The ESP Journal
Extraordinary insight into today's education information topics

Test Critics Fail the Test

Critics of Testing Don't Understand the Basics of Testing

Politimetrics of Accountability

By Glynn D. Ligon, Ph.D., ESP Solutions Group

ESP Solutions Group

Table of Contents

The Gauntlet

Critics of testing students don't understand the basics of testing. We let critics get away with bogus arguments that undermine the benefits of testing our students. Parents are misled into opposing a unique source of information about their schools—and their children. Worse, some opt their own kids out of a valuable validator of their academic progress.

Critics of state tests are doing parents and educators a disservice. I trust the critics are merely misinformed; however, their attacks are often simply not based on fact. The news media validates the critics without benefit of having a basic background in testing. The state and district testing staffs have taken such politically cautious stances that they too seldom speak as advocates for the tests they are hired to administer and interpret. I venture to say the state and district test directors agree with me that the critics are off base most of time. I don't know why we feel obligated to state our few agreements with critics' tangential points before we begin destroying their numerous and overwhelming false premises.

I'm taken aback by four observations.

- Too few professionals are taking up for the tests.
- The critics are getting away with their misrepresentations and recasting of the issues.
- School accountability systems are being undermined.
- The states are trying to do too much with their state proficiency tests.

What's needed in this debate is an unbiased, informed perspective. I no longer have a stake in this. I'm a former teacher, a former test director, and a former parent of public school students. I still have a Ph.D. in measurement and have read all the criticisms of testing. I constantly talk with parents who believe the criticisms of testing. I read the news articles about state testing and accountability.

So, here I go. I'm taking a "let's get this debate centered on the issues and facts" position.

The attack on state tests is akin to Clark Kent being bullied on the playground as a kid and not being allowed to use his powers to defend himself. Somehow, it has become politically impolite to correct or challenge the test critics without first having to agree with one of their marginal points. The test pros seem to feel obligated to begin their

response by agreeing with the test critics' red herrings that make them appear to be legitimate defenders of our schools, students, and tax payer dollars. Sorry, critics. I'm not doing that. Not being a public employee, nor representing a testing company, I'll say what should be said.

Testing, 1, 2, 3, Is Anyone Really Listening?

The purpose of this paper is to defend state proficiency testing of students. No one appears to be doing that. The critics of state testing seem to have a free run at the issues.

Critics say teachers don't get enough useful information from state proficiency tests. Sorry, critics, that's not why state proficiency tests were mandated. They were authorized for accountability.

Critics say state proficiency tests take too much time from instruction. The Council of Great City Schools documented state proficiency tests to take 2.3% of the school year. The great majority of the testing time that is lumped together and criticized is actually designed to provide teachers instructional feedback for individual students—or chosen by the students/parents for advanced placement, college entrance, gifted programs—or used by the schools for eligibility for services to address special needs, language, etc.

Critics urge parents to opt their children out of testing. Their advice is to give up the best objective, comparable, confirmatory measure of their child's performance and their school's success (or lack thereof). Some critics even recommend subjective, more expensive, and locally based alternatives.

Critics berate state proficiency tests for characteristics that the test developers have long ago remediated, often by using teachers themselves to address (e.g., removing item bias, mapping to curriculum, measuring higher order thinking skills, etc.).

Critics claim low-income and minority students and schools score the lowest on state tests. The fact is that individuals in those groups also score at the highest levels and make the largest gains. Those groups and schools may also make some of the largest gains within a district or state.

Critics claim that misuses of state proficiency tests will go away if state proficiency tests are not given. Some of their "misuses" are actual purposes of accountability (e.g., improving or closing underperforming schools, identifying poor teachers, etc.).

Critics complain that state proficiency tests are high stakes. Yes, they are. Accountability is high stakes.

When No Child Left Behind was passed, and state proficiency testing went nationwide, a fundamental mistake was made. Instead of simply fulfilling the mandate of NCLB by creating pure state proficiency tests, many states created hybrids. These hybrids do give proficiency scores for students. However, attempting to squeeze double benefit for teachers and satisfy others who railed against the increased testing burden, states promised to return skill-level feedback. That decision changed the psychometrics of the hybrid tests entirely. That decision made these hybrids less effective for both purposes.

The resolution is straightforward.

- Testing directors, superintendents, school board members (local and state), legislators, and professors need to step up, explain, and defend their state proficiency tests.
- States should have one test solely for accountability. That test should not try to give teachers any feedback other than a proficiency level and a growth score. That is the **State Proficiency Test**.
- Schools should have totally different tests to provide teachers diagnostic and prescriptive feedback. These tests might also be mandated and administered by the state—for diagnostic and prescriptive purposes to assist teachers. Those are the **Formative Tests**.

The time and cost for state proficiency tests are worthwhile for their accountability value. The time and cost for the formative tests can be debated by the teachers and test critics all they want to determine the balance between testing and instructional time.

We can resolve the current debate by focusing state proficiency tests only on proficiency.

For now, let's answer the test critics with facts.

So, here we stop and make a significant distinction. One that frames the debate and doesn't allow the critics to win by obfuscation.

"State proficiency test," means an accountability test in a major subject area (e.g., mathematics, science, English/reading/language).

- A state proficiency test is given for the sole purpose of determining the academic proficiency of a student (e.g., Proficiency Level: below basic, basic, proficient, advanced).

