Mid-Western Educational Researcher

Volume 24 | Issue 1 Article 7

September 2023

The Perfect Storm: How Policy, Research, and Assessment Will Transform Public Education

James H. McMillan
Virginia Commonwealth University

The Perfect Storm: How Policy, Research, and Assessment Will Transform Public Education

James H. McMillan
Virginia Commonwealth University

In the 2000 movie *The Perfect Storm*, as you are probably aware, there is an unusual convergence of several critical weather factors that set the stage for a destructive outcome that takes both property and lives. It has become a popular metaphor to describe how events come together in a unique way to have an exceptional influence on something, typically a negative impact. In education today there is also a perfect storm, one that won't affect property directly but will influence the lives of millions of students. It is interesting that Arne Duncan, the United States Secretary of Education, used the storm metaphor in a speech in June of 2009 (Duncan, 2009). Here is what he said:

Let me start by talking about the unique, historic, and powerful opportunity we have to transform public education. We have a perfect storm for reform: We have:

- The Obama effect;
- Leadership on the Hill and in the unions;
- Proven strategies for success; and
- The *Recovery Act* providing $100 billion.

Of course Duncan's remarks are not about anything destructive, unless you argue, like some have, that he is talking about the destruction of locally controlled education. He clearly thinks that the above factors are coming together in a positive way.

I want to focus on a different kind of perfect storm, one which is bringing several factors together that will create what I believe will be a destructive force for student learning. My contention is that there are three powerful influences that are coming together that will shape public education in the future—policy and politics, research, and assessment. What is argued is that we will soon have national standards, national tests, a national curriculum, and value-added teacher and school evaluation (see Figure 1).

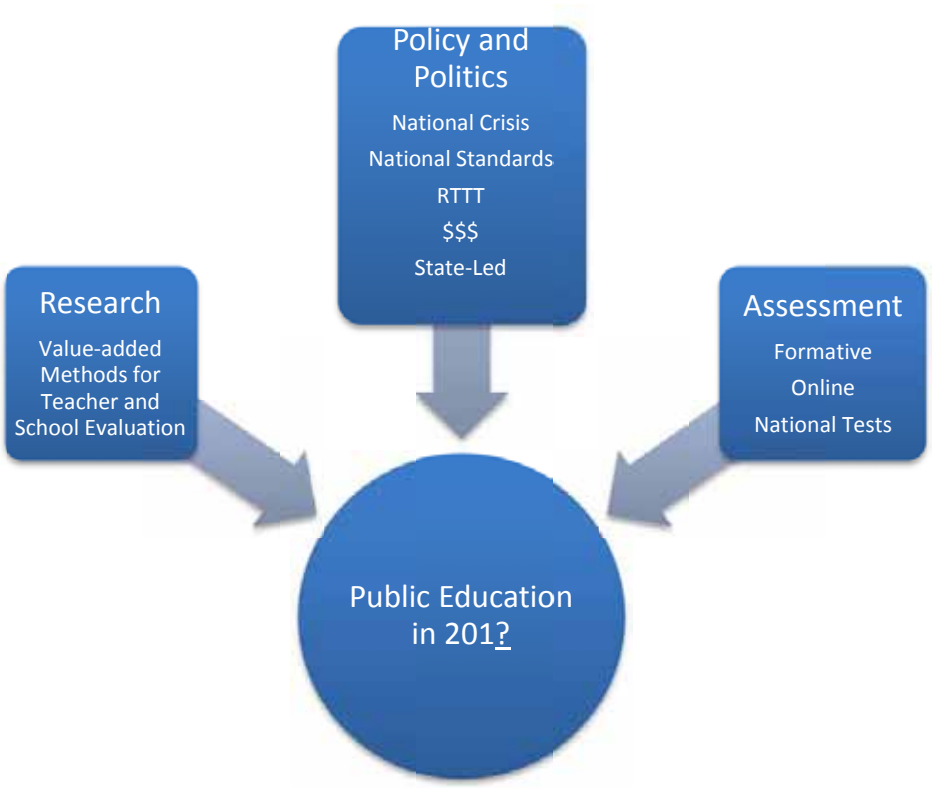

Figure 1. The Perfect Storm for Public Education

I believe the effect of these efforts is predictable based on previous experience; the effect on teacher evaluation is less clear. I will review each of these three factors, with a discussion of why they are detrimental, then list a few things we can do as educational researchers and assessment experts to mitigate the negative effects.

