but chances are the answer options may be altered, and the stem would be re-worded to more accurately represent the difficulty level/content linkage that is being assessed. As the item now stands, there is no way we can determine what thinking skills went into the selection of an option. This is a classic NRT type of item.)

Same question in a CRT Performance Assessment test in Constructed Response (open-ended) format:

Question: In a list of eight integers, 13 is the lowest member, 60 is the highest member, the mean is 32.5, the median is 29, and the mode is 18. If the numbers 8 and 65 are added to the list, explain what impact that would have on the three averages.

Desired Answer: Since 8 is lower than the other numbers, and 65 is higher, they will not change the value of the median. Since the mode is the number repeated more often, and 8 and 65 were not in the original list, the mode is not changed. Adding two additional numbers of this magnitude; however, could change the mean.

Obviously you get a lot more valuable instructional information in the Performance Assessment example. In this way, the process is identified and the student understands the principles behind all three averages.

Another Example as a NRT Question:

What was the date of the battle of the Spanish Armada?

(A) 1492 (B) 1535 (C) 1560 (D) 1588 (E) 1654

The student selected the answer 1588 which is correct. He was asked afterwards, "What can you tell me about what this meant? His response was, "Not much. It was one of the dates I memorized for the exam."

Put another way as a CRT Question:

Question: What was the date of the battle of the Spanish Armada, and explain why you picked that date?

Answer: It must have been around 1590. I know the English began to settle in Virginia just after 1600, not sure of the exact date. They wouldn't have dared start overseas settlements if Spain still had control of the seas. It would take a little while to get expeditions organized, so England must have gained naval supremacy somewhere in the late 1500's.

Standard Error of Measurement

Standard Error of Measurement (SEM) is a fact of testing relating to any multiple-choice test, but not often used by academic organizations and almost never used by the media in interpreting data. Standard Error is based on the principle that test results are a product of human behavior, and human behavior is never static. If a person were to take the same test over and over again without any intervening instruction, the scores will vary. Consequently, an individuals "true" score is not an exact number, but will fall in a range of numbers around the **obtained** score. Therefore, Standard Error is based on two statistics: the variability of the scores in a given test administration and the reliability of the test. The less the variability and the better the reliability, the smaller will be the range wherein the "true" score lies. The following is a case in point dealing with a Norm-referenced Test:

Sue Smith took the Science section of a norm-referenced standardized achievement test. She received a Standard Score of 250 and a local percentile of 51 – meaning that she performed as well as about 50 percent of the individuals at her grade level in the sample tested on the Science items, or about average. However, the Standard Error for this subtest is **14** Standard Score points, so **14** points in either direction from 250 results in a Standard Score range of from 236 to 264. This would result in a local percentile range from 32 to 64. Therefore, we can be relatively certain (with about 68% probability) that if Sue took this test many times without any intervening instruction in Science, her scores would range from 236 to 264, or from the 32nd to the 64th percentile. In other words, her score of 250 is simply the midpoint in a range of potential scores for Sue considering the degree of testing error imbedded in the test she took. Every norm-referenced standardized test score has some degree of standard error built in. The SAT, for example, has a Standard Error of at least **30** points.

Standard Error has an even greater impact on school or group norms, and can cause them to vary greatly with only an item or two difference in total score. This results in many school district scores, usually published each year in the media, as being extremely misleading without the benefit of also publishing the amount of possible error range or at least mentioning its existence. Any growth or improvement implications need to deal with the same children moving in time (longitudinal data) and use standard error to validate significance. Upward or downward trends of totally different groups of students from year to year mean very little without these considerations.

Standard Error of Measurement is also a part of tests designed under Item Response Theory. However, the application is totally different. In Item Response Theory, where the trait scores are estimated separately for each score or response pattern, SEM controls for the difficulty of the items. SEM is by item and depends on the fit of the examinees to the items.

Other types of data have a major impact on the interpretation and application of test results - whether we are dealing with NRT's or CRT's. For example, in data interpretation, there could be a large difference between averages, like mean and median - and yet these are seldom defined when addressing results. The same goes for standard deviation. In the case of CRT's, standard deviation can be important in describing the diversity of the instructional levels of the class, or student group. Also with tests designed under Item Response Theory, terms like p-values, RIT scores, Lexiles, can all have significant instructional impact at the teacher level. These things are not taught to most educators in college - but can and should be addressed in on-site professional development programs.