- The state proficiency test contains items that sample across the full range of the curriculum to give a reliable total score.
- Then by counting the students at each level, the state determines a school's accreditation rating.
- In addition, the scale score on the state proficiency test may be used in a growth model to measure how much a student gained.

Here are Some Facts.

First, state proficiency tests are for accountability—not for teachers to plan instruction, not to give teachers a profile of each student's skills. The fact that teachers don't find state proficiency tests useful is NOT a valid criticism. Yes, state performance test scores take weeks to get back to schools. Some don't get back until the next school year. However, those scores are not intended for classroom instruction; so, they are not "late" for teacher use.

Psychometrically, a test item and a test that are good for accountability aren't constructed for formative, diagnostic, prescriptive, student-level reporting for teachers. There's more on this later.

Second, state proficiency tests are NOT the only factor in accountability. State proficiency tests alone don't cause schools to close or teachers to be fired, or students to be retained in grade, or students to fail a course, or students to fail to graduate. There's typically a fail-safe mechanism in place that adds a human factor to override the test. (Note to critics: Of course, the humans don't have to override the test.)

Take away this most objective indicator, the state proficiency test, and the accountability critics will move on to attack the next one (maybe dropout rates or graduation rates).

Third, state proficiency tests don't take too much time from instruction. Proficiency testing isn't new to schools. In many districts, the state proficiency tests replaced the local annual standardized, norm-referenced achievement tests given before No Child Left Behind mandated state-administered tests. State proficiency tests answer crucial questions decision makers must know. The huge majority of "testing time" during the school year is spent on interim (e.g., formative, diagnostic) measures for teachers, eligibility tests (e.g., gifted, magnet, advanced placement, college entrance, etc.), and placement tests (e.g., special needs, language needs, etc.)—not state proficiency testing for accountability.

Fourth, state proficiency tests measure the curriculum the state mandates each student must know. The fact that passing rates on state proficiency tests are surprisingly low shows that many schools are not wasting time teaching students skills they already know. If schools are teaching skills to students that have already mastered them, that is a failure of the school, not the test.

Fifth, alternatives proposed by critics to current state proficiency tests (e.g., performance testing, portfolios, and other "authentic" methods) do not save teachers time; take away less instructional time; nor provide less expensive, objective, reliable accountability for parents, the public, and legislatures.

NCLB Didn't Begin the Criticism of Tests.

The critics didn't emerge in 2001 with NCLB. A 1991 article titled "Putting the Standardized Test Debate in Perspective" (Blaine R. Worthen and Vicki Spandel, Educational Leadership, 1991) identified these seven criticisms.

1. Tests do not promote student learning.
2. Tests are poor predictors of individual students' performance.
3. Test content is often mismatched with the content emphasized in a school's curriculum and classrooms.
4. Tests dictate or restrict what is taught.
5. Tests categorize and label students in ways that cause damage to individuals.
6. Tests are racially, culturally, and socially biased.
7. Tests measure only limited and superficial student knowledge and behaviors.

The authors' rejoinders were straightforward.

1. They offer us comparability that we couldn't get without them. Are 3rd graders learning basic math? Can 6th graders read at the predefined level of competence?
2. On their own, tests are incapable of harming students. The way in which their results can be misused is potentially harmful.
    a. Using the wrong test
    b. Assuming test scores are infallible
    c. Using a single test score to make an important decision
    d. Failing to supplement test scores with other information
    e. Setting arbitrary minimums for performance on tests
    f. Assuming tests measure all the content, skills, or behaviors of interest
    g. Accepting uncritically all claims made by test authors and publishers
    h. Interpreting test scores inappropriately
    i. Using test scores to draw inappropriate comparisons
    j. Allowing tests to drive the curriculum
    k. Using poor tests
    l. Using tests unprofessionally

The authors made one interesting characterization of the critics of the time. "Further, most critics are beginning to acknowledge that abolishing testing would leave us with many decisions still to make—and even less defensible bases on which to make them." Too bad today's critics have not captured this insight.

Critics' Recommending the Wrong Tests.

Here's an example of how critics misunderstand tests so badly that they recommend the exact type of test against which they are protesting. Texans Advocating Meaningful Assessment makes these recommendations. Each is annotated to explain how they somehow don't understand what they are demanding.

- Replace STAAR with meaningful student assessments that provide timely and useful feedback with no high stakes.
    - In grades 3-8, use assessments that provide diagnostic feedback in a timely manner to gauge how children are learning. National tests, such as Stanford, ITBS, ACT Aspire, are cost effective and proven, age-appropriate and meet federal requirements.
        - NOTE: The tests named (other than Aspire) are outcome tests, however, not diagnostic. That's why they do meet the federal and state mandates that the critics are protesting.
    - In high school, in lieu of STAAR EOCs, administer nationally recognized assessments, such as SAT or ACT, including one science test. Such proven tests are actually used by colleges and can show aptitudes for career choices.
        - NOTE: Does this reduce testing burden? The end-of-course tests in some cases replaced final exams. Adding the SAT/ACT for all students is admirable, but a different concept all together.
    - Attaching high stakes to standardized tests by requiring a certain score for grade promotion and graduation leads to teaching to the test and other corruption of classroom learning. Texas should not attach high stakes to standardized tests.
        - NOTE: Accountability is high stakes. If the test measures the Texas curriculum, then teaching to the test is teaching the curriculum.
- Limit standardized tests to no more than required by federal law.