Policy and Politics

Here we need to return to national-level policy and politics. In that same speech last June, Duncan also made the following points:

- The genius of our system is that much of the power to shape our future has, wisely, been distributed to the states instead of being confined to Washington.
- Our best ideas have always come from state and local governments.
- On so many issues … the states are often leading the way.
- We think that every state should set internationally benchmarked standards and assessments that prepare students for success in the workforce and college.

This does not sound like anything that portends an increasingly federalized system of education. Indeed, the administration has repeatedly indicated that the effort to develop national standards is not a federal initiative, citing state-led efforts of the National Governors Association and the Council of Chief State School Officers to develop Common Core Standards. The Common Core website uses the phrase "Common Core *State* Standards Initiative" (emphasis added). We now have such standards in mathematics and English/language arts (without naming specific pieces of literature to be read), which have been adopted by 34 states plus the District of Columbia. The standards are supported by common sense (yes, it makes some sense to have the same learning standards for all students), as well as by statements from influential individuals and some research. Chester Finn, Jr. based recent comments on a study undertaken by the Fordham Institute (Carmichael, Martino, Porter-Magee, & Wilson, 2010). He recently said, "The United States is approaching a set of agreed-upon national standards of a core of its K-12 curriculum, and I think that's a healthy thing for the country" (Sawchuk, 2010). (Note his inclusion of both "national standards," in contrast to language used by federal agencies, and "curriculum").

The Fordham study has received favorable press with its conclusion that the Common Core standards in English/language arts are clearer and more rigorous than current standards in 37 states, and math standards in 39 states. On July 22, 2010, CNN based their article; *National learning standards make the grade*, in part on a favorable review of the Fordham study, saying that setting national standards is "gaining momentum according to an official of an educational think tank that compiled a national study comparing standards" (Holland, 2010). All the states were given grades (with few receiving a letter grade of A that reflected the highest score) based on content and rigor, and clarity and specificity. These judgments were made by only two language arts experts and three mathematics experts, not exactly what we would hope for in rigorous, systematic, and unbiased research. Nevertheless, this study is cited as evidence that the Common Core will raise standards in most states.

While the effort to develop the Common Core has been headed by the NGA and CCSSO, adoption of them has been encouraged, one could say, by federal rules that tie much needed money for the states to agreement to use the Common Core. Both the Race to the Top grant competition, a $4.35 billion pot of money, and Title I funding ($14 billion) tie chances of funding to adoption of the Common Core. Another $320 million pot has been awarded to two organizations to develop national tests of the standards (more on that later). Also, $250 million in the Recovery Act is for improving statewide data systems, and the budget of Institute for Educational Services (IES) has been increased more than $70 million from 2009. At the state level, data systems are being developed to track students and integrate resources such as teacher credentials and fiscal information with student outcomes. In a June 8, 2009, address to the annual IES research conference Arne Duncan said:

> We want to know whether Johnny participated in an early learning program and completed college on time and whether those things have any bearing on his earnings as an adult.
>
> Hopefully, some day, we can track children from preschool to high school and from high school to college and college to career. We must track high-growth children in classrooms to their great teachers and great teachers to their schools of education.

In other words, there has been an active federal role in promoting national standards and tests. It is a policy decision with clear consequences.

Another strong political factor is that the President has emphasized the interdependence between schooling and the economic recovery, without question a serious issue for all. In July 2010, President Obama emphasized that reforming education is the "economic issue of our time... It's an economic issue when we know countries that out-educate us today will outcompete us tomorrow" (Calmes, 2010). Thus, education is in crisis and needs to be fixed (not too different from assertions made to justify NCLB). This is further supported by international comparisons.

All of these factors suggest that we may be racing to adopt rushed reforms, without careful research to know what will happen to education when these reforms are adopted. To be sure, as indicated below, research is part of the picture, and here is one area that we need to have our voices influencing.

Research on Value-Added Models and Factors Influencing Test Score Variability

There is little doubt that value-added research models will be used to judge teacher and school performance. To a certain extent, the notion that teacher effectiveness can be measured by how much their students' scores improve by the end of the year makes sense and is easy to explain. In other words, how much have students learned in this class or school? The logic of this is compelling; why not judge teachers on gain scores, not according to the same set standards for all students? Wouldn't this be fair? Teachers would be compared on a more level playing field. Perhaps, but there are significant barriers to the use of value-added models.