Summary of major reasons for NOT using Norm-referenced Tests for data feedback for instructional improvement, or for high-stakes tests:

The results tend to be score based, not skill based. How do scores help to improve instruction? What specifically will you improve? These tests are too general to identify specific problem areas. In addition, in most cases the item distractors upon which the scores are based, are designed to trap those who are not test-wise into segments of the normal curve where they do not belong. Another problem with the distracters is the possibility that they can confuse those who may be too intelligent or creative to discriminate “properly” on mundane tasks – they can rationalize almost every option. Even when skill data are available from these tests, the limited basis for making judgments about the skills make the results suspect at best. Is all this measuring learning?

Norm-referenced standardized achievement tests also have psychometric qualities that will cause them to produce predictable results with certain segments of the population. These aspects have definite implications for proper score interpretation and validation, particularly when these tests are used in instructional/program evaluation. This expertise is usually not found in educational decision-makers unless they have had special training in this area – and very few have.

These tests are timed - what is important is what is done quickly and under pressure. The premium on speed is more significant than creativity or thoroughness. Some students do not do well under these conditions even when they know the content; consequently, their special skills are being overlooked and they are being intellectually handicapped.

These tests are concerned only about whether or not the student got the right answer, and even that is clouded by the guessing factor and test-wiseness. Right answers do not necessarily indicate understanding, and wrong answers do not necessarily indicate the absence of understanding. Higher order thinking processes are **not** being tested, and many of the problems bear no resemblance to the real world.

These tests are almost entirely made up of multiple-choice items – a question format that is inherently limited and limiting – students are not allowed to generate a response. “I don’t think there’s any way to build a multiple-choice question that allows students to show what they can do with what they know” (Robert Farr, professor of education at Indiana University).

A typical norm-referenced standardized achievement test is given at specific times of the year and these times may or may not correspond directly to the scope and sequence of the content of on-site instructional programs or efforts. In short, what they often measure are the least interesting and least significant aspects of learning at the lowest level of the skill hierarchy**. If an educational organization has an objective to meet the needs of <u>all students</u>, then it is important to keep in mind that these kinds of tests are very limited in scope.** Most educators would agree that the following list contains at least a minimal number of different types of abilities (intelligence) that different students might possess, obviously in varying degrees. Almost all academic tests - particularly those used for college admission or high-stakes tests - cover a very limited number of skills in the first ability, and maybe a few of the second ability listed. The remaining nine are not touched.

1. **Logical ability: to reason, solve, analyze**
2. **Verbal ability: to communicate effectively**
3. **Visual ability: to communicate in visual terms**
4. **Musical: ability to understand musical elements**
5. **Interactive ability: to interact effectively with others**
6. **Intrapersonal ability: to understand self**
7. **Kinesthetic ability: use of body to achieve goals**
8. **Creative ability: to invent, discover, create**
9. **Moral ability: to make evaluative judgments/insight**
10. **Political ability: to conceive/understand politics**
11. **Synthetic ability: organize people, resources, concepts**

Let's look at the classroom in this context. Of the six cognitive instructional objectives usually taught (below), only a limited amount of **Knowledge** (basic recall) and a limited amount of **Comprehension** is assessed with almost all norm-referenced standardized achievement tests.

Knowledge	basic recall
Comprehension	lowest level of understanding
Application	use of abstractions in situations
Analysis	breakdown of communication into its component parts for clarity and understanding
Synthesis	the putting together of elements and parts so as to form a whole
Evaluation	judgments about the value of material and methods for given purposes

<u>Some Comments on College Admission Testing</u>

Most all college admission tests were developed under the Classical Test Theory - and carry with them all the associated problems that relate to decision-making based on any one measure. This certainly applies to the PSAT, SAT, and ACT.

No single test should ever be used as a sole measure of human behavior. This is particularly true in the process of college admission or high-stakes tests, and where the measure is limited to multiple-choice test items. Too many factors are involved in answering multiple-choice test items that **<u>do not</u>** relate to what is learned in school or what actual knowledge the student possesses about the subject in question.