    - NOTE: A clear definition of a standardized test is needed. Do they want to eliminate the tests that diagnose individual students' preparation for the state proficiency test? Do they want to eliminate the tests teachers use to diagnose and prescribe for instruction?
- Eliminate field test questions on high stakes exams.
    - NOTE: This would make tests unnoticeably shorter for students already in a testing mode. This is psychometrically the best method for ensuring valid and reliable items for future tests. A less desirable but necessary alternative would be for some students to take a longer field test to calibrate items.

Critics Use Hyperbole to the Extreme.

I won't cite the author and give him a reference. These actual quotations from his writing show how extreme and emotional the arguments of some critics stretch. Parents hear this and only need to sense the urgency in these tirades to begin thinking there's something drastically wrong with our tests.

- Because students know that test scores may affect their future lives, they do whatever they can to pass them, including…taking performance drugs (e.g. psychostimulants like Ritalin "borrowed" from their friends).
- Standardized tests don't value creativity.
- Standardized tests don't value diversity.
- Standardized tests occur in an artificial learning environment: they're timed, you can't talk to a fellow student, you can't ask questions, you can't use references or learning devices, you can't get up and move around. How often does the real world look like this? Prisons come to mind.
- Standardized tests reduce the richness of human experience and human learning to a number or set of numbers. This is dehumanizing.
- Standardized tests weren't developed by geniuses. They were developed by mediocre minds. One of the pioneers of standardized testing in this country, Lewis Terman, was a racist (the book to read is The Mismeasure of Man by Stephen Jay Gould). Another pioneer, Edward Thorndike, was a specialist in rats and mazes. Just the kind of mind you want your kid to have, right? Albert Einstein never created a standardized test

(although he failed a number of them), and neither did any of the great thinkers of our age or any age. Standardized tests are usually developed by pedantic researchers with Ph.Ds in educational testing or educational psychology. If that's the kind of mind you want your child or student to have, then go for it!

- Finally, my most important reason that standardized tests are worthless: During the time that a child is taking a test, he/she could be doing something far more valuable: actually learning something new and interesting!

Again, these criticisms fail to address the purpose or true psychometric properties of a state proficiency test. His "most important reason" would ring hollow to every store that takes time out for inventory each year, every employer who pulls an employee out for an annual evaluation and training, or every professional taking periodic certification exams.

Does it Matter What We Call Them?

Test is the name this paper chose because that's still what the newspaper headline writer uses. Parents understand test. Test is not at all pretentious or difficult to understand. Years ago, educators switched to assessment to change the connotation from multiple-choice, paper and pencil to be broader, more inclusive of performance, observation, and other grading options. The National Association of Test Directors even changed their name to the National Association of Assessment Directors. Professions use exam for their certifications, e.g., medical exam, law school exam, college entrance exam. College professors, even the test developers themselves, use the term measurement to be more inclusive of anything that provides a metric.

Hybrid State Test

On a true state performance test, there's no determination of mastery of individual skills or knowledge of objectives. That's because there are not enough items measuring a single skill to provide a reliable score for that skill. Thus, there's no individual student skills profile for a teacher.

Oh, yes, states have strayed from this, haven't they! Bowing to pressure and not understanding the psychometrics required to have the best state proficiency test, they try to create hybrid state tests. These hybrids don't sample so broadly across the curriculum. They sample a few skills each year with more items on each skill. They sacrifice validity (i.e., measuring the entire curriculum) for appearing to be "formative" tests for teachers.

However, they have so few items for each sampled skill that their reliability for each skill is low.

Still, we must include these hybrid state tests in the category of state tests. As a rebuttal to the state test critics; however, these hybrid state tests provide teachers diagnostic, instructional value.

Alternatives and Other Indicators

Some educators dislike state performance tests simply because they want to control accountability. An independent, objective, comparable state performance test is out of their control. To support that point, look at some of the alternatives, called multiple indicators, they want instead.

- School climate surveys
- Passing/failing grades
- Discipline referrals
- Parent surveys
- Classroom observations
- Performance measures graded by the teacher

I've never understood why so much weight was given to indicators that correlate with tests, meaning, predictors, rather than using the outcomes (the state proficiency tests) themselves. Let's distinguish outcomes and correlated predictors such as attendance/absences, teacher/staff turnover rate, grade retention rates, etc.

Some legislatures and local school boards are bowing to the pressure to expand their accountability systems. The tests are now only one of a growing number of indicators of performance. In some cases, the entire accountability process can be overridden by local, subjective input.

If that's what the parents want, then they are getting it. However, what the accountability systems originally intended to do was to give parents an objective, external measure of a school's performance—beyond all the subjective indicators they could already get locally if they wanted them.

How Did We Lose That Vision?

Simple, the state got off track with their test, and the test critics haven't been countered effectively with the facts.

State performance tests provide unique insights into the performance of our schools.

The loudest complaint about state performance tests is that they don't give teachers anything useful for instruction. They shouldn't. Critics of state performance tests in accountability systems say that they focus instruction on the state's curriculum objectives. They should.