The allure of value-added models is that factors such as family background, school resources, class size, previous achievement and a host of other variables can be used to isolate the effect of the teacher by comparing student expected growth (hopefully based on several years of data) to actual growth. But how this is accomplished is critical. The value-added model developed by Bill Sanders and used in several states has not been fully evaluated with an external review because part of it is "cloaked in proprietary secrecy"(Eckert & Dabrowski, 2010, p. 89). This lack of transparency and resulting appropriate external review is concerning, to say the least. It is related to another trend with value-added models. Some are developed by econometricians, individuals who can crunch numbers but may not have a good understanding of the nature of the data, limitations of the data, and consequences of reporting formats within school contexts. In the Value-Added Research Center at the University of Wisconsin at Madison researchers use the words "value-added *productivity*" (emphasis added), which suggests a business rather than education perspective. The models can be very complex and difficult to understand, and the manner in which results are reported is critical. In California, for instance, value-added scores for grades 3-5 were recently reported in the *Los Angeles Times* for the Los Angeles Unified School District; an economist and education researcher from the Rand Corporation did the analysis. Rank ordered results for every teacher (6,000 total) and school were included. The results were norm-referenced, so you could easily see how an individual teacher or school stacked up, and of course there had to be teachers at the bottom of the curve, no matter what improvement of scores. There are appropriate cautions about interpreting the results, including a statement that small percentile differences are not significant:

> Value-added scores are estimates, not precise measures, and readers should not place too much emphasis on small differences in teacher percentiles…both sampling error and measurement error contribute to the variability of the estimated teacher effects…the teacher's "true" rank falls in a range around each point estimate…the range of potential values for math was plus or minus 7 at the 20th and 80th percentiles (*Los Angeles Times*, 2010).

The problem is that the initial results were not reported with the standard error of measurement intervals, only one year of data was reported, and no other indicators of teacher effectiveness were included. Reporting data for the value-added system in Tennessee is obtuse and difficult to understand. Researchers at Vanderbilt University (NCPI, 2009) have used a simplified value-added model for linking student test scores with performance pay, but there is still a need to report results so that interpretations are appropriate.

Another consideration is how well value-added normative data fit with standards-based education. There is a clear record of research about the implications of norm-referenced evaluation. The logic of standards-based education, which has become the basic model of school reform, is criterion-referenced. But in standards-based models, student background is not controlled. If schools with high and low socioeconomic student populations show the same achievement, it is difficult to know if the standards are too easy, teachers in the low SES schools are terrific, or if teachers in the high SES schools are terrible. There is some development of status-based accountability based on test scores, as well as efforts to combine norm-referenced value-added data with status data (e.g., in Colorado) (Betebenner & Linn, 2009).

It will be interesting to gauge public reaction to reporting value-added results. In the September 2 issue of the *Wall Street Journal*, an editorial was titled "Teachers for Cover-ups." It targeted the Los Angeles teachers' union for objecting to the reporting of the scores, printing "Unions tell the *L.A. Times* to stop reporting test results." As could be expected, the *Wall Street Journal* defended the reporting of the scores and ranking. My hunch is that value-added results will be embraced by most non-educators and some educators, even with the caveat that standardized test scores signal but one of many important schooling outcomes, but we will see.

I believe there are several important issues with value-added models, beyond reporting of results, that need further research. One is preparing tests with sufficient "stretch" so that there is not a ceiling effect (Koedel, 2010), something that is common with standards-based assessments. This is needed to allow high scoring students room to improve. But to do this has obvious implications for the make-up of the test.

Another research-related factor to consider is how much teachers can actually influence the variability in student performance on standardized tests. Consider all the factors that influence student achievement on these tests that are outside the control of the teacher (e.g., general ability, native language, friendships, parental support, siblings, previous achievement, attendance, summer experiences, curriculum, district testing policies). This doesn't leave much that differences between teachers can influence. Schochet and Chiang

(2010) claim research has shown that 90% of the variability in student achievement is determined by student-level factors other than what the teacher can control (at the same time many claim that the teacher is by far and away the most important school-related factor to student achievement). Consequently, a limited amount of the remaining variability can be attributed to the unique contributions of an individual teacher (as differentiated from what any teacher provides). While value-added models help adjust for such differences, there is simply no way to fully account for these differences in a systematic manner.