As stated before in this document, three important reasons are:

Requirement of normal distribution - spread of scores & ranking
Testing more of what is brought to school than what is obtained in school
Full of multiple-choice test items

An excellent article by Rebecca Harris (age 17) from the Salem, Oregon Statesman Journal (October 2003) reports that the SAT tests school-learned math and reading skills – but ultimately, these skills are only tools used to assess “math and verbal reasoning”. She states, "This 'reasoning' is, ostensibly, some kind of logic, but more than anything, it requires speed – and a brain that thinks like a test writer". Rebecca hit the nail on the head.

There is no question that some students do well on multiple-choice test items when their knowledge of the subject being tested is minimal. Conversely, there are some students who do very poorly on multiple-choice test items when their other academic data are excellent. Hence, the necessary national focus on teaching test-taking strategies.

Lest we forget, the correlation between test scores and family income/ethnicity is very significant. A recent study (Orlich & Gifford in References) which is due to be presented this fall in Washington, D.C., strongly supports very high correlations between family income and test performance, particularly on the SAT and ACT. This research is a powerful support of many other similar conclusions. In short, your odds of doing better on the SAT and ACT depends largely on your family wealth and ethnicity. This has a great impact on selection criteria, scholarship aid, and educational opportunities for the poor.

381 students were admitted to University of California (Berkeley) in 2002 with SAT scores falling between 600 and 1000 – well below the 1330 average for entering freshmen. (Remember, this was based on the old SAT I. The New SAT was revised in 2005, so the current score scale has changed. However, with the exception of two subtests, it is the same test.) What about the 381 who had the low SAT scores? They were in the top 4% of students who completed UC course requirements and graduated in the top 4% of their class; showed promise as community leaders, athletes, musicians or artists, or had overcome hardships. These data are very consistent with the research that shows there is a weak correlation between the SAT or ACT and college or life success.

Time limits on the college admission tests have always been a problem for many students. ETS reports that they have always been strictly timed. Yet the College Board concedes the time limit isn’t intended to measure how students perform under a deadline – rather the restriction merely serves a logistical purpose. These time constraints place unnecessary time pressures each year on the more than 2 million students whose scores can have a major impact on their college careers. The approach also runs counter to the SAT’s goal of predicting how students will fare in college, which typically provides ample time to complete coursework and exams. An even bigger concern is that students with learning disabilities receive 90 extra minutes to finish the SAT. Starting the fall of 2004, colleges no longer had to inform ETS as to which test takers get the extra time, a change the College Board made after disability advocates threatened discrimination suits. But the new policy also creates an incentive to make bogus disability claims. Test administrators admit that some families pay private psychologists to declare their child disabled. A solution? Give everyone an extra 90 minutes! An even better solution, turn off the clock. Of course, that

would defeat "standardization". At present, the SAT creates an opportunity for some to cheat and prevents others from fully demonstrating their abilities.

Robert Schaeffer, the education director for the National Center for Fair and Open Testing (FairTest) claims that academic ability is far broader than what the SAT measures, and that there is no evidence that any test, no matter how it is constructed, will tell how well students will do in college any better than high school performance. This is heavily supported by recent research which indicates that GPA is one of the best predictors of college performance.

Some new research: Robert Sternberg of Yale University is working on a project, funded by the College Board, to revamp the SAT so that it will include exercises that are designed to measure analytical strengths, creative ability, and practical reasoning. So far, the preliminary results showed that it does do a better job of predicting actual college success for a wider range of students than the current SAT. Tom Fischgrund, author of the new book “Perfect 1600 Score: The 7 Secrets of Acing the SAT”, (that would relate to the SAT I test that was revised in March 2005) says that many students he has studied in the past several years who have made perfect scores on the SAT, have a full range of characteristics for being successful in college and in life. The main difference between students with perfect SAT scores and those with average SAT scores was that the former tended to read more, made more effort to prepare for the exam, and often enrolled in a couple of SAT review classes.

The following are comments regarding the use of SAT/ACT data for college admission from three highly respected organizations.

National Research Council, National Academy of Sciences:

> “SAT and ACT scores are estimates of student performance with substantial margins of error, not precise measures of ‘merit’ – even academic merit. Consequently, the assumption that either test measures the criterion that should bear the greatest weight in admissions is flawed.”

Standards for Educational and Psychological Testing:

> “In elementary or secondary education, a decision or characterization that will have a major impact on a test taker should not automatically be made on the basis of a single test score.”