Somehow, we let these two misguided criticisms frame the discussion instead of the original, unique, and necessary purpose for state proficiency testing.

State proficiency tests arose when schools got more money and the state wanted to measure the return on its investment. The state wanted to know which schools were delivering effective instruction to students.

No matter how many committees have been formed and how many times a study has been conducted, there are only a precious few direct indictors of student success. Let's make that only a precious few objective indicators of student success.

- Graduation rate
- Promotion rate from one grade to another
- Credits earned by age
- Success after graduation (must be defined)
- State proficiency test scores
    - Growth scores

No, report card grades, teacher credentials, discipline rates, school climate, demographics, per-student spending, participation, etc. are not objective student outcome measures.

School boards, legislatures, and Congress legitimately asked for evaluation of the investment they made in the schools—and the public trust in them to deliver quality education. They looked for something objective, affordable, and comparable across schools and districts. But first, they identified four basic questions. These questions looked at OUTCOMES.

- Are students on pace to graduate?
- Are students learning the skills and knowledge required to be successful when they graduate?
- Are students graduating?

- Are students successful after graduation in college or the workplace?

Notice, these questions don't include whether parents are happy with their schools, if students feel safe, if teachers are qualified, if attendance is high, how much money is spent, how much time is devoted to instruction, etc. Those would have been contexts, inputs, or processes. Those are only correlated to success, not direct measures of success.

Oh, yes, one might argue that schools don't get assigned students of the same starting performance levels. True. So, let's allow in our accountability system for academic growth. However, we won't adjust for factors such as income, race/ethnicity, or gender. The only predictor allowed is entry test score. Then every student is measured compared to all others starting at the same performance level in our growth model. That's fair—and supported by research.

In reality, growth only indicates progress for the individual student if the growth exceeds that of similar students. Also, the goal for a student is always to attain an actual performance level that is acceptable.

Growth scores can be aggregated for schools to determine below, average, or above expected levels or targeted goals.

Testing Time in the Schools

Protesters against the amount of testing in the schools probably are not aware of two significant facts.

> First, state proficiency tests make up a small fraction of the testing they criticize. There's more depth to testing in schools than they acknowledge. The Council of Great City Schools surveyed their members to document 8.6 "state" tests annually, which includes the hybrid tests and others far beyond what this paper defines as the mandated state proficiency tests for accountability. ("Student Testing in America's Great City Schools: An Inventory and Preliminary Analysis," October 2015)

Even with this broader brush, their timing came to the equivalent of 4.2 days or 2.3% of the school year for testing.

> Second, once students graduate, there's even more testing. In the real world, we take tests for additional education, certification tests to be employed, to stay employed, to perform

services, to enjoy privileges, etc. Testing is an omnipresent activity within our society that maintains our standards for health, safety, and quality in almost every aspect of our lives. To exempt schools from testing for accountability is inconsistent with everything else we require.

Performance Vs. Formative Tests

Accountability tests are designed to measure a performance level—pass/fail, or maybe basic, proficient, advanced. The tests that do this well have items that divide test takers into proficient and nonproficient. Test reliability is determined by how precisely the test measures pass/fail, the same way, each time a person takes it.

Now a formative test, the one teachers find useful, the one state test critics say they want, to be useful, has to do the same job for individual skills. That means having a mini-accountability test for each skill to be measured. So imagine that the formative test has to have a sufficient number of items for each skill it wants to measure. Now there's the difference, a formative test can only reliably measure a limited number of skills because the test can't be too long.

A state's accountability test can't be a formative test because its job is to measure broadly the state's curriculum. To be useful to teachers—reliably—it would have to be way too long, have way too many items for each skill.

A formative test can't be a good accountability test because it doesn't cover enough skills across the entire curriculum.

Oh, yes, many states try. They sample skills each year, each grade level. Then they give teachers reports for individual students to plan instruction. They sacrifice having an excellent accountability test. They sacrifice having an excellent formative test. They practice politimetrics.

Simply put, we need two different tests. One for accountability. One for instruction.

When critics denigrate our accountability tests for not helping teachers plan instruction, they simply don't know about which they are talking. They often are practicing two devious debating tactics.

- Pandering to an uninformed constituency on an emotional issue

- Arguing that the target of their attack doesn't do something it wasn't intended to do in the first place.

Let's start at the real beginning and ask the questions that everyone wants answered. See Figure 1. Isn't it interesting that we aren't all asking the same questions? Doesn't it now make sense that it takes more than one type of test to satisfy everyone? Using the earlier distinction between accountability tests and formative tests, Figure 1 shows which one answers each question. Notice I didn't say which one answers each question "better."

This isn't sitting down to a seven-course meal and not knowing which piece of silverware to use when you see 12 pieces laid out on three sides of your plate. You can still eat your food, and it tastes the same. This is a doctor using the right X-ray, MRI, EKG, etc. The doctor gets back very different data for very different purposes, yielding critically different insights.

Performance tests deliver reliable accountability data.

Formative tests were historically under-valued and under-funded in education. Formative assessment is what really helps teachers focus their instruction on students' immediate needs.

A major complaint educators have about statewide accountability assessments is that they make poor formative assessments for teachers. True, but the real problem is we can't seem to let accountability assessments simply do their job without faulting them for not being formative assessments as well. We should all be demanding separate assessments—one designed to be an excellent accountability measure, and many designed to be excellent formative assessments. But no, educators who disagree with the money and time invested in accountability measures have lobbied politicians to stretch the use of those assessments beyond the capability of a well-designed accountability test.