Assessment

There are several developments in the assessment field that will fuel national assessments. These developments are driven by an unprecedented convergence of three factors: substantial federal funding, "voluntary" participation in determining common state standards, and advances in technology. Ironically, research on the impact of formative classroom assessment has generated interest in making large-scale tests more responsive to student learning and relevant to instruction. This is clearly reflected in the RTTT funding of $350 million in grants to support the development of a "new generation" of "multi-state" comprehensive assessments. Two groups have been funded with approximately $160 million for four years for development of the comprehensive systems (SMARTER Balanced Assessment Consortium [31 states] and the Partnership for the Assessment of Readiness for College and Careers [PARCC; 26 states]). The "new" assessment systems must go beyond summative assessment and include an integrated set of performance assessments, as well as interim assessments that are described as "through-course," accomplished during the school year. While this new emphasis on formative assessment is noteworthy and appropriate, it will be interesting to see how it can be achieved.

At issue is whether it is possible to use benchmark testing for what has been carefully and clearly defined as a process or series of steps used in formative assessment (Wiliam & Leahy, 2007; Brookhart, 2007; Popham, 2008). Consider the 2006 definition used by the Council of Chief State School Officers:

> Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes.

Note this definition includes *during instruction*, *providing feedback*, and *ongoing teaching*. These are characteristics not often associated with large-scale testing. Wiliam and Leahy (2007, p. 31) point out "a 'formative assessment' that predicts which pupils are likely to fail the forthcoming state-mandated test is not formative unless the information from the test can be used to improve the quality of the learning within the system." Popham (2008, p. 6) has recently made the same point in his definition of formative assessment, which emphasizes that formative assessment is a "planned process" in which evidence is used so that teachers "adjust their ongoing instructional procedures" or students "adjust their current learning tactics." It is assessment with these characteristics that, according to the research, improves student learning.

It seems to me that what is being proposed is quite different from what is defined as formative in the context of on-going instruction. I'm not sure what to call it to differentiate it from a more instructionally relevant definition. Maybe something like "quasi-formative" would work, or maybe such assessments should be called "summative/formative" tests since they look like mini-summative tests that can provide limited feedback to teacher and students. Maybe we will all be pleasantly surprised, but the task is daunting.

The difference between what the classroom assessment literature contains about formative assessment and these "new" assessments is important because the evidence that formative assessment makes a difference in achievement is based on the definition that includes on-going, feedback, and immediacy. Empirical evidence that formative benchmark testing has a positive impact on student learning is both limited and mixed. For example, some research suggests that targeted instruction can lead to improvements in student test scores (Lachat & Smith, 2005; Nelson & Eddy, 2008; Trimble, Gay & Matthews, 2005; Yeh, 2006) as well as proficiency in reading and mathematics (Peterson, 2007). However, empirical investigations that utilized quasi-experimental approaches have found no significant differences between schools using benchmark assessments and comparison schools not using such tests (Henderson, Petrosino & Guckenburg, 2008; Niemi, Wang, Wang, Vallone, & Griffin, 2007). There is also little evidence that interim tests can be used to determine whether students are on track to successfully complete the end-of-year assessment (Brown & Coughlin, 2007).

The rhetoric of "new generations" assessments is appealing with its emphasis on interactive assessment items that require "higher order" thinking skills and the use of artificial intelligence to score open-ended responses. Both proposals include the development of online digital resources to improve teaching and learning, including professional development materials, all aligned to national standards. There is even consideration of combining interim assessments with a year-end assessment to reach a final student score.

The list of objectives upon which the new assessments are based is impressive if daunting (Center for K-12 Assessment & Performance Management, 2010):

- Aligned with national standards.
- Lower cost (hence online tests).

- Formative as well as summative.
- Fast turnaround (hence online tests).
- Use of adaptive test delivery.
- Assessment of problem solving with multi-step simulations.
- Greater accommodations for students with disabilities and ELL students.