1999 College Board publication (the authors of the SAT) “College Bound Seniors”:

> “The gender gap is continuing to grow wider with females now falling 43 points behind males.” Since both first year college and senior high school women tend to make higher grades than males, this negatively impacts scholarship aid and denies females equal educational opportunities."

Why do women score lower? One possibility. Men use surface level thinking, women use analytical thinking - speed is a significant factor - especially in reading passages. Women tend to read the passages - it is a time killer.

Creativity and the SAT

In defense of the SAT, there are aspects of this test that do have merit in some situations. Tom Fischgurnd, Ph.D., in his book "Perfect 1600 Score: The 7 Secrets of Acing the SAT", did an excellent job of outlining one important potential application. The following sums that up:

We don't normally think of creativity as a form of scientific thinking, but that's precisely what it is. In fact, the same analytical thinking that fuels creativity is the type of thinking needed to figure out SAT questions. Some perfect score students say they aced the SAT because they were able to **get inside the heads of the question writers**. "After taking practice tests and analyzing the questions I got wrong, I figured out the correct way to answer the questions," says Matthew S. "We're always taught in school to look at things from different angels, but there's a certain logic to the SAT questions. I learned it and gave the answers I knew the College Board was looking for. The second time I took the SAT, when I got a perfect score, I knew which section was experimental just by the types of questions that were asked. I was so certain of it that I left this section completely blank." (I think that is great for Matthew - but I do not recommend that for everyone.) What's interesting is that Matthew **used his creative skills to crack the code of the SAT.** Most of us wouldn't think of this as an act of creativity, but that's exactly what it was. Matthew employed his analytical way of thinking and meshed it with a practical approach to the test. When he thought his answer seemed too obvious for a difficult question at the end, he applied a fresh creative approach to the question and took another stab at it. I would argue that Matthew is a case in point that the SAT can, indeed, test a student's creative thought processes.

There are some non-traditional colleges out there who rely very heavily on the SAT for this reason. They are looking for students who can, as I like to say, "think out of the box". This, I believe, is an excellent option - perhaps not as a single criterion - but certainly one great possibility.

Is There a Better Way to Get There?

The "Basic" Plan

Before any state, regional, or local plan can be considered, it will be necessary to understand the federal implications. One major concern would be to not over-test. Also, the federal system is looking for data that reflects higher-order thinking skill, scores, and several "diagnostic" reports. One could be testing all year! It seems strange that the Feds don't make more use of the National Assessment of Education Progress (NAEP) as their means of getting statewide and national data. That system is already in place - but could perhaps use some enhancements. This is based on multiple-matrix sampling so that no student, school or district would be tied up for long periods of testing time. If NAEP was more "formalized", then that would leave the local districts and schools to design assessments based on their own curricular efforts.

Several states, including Idaho, have resorted to the Item Response Theory as a means to solve both state-wide data and also provide some instructional feedback at the local level. I would see this as involving instructional administration and staff contribution on the content development of the instruments, training on understanding the statistical concepts involved, and training on interpreting the results - both for instructional feedback and reporting to parents. This section of the document will focus on the use of state-wide high-stakes tests. Most states already have, or

should have, something in operation, so this section might be best used to check if everything is up to par.

Some time ago, Robert L. Linn, Distinguished Professor at the Center for Research on Evaluation, Standards, and Student Testing at the University of Colorado, wrote an article for the American Educational Research Association (AERA). His comments were offered as a way of enhancing the validity, credibility and positive impact of assessment and accountability systems while minimizing their negative effects. I think these 6 points are well taken, and I have added a few of my own in brackets.

1. Provide safeguards against selective exclusion of students from assessments. This would reduce distortions such as those found for Title I in the fall-spring testing cycle. One way of doing this is to include all students in accountability calculations. [An important point since some sort of selectivity is practiced in practically all school systems and tends to distort the results.]

2. Make the case that high-stakes accountability requires new high-quality assessments each year that are equated to those of previous years. [This is not known or understood by most educators and the general public, but a very valid point.] Getting by on the cheap will likely lead to both distorted results (e.g. inflated, non-generalizable gains) and distortions in education (e.g. the narrow teaching to the test).