There is plenty of money to have two separate assessment programs—one to rate schools and one to diagnose and prescribe instruction. Plenty of money if we automate test administration, scoring, and reporting. Plenty of money if we apply extreme security and confidentiality standards only to the accountability assessments, not to the formative assessments.

po-lit-i-met-rics \p•ˌli-t•ˌme-triks\ n pl but sing or pl in constr (ca. 1972) 1 : the quantitative study of political groups, institutions, nations, and international systems 2 : statistics and indicators that are determined by a combination of scientific, mathematical, and political processes 3 : the art or science of determining high-stakes measures and criteria for accountability, esp. in the field of education

What do the words decimated, income tax rate, and proficiency level have in common? These are all terms derived through a combination of political and psychometric decision making. Politimetrics are used to determine each.

Decimated refers to drawing lots to select one in ten soldiers to be executed. While the measurement of one in ten is rather precise, the setting of 1/10th as the cut point was rather political—enough to make a point, but not too many to wipe out a useful unit. The income tax rate is set mathematically to generate a target revenue, but the rate is also politically determined by a vote of Congress and a signature from the President—to curry favor or avoid retaliation by the voters. The determination of proficiency levels on an assessment is informed by a projection of how many students will perform within each level, but ultimately a political body adopts the official cut scores.

Separating psychometrics, accountability, and annual objectives for adequate yearly progress from the political context within which education lives is impossible.

Some significant politimetrics of our time are:

- 100% of students proficient by 2014
- The National Assessment of Educational Progress' (NAEP) standard for being proficient rather than basic
- Criterion scores for eligibility for Title 1 services
- Formula for calculating a dropout rate
- Average daily attendance (rules for excused absences, tardies, etc.)
- Persistently dangerous school
- Highly-qualified teacher
- Percent of students by race/ethnicity
- Age requirement to enter kindergarten
- Percent expenditures on instruction
- Income guidelines for National School Lunch Program eligibility

ESP Insight
Politimetrics didn't work well when Congress decided 100% of students must be proficient by 2014.

The governing body exercising its politimetric responsibilities may be a local school board, the Office for Civil Rights, a state legislature, a school parent advisory committee, a state school board, or Congress. The result is that the comparability, validity, and reliability of our education statistics are susceptible to politics. Many of us have worked hard to raise the level of data quality within the education statistics arena. However, a major component of quality is definition—especially setting a standard and the process for measuring that standard. Policy and politics play a significant role in data quality and our perception of data quality in education's metrics.

Let's revisit the list of education politimetrics and rate each by our level of confidence in them.

I made up these ratings for the sake of discussion. Agree or disagree with these ratings, the fact is, politimetrics are simpler for some measures and certainly some politimetric decisions have more face validity than others. One could argue that anything rated above 50% in Figure 1 may be over-rated.

Figure 1: Confidence Levels in Education's Politimetrics

Politimetric	Perceived Confidence by Educators
100% of students proficient by 2014	0%
Formula for calculating a dropout rate	5%
Persistently dangerous school	20%
Percent expenditures on instruction	50%
Average daily attendance (rules for excused absences, tardies, etc.)	60%
Percent of students by race/ethnicity	65%
Highly-qualified teacher	70%
Criterion scores for eligibility for Title 1 services	75%
Income guidelines for National School Lunch Program eligibility	80%
NAEP's standard for being proficient rather than basic	90%
Age requirement to enter kindergarten	95%

Origins of Politimetrics

At one time, a brief time, I thought I had created the term politimetrics. However, credit goes to Thomas Gurr ("Politimetrics: an introduction to quantitative macropolitics," Prentice-Hall, 1972). He thought of politimetrics more as statistics about political entities. My notion of politimetrics is more as the artful combination of psychometrics and policy—how we arrive at tolerable criteria for accountability.

In this paper, I've expanded the term even more to encompass the whole arena of tests, the standards they measure, the rigor they impose, and the uses to which the scores are applied—appropriately and inappropriately.

In the process, I've challenged the establishment and those people who have become established in the testing and accountability world. So why not start with someone who has become one of the most respected authorities in the testing and evaluation field for education. Dr. Eva Baker, past President of the American Educational Research Association and Director of the Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Eva Baker doesn't seem to understand accountability because of politimetrics. Politimetrics has worked to influence her phrasing if not her thinking. Being a leader of CRESST and as past President of AERA, she must be ever vigilant to the politics of psychometrics as presented by the pending decisions of Congress, the rambling priorities of a major education research association, and the leanings of her in-crowd. She does not have the luxury of seeing accountability from a simple, clear perspective. She appears to be obligated to couch every thought she issues publicly inside a complete context of political propriety (aka correctness).

ESP Insight

Politimetrics for education is more the artful combination of psychometrics and policy—how we arrive at tolerable criteria for accountability.

This paper sets out to do one simple thing—separate accountability from all of the politically proper, politically expedient, politically encrusted context that tests and accountability have accumulated around them. This paper intends to call for state proficiency tests for accountability to be singular in design, purpose, and use. When we say accountability proficiency test, we should think of only one thing—a standardized test that provides a measure solely for the evaluation of student learning progress. In the process, a compelling case will be presented in support of true formative data and assessments.

