

## Best Practices for Next-Generation Assessments

Malbert Smith III, Ph.D., President



MetaMetrics.

1250L

On Dec. 10–11, 2009, the National Academies’ Board on Testing and Assessment (BOTA) held a two-day conference on “Best Practices for State Assessment Systems.” The conference’s overarching theme was the need to move beyond current assessment systems in favor of more innovative approaches and technologies. Interest in this topic is being fueled by a number of trends, including the Common Core Standards Initiative; RTTT funding; concerns about international benchmarking and assessments; college and career readiness; the belief that NCLB assessments have narrowed the curriculum; and U.S. Education Secretary Duncan’s statements that *we can do better than the bubble sheet*. The collective force of these trends and the new funds committed to “innovative assessments” has created a *zeitgeist* for change.

The conference presenters did an outstanding job of painting a picture of the many issues that will need to be addressed, including the costs (one-time versus on-going); psychometric issues; the need for different item types; the tension between curriculum and assessment staff; and the role assessment systems should play in education. A recurring theme was to be mindful of past lessons to avoid the fate that some of our previous innovations have experienced.

Some 30 years ago, legendary assessment guru Oscar K. Buros reflected on the past 50 years of testing (Buros, 1977). His concern about the lack of progress made in the testing field was punctuated in the following: “If you would examine these books and the best of the achievement and intelligence tests then available, you might be surprised that so little progress has been made in the past fifty years—in fact, in some areas we are not doing as well. Except for the tremendous advances in electronic scoring, analysis and reporting of test results, we don’t have a great deal to show for fifty years of work. Essentially, achievement tests are being constructed today in the same way they were fifty years ago—the major changes being the use of more sophisticated statistical procedures for doing what we did then—mistakes and all” [p. 10].

OK, no pun intended, what major advances in testing have we witnessed since Buros’ critique over thirty years ago? Clearly, the testing field has advanced in many ways—with computer-adaptive testing, IRT models, latent variable theory and robust simulation models—as our computing power has exploded. Yet, many of the basic assumptions about how and when testing should be done, and the inferences we make from them, have changed very little. While testing is done much more frequently than 30, or even 80, years ago, the ultimate question that needs to be answered is “are we getting better information from the tests that we administer?” Unfortunately, I’m not confident that many of us would answer this question in the affirmative.

As we look back at these lessons learned, we need to look even further into the annals of history than Buros did. Consider, for example, some of the past measurement issues in the hard sciences. A major theme across the sciences is the acceptance and unification of common scales and metrics. In the late 1600s, there were literally dozens of instrument makers (think test publishers) claiming to measure temperature. Each had an eponymous scale that was correlated with the others (similar to the current state of educational measurement constructs). Regardless of whether these instruments used water or mercury, the real breakthrough in measuring temperature was the realization that our scientific

understanding of temperature could guide the instrumentation and acceptance of a common scale. This resulted in the acceptance of Daniel Fahrenheit's scale in 1714 and Anders Celsius' scale in 1744. And if the British and French could have sorted out their differences, today we would use a single scale.

When it comes to the measurement of reading, writing and mathematics, we have the same potential for unification as we had with temperature. The underlying measurement issues in education have nothing to do with cosmetic differences in item types, but, rather, in the towering Babel of too many assessments measuring the same construct on different and non-exchangeable scales. Therein lies the confusion on what proficiency means on NAEP versus a specific state test, PIRLS or other testing instrument. Just imagine the confusion in the healthcare industry if we had as many different metrics and scales for measuring temperature as we have for reading.

Unfortunately, today there seems to be more attention and debate on the item types within a test than on the construct being measured—the metric or the utility of the scale. The focus on item type probably is the result of falsely equating the item type with the construct of interest. Often people will look at the item type as illustrative of how the construct should be taught; practice on items like these to improve your reading ability. Item types can vary tremendously across many dimensions—from task complexity to costs to psychometric issues—but they still are just proxies for the measurement of the attribute or construct. We can have many different item types that measure the construct of “reading ability,” without privileging any one of these as the most relevant.

The first breakthrough in a new era of meaningful assessments rests upon the idea that reading, writing and mathematics can be measured on vertical (developmental) scales. As E.L. Thorndike stated, “Whatever exists at all, exists in some amount” (Osterlind, 1997). Constructs and measures can be explained using common, vertical scales which facilitate communication and clarity. A primary goal of education is to foster growth, and it is time that we routinely measure individual student growth.

The second breakthrough is predicated upon the premise that just like we can order students from low to high across the vertical scales of reading, writing and mathematics, we also can order instructional content along the same vertical scales.

Placing students and instructional tasks (e.g., readers and books) on the same scale enables us to bridge the curriculum divide. Test and resource publishers can link their products to these underlying scales in such a way that educators can connect assessment with day-to-day instruction in the classroom.