3. Don't put all of the weight on a single test. Instead seek multiple indicators. [He is advocating getting away from an aggregate score and at least use subtest or skill scores.] The choice of construct matters and the use of multiple indicators increases the validity of inferences based upon observed gains in achievement.

4. Place more emphasis on comparisons of performance from year to year than from school to school. This allows for differences in starting points while maintaining an expectation of improvement for all. [Advocating use of longitudinal comparisons – same kids moving in time – rather than making inferences from group data composed of different children each year – an extremely valid point but almost never done. Comparing any test scores on totally different groups of children each year and making inferences about progress is misleading, meaningless and irresponsible.]

5. Recognize, evaluate, and report the degree of uncertainty in the reported results. [For example, scores of any type of multiple-choice item tests should never be displayed without their associated definition or consideration of standard error.]

6. Put in place a system for evaluating both the intended positive effects and the more likely unintended negative effects of the system.

Assessment Implementation - Questions to be Answered

Who will develop the instruments? Will they be commercial NRT or CRT instruments? If locally developed – who? Will they include proper psychometric considerations? Is IRT used in the development? Is multiple matrix sampling a consideration – particularly with respect to state or national data if political information is all that is desired?

Will they provide only “global” or political type of score information, and will they allow for specific skill analysis?

Will test items be limited to multiple-choice – which introduces many other factors in trying to determine what is being assessed?

Who will take the tests? At what cost, and to whom?

Who will determine cut-off scores? Will they be based on total test, subtest, or specific skill data? How will they be determined?

Will measurement error be taken into consideration? Will tests be recognized as fallible assessments of human behavior – in light of the fact that human behavior is variable, not static, and has never been thoroughly or accurately assessed, particularly at any one given time?

Remediation

Test cut-off points for remediation – what are they? Will they include SEM?
There is considerable history of a lack of agreement on specific standards of behavior.
Decisions on cut-off points involve major cost considerations.
Cut-off scores involve value decisions – what is enough for any one student?
What level of skill is involved: Total, Subtest, specific Skill(s)?
What does the score(s) describe? A math item may require a student’s ability to comprehend what is written, perform analysis, as well as the ability to compute.
What type of scores will be used and how will they be interpreted?
Does ability or aptitude enter into remediation – or are scores based only on achievement?

If remediation is involved, then remediation should be based on as many variables as possible to determine both the legitimacy of the deficiency and the direction of remediation.

Reports

What will be the specific content of the reports?
Will they be based on individual cases, groups, or large numbers of students and programs?
Is instructional improvement a consideration? Improvement of administrators, teachers, students – how will that be handled?

If the intent is to improve deficiencies besides just being punitive, then the recipient should be the student and those who are responsible for their remediation.

General student considerations:

Does, or should, everyone have to be “minimally competent” in every area tested? Is being poor in math indicative of a person who is a failure in life? Do we relegate all those who do not meet “minimal standards” as failures?

What about those students who have difficulty demonstrating their knowledge through multiple-choice tests? They definitely exist. Can they be tested in other ways – i.e. essay or oral exams? Are these “other ways” even considered as alternatives?

Will the minimum standard become maximum – with teachers over-teaching the minimum requirements at the expense of other areas – already a concern in many districts?

After teaching the basic decoding skills – what is defined as functional literacy? If you teach a foreign born person (who has no understanding of English) basic vocabulary and decoding skills and present him with an English dictionary, how would you expect him to translate "He works around the clock"?

Positive and negative outcomes of minimum competency or proficiency tests:

Positive	**Negative**
Students	
Award of meaningful diploma	Arbitrary denial of diploma
Early identification of need	Restriction of learning outcomes
Provision of remedial help	Attachment of negative label
Monitoring of progress	Restriction of employment
Teachers	
Information for instructional mgt.	Unfair evaluation
Clear instructional goals	Loss of academic freedom
Increased instructional support	Increased workload
Increased opportunity for Individualized instruction	Restriction of curriculum
Administrators	
Increased funding	Need for more funds
Renewal of trust in schools	Law suits
Information for program evaluation	Poor publicity
Clear demonstration of need	Need for more staff
Clearly articulated goals	Increased public pressure
Parents	
Literate children	Increased taxes
Information about child's progress	Denial of diploma to child

Ways to increase the probability of positive outcomes and decrease the probability of negative outcomes (not sure of the source, but like the process):