As I listened to Dr. Baker's Presidential Address at the 2007 AERA Annual Meeting, I made notes on the five "accountability fixes" she proposed. I found it interesting to read her "expanded version" entitled "The End(s) of Testing" available on CRESST's website. The fixes are now six, and they are called "mitigations" and "tactics."

The tone of the expanded version is much less critical of accountability than was the address. Attachment A is an open letter to Dr. Baker written after the address. Do I really think Eva Baker doesn't understand accountability? No. That's why the title

Of my original response said "Why Eva Baker Doesn't Seem to Understand Accountability." She understands accountability, but like so many others in education, she uses the term rather expansively. She's managing the politimetrics. She's trying to squeeze formative dollars out of the accountability budget. I chose her so this letter could argue the opinions and conclusions rather than the agreed-upon facts or Dr. Baker's qualifications. I also know she has a bigger platform if she chose to respond.

The main point of the open letter was simply to say that accountability testing is different from formative assessment. They have different purposes, different psychometric requirements, and different policy foundations.

ESP Insight
Education needs to separate accountability from all of the politically proper, politically expedient, politically encrusted context that assessments and accountability have accumulated around them.

The distinction between politimetrics and political correctness is important. A politimetric decision may or may not be politically correct. The decision is politimetric because it is a compromise between the statistical or psychometric factors and the political ones. In this context, political means policy in general more than government in particular. Politimetric decisions can be correct without being politically correct.

In 2000, when I was consulting with the Governor's Office in Colorado during the creation of their school accountability reports, the question arose of where to set the dividing lines for CSAP (Colorado's statewide assessment) performance between school ratings. Initially, Governor Owens wanted to use A, B, C, D, and F, but eventually agreed to descriptions (excellent, high, average, low, and unsatisfactory). Today in 2018, several states have adopted the A-F ratings.

ESP Insight
NCLB was the piñata hanging from every AERA meeting room chandelier.

As the Governor's policy advisors and members of the Legislature debated the relative merits of various methodologies, I asked "How many schools can Colorado tolerate being unsatisfactory? How many schools will the public accept as being excellent? The answers were 8%

excellent, 25% high, 40% average, 25% low, and 2% unsatisfactory. With the policy determined, the psychometrics, statistics, and mathematics of establishing the rules were straightforward. Yes, in year one the cut points were arbitrary (based upon actual performance of all schools). Critics complained that the system was normative, a pejorative word used to discredit accountability systems deemed as dooming a set percentage of schools to failure.

The reality is that after year one, any number of schools could be rated excellent or unsatisfactory as they changed their performance. That accountability system remained in use for over eight years without substantive modifications. A major reason for its persistence was the face validity of the published schools' ratings. The creative blending of psychometrics, statistics, and policy resulted in an accountability system that worked. Subsequent governors and legislatures have had their opportunities to apply their own politimetrics to the next generations of accountability rules.

Is the infusion of politics into accountability anathema to valid ratings? Not at all. In fact, without the balance of policy makers in the design, accountability systems would be inflexible, statistical theories. Data-driven decision making (D3M) isn't just about the numbers. When policy decisions are made, the facts are balanced with the politics. National politics may be over-sensitive to the political dynamics with all the polling that goes on before Congress or presidents and candidates claim their policy ground.

Growth Models

ESP Insight

A major reason for the persistence of Colorado's school accountability report cards is the face validity of the published schools' ratings.

Growth models represent another generation of politimetrics in education. The idea is very simple—recognize schools making gains on assessments. The implementation has become very complex. Hierarchical linear models (HLM), which few educators understand and most statisticians I have met trust too blindly, are being touted as the most sophisticated way to tease out gains. With the error measurement of tests, the mobility of students, the small cell sizes for subgroups, and the resistance of student performance to rapid/sustainable improvement, HLM frequently splits hairs as it combs through data to find statistically significant differences that translate into tiny practical advantages.

The preceding statement was a blatant generalization that does not recognize the existence of clear academic gains within effective schools. The admonition in the statement is for us not to get our hopes too high

for what growth models show. Many schools full of low-performing students are really ineffective with academically disadvantaged students. Many schools full of high-performing students are restricted in how high they can perform because of assessment ceilings. Many—not all. The pursuit of those exceptions is both noble and necessary. Even for those low-performing schools that achieve miracles with their students, the ultimate goal doesn't change. Maybe they still need assistance to reach that goal. The prime objective of No Child Left Behind and most state accountability systems was to establish a goal line that is the same for all students regardless of how unlevel their playing fields are.

When considering growth models, the measure of growth must be as objective, numerical, reliable, valid, and comparable as possible. Again, we get back to needing a true accountability measure for the task. Formative assessments have a single shortcoming related to growth. They are most useful when they focus on a limited number of specific skills and objectives about to be taught. This characteristic makes them poor measures across grade levels and school years. An accountability test should measure skill and objectives across multiple years to avoid floor and ceiling effects—and to fit the assumptions of emerging growth models.

Contrasting Accountability and Formative Assessments

Reading through Figure 2 makes one wonder why anyone ever tried to make accountability and formative assessments the same. The reason is simple actually. Educators always want to squeeze every ounce of utility out of their efforts.