The de minimus psychometric standards of validity and reliability are prerequisites for any assessment. Innovative assessments can adhere to these same psychometric principles while making the following features possible:

1. Assessment and instruction can be usefully blurred, proving that it is possible to “mine the exhaust” of the instructional experience for assessment data as the student engages in instructional tasks. Assessment and instruction are two sides of the academic coin.
2. Computer-adaptive engines can be applied to instructional content, just as they are applied to the test item bank. Both the creation and delivery of content and test items are targeted to the individual.

3. Assessment engines can connect day-to-day progress with year-to-year summative tests by reporting on common developmental scales. Having multiple measurements on a common scale across time and various assessment instruments permits a more reliable and stable estimate of the learner's true ability. We can have more confidence in the inferences that we make about a student's current status and growth trajectory when we rely on multiple measures across the year, as opposed to a single administration of a high-stakes assessment.
4. Test items can be created "on the fly" as students interact with instructional content. Machine-generated test items can be created and discarded as needed throughout the experience of the student. The storehouse of value is in the underlying scale, not in a secure set of test items.
5. Scoring and reporting can be immediate for students, teachers, parents and policymakers. The learning experience and the assessment data mined from the experience is not constrained by calendar, time or location. Delivery is accessible 24/7 via the Web.
6. Perspectives and monitoring can be longitudinal across the developmental lifespan of the student for each construct. As LEA's move from K-12 to P-20 systems of accountability, the importance of optimizing growth for each individual student requires the monitoring and documentation of longitudinal data. Within these utilities, growth over the lifespan of the learner can be measured and expressed with unparalleled precision (Williamson, 2006).
7. The focus is "student-centric," as opposed to "teacher-centric." A student-centric approach breathes life and reality into the ideal of individual educational plans (IEPs) by paying attention to the critical components of skill acquisition: targeted practice, real-time corrective feedback, intensive practice, distributed practice and self-directed practice.

Innovative assessments built on these metrological principles unlock the realities that are manifested in the hard sciences. We can look at longitudinal data across the developmental span of the learner. By building formative and summative assessments on a common scale, educators, families and policymakers can connect day-to-day learning activities with year-to-year activities. The storehouse of value is no longer locked up in a specific item type but in the value of the scale. Items can be constructed through theory and on-the-fly through artificial intelligence.

An area that needs to be examined in more detail is the mode of assessment. Dirk Mattsen offered the example of assessment delivered through mobile devices and interactive games. Truly innovative assessments will blur the distinction between assessment and instruction. Today, by using existing technologies and frameworks, it is possible to mine the exhaust of an instructional experience and extract assessment data. Individuals who "play" a video game are being assessed in real-time, although they are generally oblivious to this assessment. As they get better, they progress to higher levels within the game. There is no reason why the same instructional and psychometric considerations can not be applied to the teaching and measurement of reading, writing and mathematics.

In conclusion, Stephen Lazar began his presentation with a quote from Ecclesiastes: "What has been will be again, what has been done will be done again; there is nothing new under the sun" (Ecclesiastes 1:9 New International Version). The statement certainly is true when we examine the educational and measurement issues confronting us today. Nearly 90 years ago during the dawn of educational measurement, there was tremendous optimism about the role of assessment. Wilson and Hoke wrote: "The college instructor blames the high school teacher, the high school teacher complains of the grade

teacher, each grade teacher above first grade finds fault with the poor work of the teacher in the grade below, and the first grade teacher in turn is chagrined at the shortcomings of the home training. Must this go on indefinitely? Whose opinion shall prevail? Is it not possible to get away from personal opinion to an agreed-upon consensus of opinion? May we not replace the constantly conflicting subjective standards with definitely defined objective standards?" (Wilson & Hoke, 1921).

This time, let's hope we get it right!

### **References**

Buros, O.K. (1977). Fifty years in testing: Some reminiscences, criticisms, and suggestions. *Educational Researcher*, 6(7), 9-15.

Holy Bible: New International Version. Authentic Media, 2008.

Osterlind, Steven J. (1997). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance and Other Formats (Evaluation in Education and Human Services)*. New York: Springer.

Williamson, G. L. (2008). A Text Readability Continuum for Postsecondary Readiness. *Journal of Advanced Academics*, vol. 19 (4), 602-632.

Wilson, G.M., & Hoke, K.J. (1921). *How to Measure*. New York: The Macmillan Company.

### **About the Author**

Malbert Smith III, Ph.D. is president of MetaMetrics, an educational measurement and research organization. MetaMetrics' renowned psychometric team develops scientific measures of student achievement that link assessment with targeted instruction to improve teaching and learning.



1000 Park Forty Plaza Drive, Suite 120, Durham, NC 27713  
[www.Quantiles.com](http://www.Quantiles.com)

MetaMetrics®, the MetaMetrics logo and tagline, Lexile®, Lexile Framework®, Lexile Analyzer®, the Lexile logo, Quantile®, Quantile Framework® and the Quantile logo are trademarks of MetaMetrics, Inc., and are registered in the United States and abroad. The trademarks and names of other companies and products mentioned herein are the property of their respective owners. Copyright © 2010 MetaMetrics, Inc. All rights reserved.