At issue with all of this is whether the new assessments will reflect older or newer research about how learning occurs and cognition. Traditional large-scale assessments tend to reflect learning theory that emphasizes fragmented knowledge and limited conceptions of cognition. More recent research on learning and cognition has emphasized the mental structures needed for problem-solving and the organization of knowledge so that it is useful. Knowledge is constructed and stored so that it can be easily retrieved, depending on context, the nature of the task, and previous learning. It is a matter of knowing when, where, and how to use knowledge, not simply demonstrating what is known and understood. Hence, students need to develop sophisticated understandings of how core concepts and explanations are applied to decision-making and problem-solving. Research on constructivism and learning progressions provides a basis for developing assessments on this more sophisticated idea of learning (Pellegrino, 2009).

Can the currently funded assessment development projects reflect more contemporary theories of learning and cognition? It will depend in large part on the nature of the standards that are assessed. The current plan to utilize through-course assessments throughout the school year is a step in the right direction, as is the emphasis on more constructed-response and performance assessments. It will be interesting to see if this emphasis reflects more recent learning theory or whether it becomes a series of mini-summative assessments, like what is now occurring with interim assessments.

Error (there is more than what you are led to believe)

There are several sources of error, both systematic and random, that must be considered for the next generation of accountability tests. For many years we in the research and measurement community have known about the deleterious effects of using standardized test scores to judge teacher effectiveness. One of the best insights was offered by Donald Campbell (1979). His conclusions have become known as "Campbell's law," and it is relevant for many fields, including education (Rothstein, 2008). This is what he asserted:

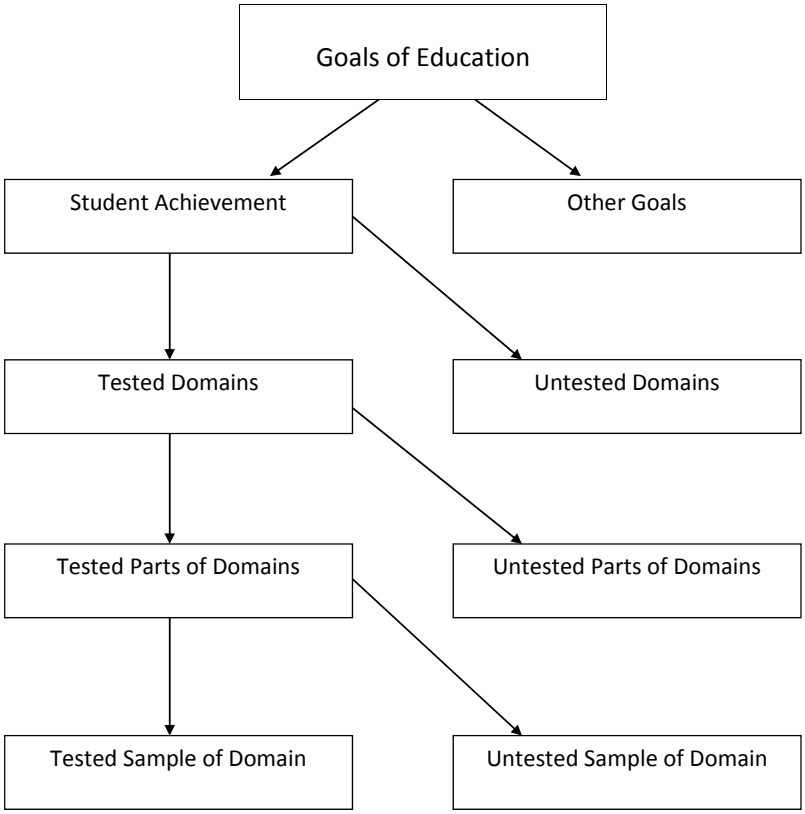

Figure 2. The Effects of Sampling That Narrow What is Tested (Adapted from Koretz, 2010).

> The more any quantitative social indicator is used for social decision-making [e.g., teacher effectiveness], the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor (p. 85).

Furthermore, he stated that:

> From my own point of view, achievement tests may well be valuable indicators of general school achievement under conditions of normal teaching aimed at general competence. But when test scores become the goal of the teaching process they both lose their value as indicators of educational status and distort the educational process in undesirable ways … Achievement tests are, in fact, highly corruptible indicators (p. 85)

A primary cause of Campbell's law is that there is incomplete and imperfect measurement of *desired* outcomes. The factor that makes this measurement incomplete is sampling error. The amount of sampling error is realized by considering all stages of sampling to get to the final test. Figure 2 illustrates these steps. At the outset only certain goals of education are selected, namely mathematics and reading/language arts achievement goals. Then achievement goals are limited to those domains of achievement that are sampled. Once domains are identified, parts of the domain are sampled, and then there is a sample of each part.