Extracting formative data out of the accountability turnip is understandable. Unfortunately, accountability proficiency tests are not up to the formative task.

Politimetrics has loosened the focus of accountability state proficiency tests by pandering to the proponents of formative assessments. What makes a good formative assessment does not make a good accountability test. Does your state's assessment go on the long list of those that would be more precise measures for accountability if they had not been developed to also provide objective-level proficiency scores for individual students?

ESP Insight

Educators always want to squeeze formative data out of the accountability turnip.

About 1990, Darvin Winick, now Chair of the National Assessment Governing Board, led a study group in Texas to recommend the next generation of state proficiency tests for accountability. As a member of that group, I recall agreeing with Dr. Winick that a nationally standardized test best fit the requirements for a single measure for evaluating the achievement of Texas students. In addition, we agreed that this test would never satisfy the need for formative, diagnostic data for teachers, so a separate diagnostic test should be developed aligned with Texas' curriculum standards. Those discussions and the insights about separate measures have remained valid through today. Unremarkably, Texas' TAKS became another generation of state assessments that try to be both accountability and formative measures at the same time.

Figure 2: Contrasting Formative and Accountability Assessments

Accountability Assessment	Formative Assessment
Representative items from across all knowledge and skills	Selected objectives representing knowledge and skills to be taught now
About 50% items correct by average student provides maximum measurement precision	About 75% correct by a proficient student provides expectation of success on the accountability assessment in the future
More total items on assessment for reliability of the overall proficiency level of the student	More items per individual objective provides confidence in diagnosis of areas in need of instruction
High security and confidentiality to protect the integrity of the test items and the results for individual students	No need for security because the whole idea is for teachers to use the items on demand
Scheduled administration times or windows for comparability	On demand administration to coincide with instructional planning
Timely scoring and reporting for decision making	Immediate scoring and reporting for diagnosis and prescription
Fresh items with only a few reused for alignment and equating	Reusable items for next groups of students as long as alignment with standards is maintained; released items from accountability assessment
Major concern about cheating	No concern about cheating; no incentive for teachers or students to cheat
Content measured is the same for all students	Content measured is what each student needs at the moment
On-line administration supports security and lowers costs	On-line administration supports the on-demand nature of formative assessment and lowers costs
Vertical scaling desired for measurement of growth	Measurement of current status on objectives for diagnosis
Politimetric establishment of cut points for proficiency	Teacher decision of cut points for prescription of interventions

Standards-Based Grading

Enhanced formative assessments are basic to the adoption of standards-based report cards for parents. In a previous Optimal Reference Guide (Using Assessment Results to Get Performance Results, 2006), Dr. Evangelina Mangino reported that parent reports from accountability assessments are too hard to understand, and that scale scores are meaningless to teachers because they do not provide a context for interpretation. "Teachers do not use (accountability) test reports as much as they might because they are overwhelmed by the quantity and complexity of the reports. Training on the vast scope of these reports is not realistic. There needs to be a better targeting of the really useful information to teachers in a simpler format and at the best time."

ESP Insight

No need to worry about security and cheating on a formative assessment. That alone saves dollars and time.

This challenge has been addressed by ESP in the development of Ed-Fi dashboards for standards-based grading. Ed-Fi is an interoperability standard for connecting data across sources. Displaying state performance test results for individual students, schools, and districts in dashboards in a useful interface addresses this for teachers and other educators. Displaying formative assessment data in a timely manner from assessment vendors is the key to use for decision making in the classroom, the school, and the district.

That is a difficult challenge for state performance tests. That is the prime objective for formative assessments. In fact one of the final recommendations of Dr. Mangino's study was to support on-line diagnostic testing.

Conclusion

Stand up to the test critics. Tell them we must have accountability. There is no better accountability tool than a state proficiency test.

Support our teachers with a set of comprehensive formative assessments.

Those are two different things.

Some of our states, led by some of our professors, educators, and legislators, have given us form accountability instead. Form accountability doesn't do either formative assessment or accountability well.

Attachment A – The Original Open Letter to Eva Baker

Dr. Baker,

I enjoyed your 2007 AERA presidential address--except for the part where you suggested "accountability fixes."

The real world, Congress, state legislatures, and the public are serious when they criticize education for being reluctant to be accountable. We must be cautious when suggesting moving from tests to softer, subjective "accountability" measures.

In your keynote address, you laid out five "accountability fixes" for the No Child Left Behind Act. Unfortunately, you could not have been more wrong about what needs to be done for accountability. Simply put, most of the fixes do not belong in accountability. They belong in a school improvement system. The distinction between the two escapes most educators. Of course, the purpose of accountability is to verify that the resources being invested in education are delivering the expected benefits--successful schools. This is very different from telling schools where individual students need to improve.

As I listened to your address, I reacted to each of your proposed accountability fixes.

- Fix 1: More Indicators

We get very confused by having more and more indicators to interpret. The fact is NCLB already mandates multiple indicators rolled into a single adequate yearly progress rating. As long as additional indicators get combined into a single rating rather than present a confusing and conflicting array of separate indicators, this is a great idea. Otherwise, when it comes to indicators, the more the murkier.

- Fix 2: Opportunity to Learn

Opportunity to learn is a process indicator, not an outcome indicator. We will not be satisfied knowing whether or not students were taught, we want to know if they learned. States should definitely monitor opportunity to learn as part of their overall implementation of NCLB.