Planning the process
- Warn in advance of test use
- Involve relevant groups in planning – including parents
- Preview results
- Allow for variety to item types and testing alternatives

Selecting test content
- Ensure face and curricular validity
- Review for relevance and lack of offensiveness
- Include open-ended responses for item types that cover a broader range of skills
- Do not limit tests to basic recall or low-level skills

Setting Standards
- Provide for remediation
- Focus on minimums – but allow for measurement error
- Involve relevant groups
- Use a variety of sources of data

Verifying quality
- Select item difficulty
- Calculate reliability of classification
- Determine face and divergent validity

Using the tests
- Administer tests early to provide skill information
- Provide remedial help
- Allow multiple opportunities to take the test
- Allow for alternative methods of assessment
- Maintain test security
- Monitor outcomes – do longitudinal follow-up
- Ensure tests do not become a mandate for grade retention
- Train educators to understand test scores and how to interpret them

Other considerations involving legal implications:

(Note that the following comments are adapted from Merle Steven McClung, educational law consultant and staff attorney for the Center for Law and Education in Cambridge, Massachusetts)

Be sure your competency program is phased in gradually over a sufficient period of time to allow both teachers and students to adjust to new requirements.

Be sure that the tests you are using actually test what the schools have taught.

Make sure the impact of the testing program is not disproportionately heavy on one or another racial or cultural group. Any functional competency test, and the curriculum on which the test is based, should reflect all aspects of our pluralistic society – or at least the extent of diversity reflected by the student population.

Satisfactory test performance should be just one minimum standard, to be used in conjunction with other criteria (course credits, absenteeism, etc.) in determining eligibility for graduation.

There should be representative community-based participation in making of decisions about promotion/graduation requirements, since these decisions obviously involve many hidden assumptions about educational goals, performance levels, grouping or tracking, discipline, etc.

What Else Can We Do?

Diversify! In some states, it is essential to use the test that has been designed, used and accepted by the federal establishment. That is a given and apparently will not include NAEP testing. However, it will be very difficult for some states to accomplish everything the state or feds want to see in one test. It may also be very shortsighted for some school districts or individual schools to adequately use this instrument to assess some local instructional goals that they feel are also important. An excellent alternative would be to design a matrix of assessment tools to satisfy local needs that would include the state test as a part of that matrix. Some possible alternatives would be, 1) subject rating scales, 2) performance assessment instruments, 3) student portfolios, 4) written essays and oral presentations, 5) senior/student projects, 6) extracurricular activities, 7) volunteer activities, and 8) honors, AP, awards, etc. I cannot imagine a state department of education refusing to accept this type of additional information, if not passing one multiple-choice item test was the case.

Since so many of the high-stakes tests, and certainly the college entrance exams, are loaded with multiple-choice test items, it would be helpful to continue to use these occasionally in classroom testing. It might be a good idea sometimes to question students as to why they answered a multiple-choice test item incorrectly. However, the constructed-response type (open-ended) items at the local level will certainly provide far better reliable and valid information.

Charles J. Clock, Ed.D
Educational Evaluation, Measurement, Statistics

cdclock@adelphia.net
(208) 777-8083
Post Falls, Idaho

June 22, 2006

Useful References:

Bloom, B.S., Madaus, G. F., and Hastings, J. T., Evaluation to Improve Learning, New York: McGraw-Hill, 1981

Hambleton, R.K. & Swaminathan, H., Item Response Theory - Principles & Applications, Boston, Kluwer Nijhoff Publishing, 1985

Kohn, Alfie, The Schools our Children Deserve, New York, Houghton Mifflin, 1999

Lemann, N., The Big Test - The Secret History of the American Meritocracy, New York, Farrar, Straus and Giroux, 1999

Orlich, D.C. & Gifford, G., "Test Scores, Poverty and Ethnicity: The New American Dilemma", Paper to be presented at October 20, 2006 Phi Delta Kappa "Summit on Public Education", Washington, D.C.

Popham, W.J., TESTING! TESTING! What Every Parent Should Know About School Tests, Boston, Allyn and Bacon, 2000

Sacks, Peter, Standardized Minds, Cambridge, MA: Perseus Books, 1999

Wiggins, Grant P., Assessing Student Performance, San Francisco: Jossey-Bass Publishers, 1993