When you consider teacher effectiveness, similar sampling takes place, except now the achievement results, based on incomplete sampling, are used as an indicator of teacher effectiveness. In other words, only a *sample* of how "effective" the teacher has been is measured. There is error associated with sampling, and many desirable teacher benefits are not included (e.g., influencing a student to stay in school, developing a positive attitude toward reading, enhancing prosocial skills).

The sampling dynamic leads to the corruption of education by shifting resources allocated to tested subjects. Koretz and colleagues call this *between-subjects reallocation* (Koretz & Hamilton, 2006; Koretz, McCaffrey, & Hamilton, 2001), and summarize evidence to document the effect.

The amount of error that results from sampling must be added to two additional sources of error – measurement error and cohort effects. Measurement error is well described if under-reported. Typically a single source of measurement error is included, and that is most commonly internal consistency. Even high internal consistency reliability estimates, however, result in a fair amount of error in making final determinations such as pass/fail, or for teachers – adequate/inadequate. This is illustrated nicely with some data from the Virginia Standards of Learning test results. According to technical manuals, the overall amount of likely misclassification is typically about 10% for 5th grade math. There is a 4% false negative result, just attributed to the measurement error. If similar statistics result when making decisions about teachers (adequate/inadequate), 4 of 100 teachers could be unfairly terminated.

A recent IES report addresses misclassification error rates using value-added data when measuring teacher and school performance (Schochet & Chiang, 2010). Using simulations, they estimate that the total percentage of misclassified teachers using three years of data is about 26%. That is, about 26% are false positives and false negatives. One in four teachers are misclassified.

Cohort effects are very difficult to control. Obviously, in any given year a teacher may have more or less able and motivated students. Students seem to come together as a group in some classes but not others; some students "lose" more knowledge over the summer than other students. Teacher-student relationships vary. More students are absent for some classes. There is more in migration of students for some classes. Changing the criteria for student assignments to different teachers may be important. Every teacher knows that every class is unique, even if there is random assignment of students to each teacher. These factors are identified by Kane and Staiger (2002) as random differences across classrooms. Cohort effects are very real and are only adjusted by presenting many years of data, with the assumption that these effects eventually even out.

Another consideration that results in error in our conclusions about student learning and teacher effectiveness is the well-documented *test inflation* factor. Test inflation occurs when increases in scores do not match increases in actual student knowledge and understanding. As we have seen with NCLB, percentages of students judged to be proficient keeps climbing (though now we're seeing some ceiling effects). The question is whether the increase in scores is an indicator of student achievement or reflects on many factors that result in higher scores without the associated gain in achievement. This is essentially a validity issue. What inferences are appropriate about student learning?

Research on test inflation has documented large exaggerations of improving accountability test scores (Koretz, 2008). The best recent illustration includes examples of studies that show more improvement on state-level high-stakes test scores than on NAEP. For example, research on scores from Kentucky in the 1990s showed significant gains on the KIRIS over three years, with no improvement on NAEP. Similar patterns were found in Texas. But even standardized achievement tests many years ago showed test inflation when scores at the end of several years use of the same standardized test resulted in lower scores on the newly standardized version of the test (which then would show gradual improvement each year). What happens is that over time teachers focus instruction on what is on the test, use

classroom test items that are similar to what is used on the high-stakes tests (e.g., more multiple-choice items), tend to use test items that they remember from the high-stakes test, enhance students' test-wise capabilities, cheat, coach, read items to students, give extra time to finish the test, teach writing to be consistent with the scoring rubric, and excessive drilling on knowledge tested. The goal is higher test scores, not greater student knowledge, understanding, and problem-solving ability. Teachers may also focus instruction on "bubble" students, ones who are close to passing, with less emphasis for very high performing as well as very low performing students.

Three things seem inevitable – 1) there will be more testing; 2) the stakes will be higher; and 3) there will be greater standardization across states. This will inevitably lead to more test prep and teaching to the tests. The prospect of a school and state performing poorly on national tests will generate considerable motivation to do whatever is needed to improve test scores, leading to test score inflation and less emphasis on what is not on the test. The current considerable influence of test-based accountability on teaching and learning seems poised to become even more powerful. There will be significant pressure on teachers to focus on what is tested.