- Fix 3: Performance Assessment

Have we already forgotten that performance assessments withered as accountability measures because they are too costly, unreliable, and rater- biased to be practical? Beyond limited constructed-response

items, writing samples are the signature survivor of performance measures within statewide assessment systems.

- Fix 4: Formative Assessment

Wait a minute. Aren't these accountability recommendations?

- Fix 5: Prioritized Standards

Great idea—for formative assessments like James Popham advocates. However, we should be expanding the scope of the content of our accountability assessments.

The message in your fixes is that accountability should be more like formative evaluation. You continued with a call for an accountability system that leads to instructional decisions. You criticized current accountability systems as having an absence of feedback for teachers. This is formative not accountability assessment. This just isn't a reasonable expectation for an accountability system. We should define accountability appropriately and narrowly. We must accept the expense and burden of accountability. We can then construct accountability assessments that measure a broad range of knowledge and skills. We can have more affordable, shorter forms that are tightly aligned with the full core academic standards.

In the process of creating a truer accountability solution, we should keep a focus on the need for formative data.

First and foremost, we need two different assessment programs—one for accountability and one for formative decisions. Your fixes perpetuate the same mistake NCLB codified in 2001. They call for accountability results to be useful to teachers. This is not likely to happen and sets up accountability assessments to disappoint teachers. I'm also critical of the policy makers for buying into the notion that if the tests are not useful for teachers to plan instruction, then they are failures.

I see that we are trying to satisfy everyone with a single assessment and accountability system. What we need is to satisfy the accountability requirement. Then we need to have a separate, differently designed and crafted system for formative evaluation.

To say the NCLB's accountability can be fixed by making it into formative evaluation is just wrong. Accountability can be fixed by separating it cleanly from the formative evaluation process. Then we can set about to build the infrastructure and processes to do formative evaluation and assessment right. The scope of NCLB is far beyond accountability. Formative goals fit, but formative goals are not accountability fixes.

Accountability assessments are like stock prices for a corporation. There is an incredible array of components that can be analyzed to discover what went right or wrong with a corporation and how to improve, but the accountability function is not tasked with that diagnosis and prescription. The stock price is not very helpful to management and workers to design improvements, but it is the essential way to value the worth of a company. Shareholders are not satisfied to know that there was an "opportunity to earn," or that performance evaluations were high for all employees, or that the corporation focused on a smaller set of standards for the year. They want a higher stock price.

Want a sports analogy instead of a business one? A professional baseball team is ultimately judged by its won/lost record or by championships won. A .333 won/lost record doesn't tell anyone what needs to be done to improve, but it is a clear accountability measure. Separately, management (or the fans) must analyze RBIs, ERAs, LOBs, BAs, and HRs. If you have no idea what those are, that's fine, because you know .333 is bad. Let the fans argue the statistics and management rebuild the team.

OK, so here's the education example. Parents see that their school has a 33.3% proficiency rate. Bad. They won't be satisfied knowing that their children had an opportunity to learn 90% of the standards, or students averaged over 85% on formative assessments, or the teachers reported student performances to be acceptable on report cards. Parents know there's something wrong. Policy makers know that the school must improve. That's accountability.

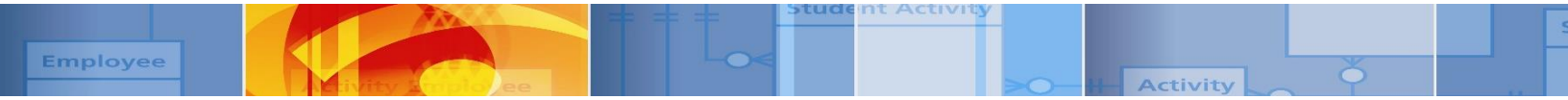
What to do is the next step after accountability. If you want to roll all assessment together into a complex system of "form accountability," that is wrong. Instead, we need to separate them even more. Formative assessment—accountability assessment. Two different types of tests.

How can that be simpler?

I am ready to support an increase in formative information for teachers. Doing that requires information systems changes far beyond over-analyzing accountability test results.

Sincerely,

Glynn D. Ligon, Ph.D.



About ESP Solutions Group

ESP Solutions Group provides its clients with *Extraordinary Insight™* into P20W education data systems and analytics. Our team is comprised of industry experts who pioneered the concept of "data-driven decision making" and now help optimize the management of our clients' state and local education agencies' information systems.

ESP personnel have advised school districts, all state education agencies, and the U.S. Department of Education on the practice of P20W data management. We are regarded as leading experts in understanding the data and technology implications of ESSA, SIF, Ed-Fi, ED*Facts*, CEDS, state reporting, metadata standards, data governance, data visualizations, and emerging issues**.**

Dozens of education agencies have hired ESP to design and build their longitudinal data systems, state and federal reporting systems, metadata dictionaries, evaluation/assessment programs, and data management/analysis and visualization systems.

To learn how ESP can give your agency *Extraordinary Insight™* into your P20W education data, contact us at (512) 879-5300 or info@espsg.com.

This document is part of *The ESP Journal* Series, designed to help decision makers analyze, manage, and share data in the 21st Century.

(512) 879-5300
www.espsolutionsgroup.com