Surviving the Storm

The movement to national standards and tests is powerful. We are now desensitized to high-stakes testing and have the technical capability to use complex approaches to teacher and school evaluation. So if the "Perfect Storm of Reform" is coming, what can we do to minimize the destruction it could wreak on student learning? I believe the following are things we can do with assessment and research that can have a positive impact.

Assessment Development

- Get involved immediately in the construction of high-stakes tests to ensure that these tests are developed with sufficient attention to validity, reliability, and fairness (the three pillars of educational measurement), and that important, high-level standards for learning are assessed (e.g., inference, problem-solving, deep understanding). This should include developing tests that provide the correct types of evidence that can be shown to be appropriate for evaluating teachers. We also need to get involved with state tests and reporting options.
- Employ multiple methods of assessment, even if this is less cost efficient, perhaps on a sampling basis (matrix sampling).
- Emphasize the need for standards and tests to be compatible with contemporary learning theory.
- Become involved in state test design and reporting options.
- Monitor the integrity of data systems and encourage data that can examine trends over several years.

Evaluating Teachers and Schools

- Measure, "count," and report everything that is important in defining teacher effectiveness.
- Emphasize that value-added models of teacher effectiveness are at best only a general indicator of teacher effectiveness, and that more assessment may be warranted as a follow-up to verify and identify more specific areas of concern. There is error that needs to be accounted for, and using norm-referenced analyses may distort the differences between teachers.
- Report all important school data together, not just value-added scores, to provide context and a balanced perspective on school effectiveness. Context would include such "input" factors as student socioeconomic status, size, teacher characteristics, and special programs. Contextual information should also be presented in displaying teacher effectiveness data. Do not come up with single grades for schools.
- Consider results from a single test as an *indicator* or *snapshot* that requires further evidence.
- Combine value-added with status-based approaches.
- Monitor unintended consequences and factors influencing test inflation.

Reporting

- Report and explain confidence intervals and standard errors of measurement. These are not so technical that parents and others can't understand. The concept of margin of error is well understood once explained (hopefully, though this is based only my own experience).
- Avoid reporting of scores of small student subgroups.
- Avoid reporting of single year "growth." Use several years of data longitudinally to indicate stability over time.
- Be suspicious of large gains in any one year.
- Use plain, nontechnical language
- Present concise summaries.
- Utilize graphs and charts.
- Provide guidance for how to use the results.

Other

- Involve parents in the development, reporting, and use of assessment results.
- Conduct research on the impact of assessments on instruction and student achievement.
- Provide on-going teacher and administrator professional development to ensure accurate, uniform understanding of how to use results.
- Keep a close eye on econometricians and other quasi-educators.
- Use policy issues as examples in instructing preservice teachers and school administrators, and focus

professional development on assessment and research principles and issues that are critical in the appropriate interpretation and use of assessment data.

- Gather data that are locally relevant and meaningful.

Summary

In summary, bring on the storm! We are equipped and motivated to fight for what is right for our students and the system of education in our country. We can't be complacent during this critical time of establishing national standards and national tests. By understanding and communicating important principles of research and assessment we can work with politicians and others to influence policy. The next few years will be both exciting and daunting, but just as we tell our students to be engaged in learning, we need to be engaged in efforts to establish policy that will affect our profession and students.

References

Betebenner, D. W., & Linn, R. L. (2009). *Growth in student achievement: Issues of measurement, longitudinal data analysis, and accountability*. Paper presented at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda. Retrieved from http://www.k12center.org/publications.html

Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.). *Formative classroom assessment: Theory into practice* (pp. 43-62). New York: Teachers College Press.

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region*. Washington, DC: U.S. Department of Education, Institute of Educational Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Available http://ies.ed.gov/ncee/edlabs

Calmes, J. (2010). *Obama defends education program*. Retrieved from http://www.nytimes.com/2010/07/30/education/30obama.html?_r=2&ref=todayspaper&utm_source=Newsletter

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*, 67-90.

Carmichael, S. B., Martino, G., Porter-Magee, K., & Wilson, W. S. (2010). *The state of state standards—and the common core—in 2010*. Washington, DC: The Thomas B. Fordham Institute.

Center for K-12 Assessment & Performance Management. (2010). *Enhancing assessments to support teaching and learning: Next generation assessment systems proposed under Race to the Top program.* Retrieved from http://www.k12center.org/rsc/pdf/15051-K12Cntr-RTTTbro-Digital.pdf

David, J. L. (2010). Using value-added measures to evaluate teachers. *Educational Leadership, 67*(8), 81-82.

Duncan, A. (2009, June). Address to the Governors Educational Symposium. Retrieved from http://www.ed.gov/news/speeches/2009/06/06142009.pdf

Duncan, A. (2009). *Robust data gives us the roadmap to reform*. Washington, DC: U.S. Department of Education.

Eckert, J. M., & Dabrowski, J. (2010). Should value-added measures be used for performance pay? *Phi Delta Kappan, 91*(8), 88-92.

Henderson, S., Petrosino, A. & Guckenburg, S. (2008). *A second follow-up year for "Measuring How Benchmark Assessments Affect Student Achievement"* (REL Technical Brief, REL 2008-No. 002). Regional Educational Laboratory Northeast & Islands.

Holland, S. (2010). National learning standards make the grade. Retrieved from http://www.cnn.com/2010/US/07/21/national.learning.standards/index.html?hpt=C2

Kane, T., & Staiger, D. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16*(4), 91-114.

Koedel, C., & Betts, J. (2010). Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy, 5*(1), 54-81.

Koretz, D. (2008). Measuring up: What educational testing really tells us. Cambridge, MA: Harvard Press.

Koretz, D. (2010). *Implications of current policy for educational measurement.* Princeton, NJ: Educational Testing Service.

Koretz, D., & Hamilton, L.S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531-578). Westport, CT: Praeger.

Koretz, D., Mccaffrey, D., & Hamilton, L. (2001). *Toward a framework for validating gains under high stakes conditions* (CSE Technical Report No. 551). Los Angeles: University of California, Center for the Study of Evaluation.

Lachat, M. & Smith, S. (2005). Practices that support data use in urban high schools. *Journal of Education for Students Placed at Risk, 10*(3), 333-349.

Los Angeles Times (2010). *Los Angeles teacher ratings: FAQs & About*. Retrieved from http://projects.latimes.com/value-added/faq/#difference_value_added_api

National Center on Performance Incentives (NCPI). (2009). *Frequentluy asked questions: About our work in metropolitan Nashville Public Schools*, retrieved from www.performanceincentives.org/about_ncpi/faq_nash.asp

Nelson, M. & Eddy, R. (2008). Evaluative thinking and action in the classroom. In T. Berry & R. Eddy (Eds.). *Consequence of No Child Left Behind for Educational Evaluation: New Directions for Evaluation, 177*, 37-46.

Niemi, D., Wang, J., Wang, H, Vallone, J., & Griffin, N. (2007). *Recommendations for building a valid benchmark assessment system: Second report to the Jackson Public*

Schools. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Pellegrino, J. W. (2009). *The design of an assessment system for the Race to the Top: A learning sciences perspective on issues of growth and measurement*. Paper presented at the exploratory seminar: Measurement challenges within the Race to the Top agenda. Retrieved from: http://www.k12cwentere.org/publications.html

Peterson, J. (2007). Learning facts: The brave new world of data-informed instruction. *Education Next, 7*(1), 36-42.

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Rothstein, R. (2008). *Holding accountability to account: How scholarship and experience in other fields inform exploration of performance incentives in education*. Nashville, TN: Vanderbilt University, National Center on Performance Incentives.

Sawchuk, S. (2010, August 8). States rush to adopt common standards as RTTT content ends. *Education Week*. Retrieved from www.edweek.org

States rush to adopt Common Standards as RTT contest ends. *Education Week*, *August 11*, 8.

Schochet, P. Z., & Chiang, H. S. (2010). Error rates in measuring teacher and school performance based on student test score gains. Washington, DC: National Center for Educational Evaluation and Regional Assistance (NCEE 2010-4004), Institute of Education Sciences.

Trimble, S., Gay, A., & Matthews, J. (2005). Using test score data to focus instruction. *Middle School Journal, 36*(4), 26-32.

Wiliam, D., & Leahy, S. (2007). A theoretical foundation for formative assessment. In J. H. McMillan (Ed.). *Formative classroom assessment: Theory into practice* (pp. 24-42). New York: Teachers College Press.

Yeh, S. (2006). High-stakes testing: Can rapid assessment reduce the pressure? *Teachers College Record, 108*(4), 621-661.