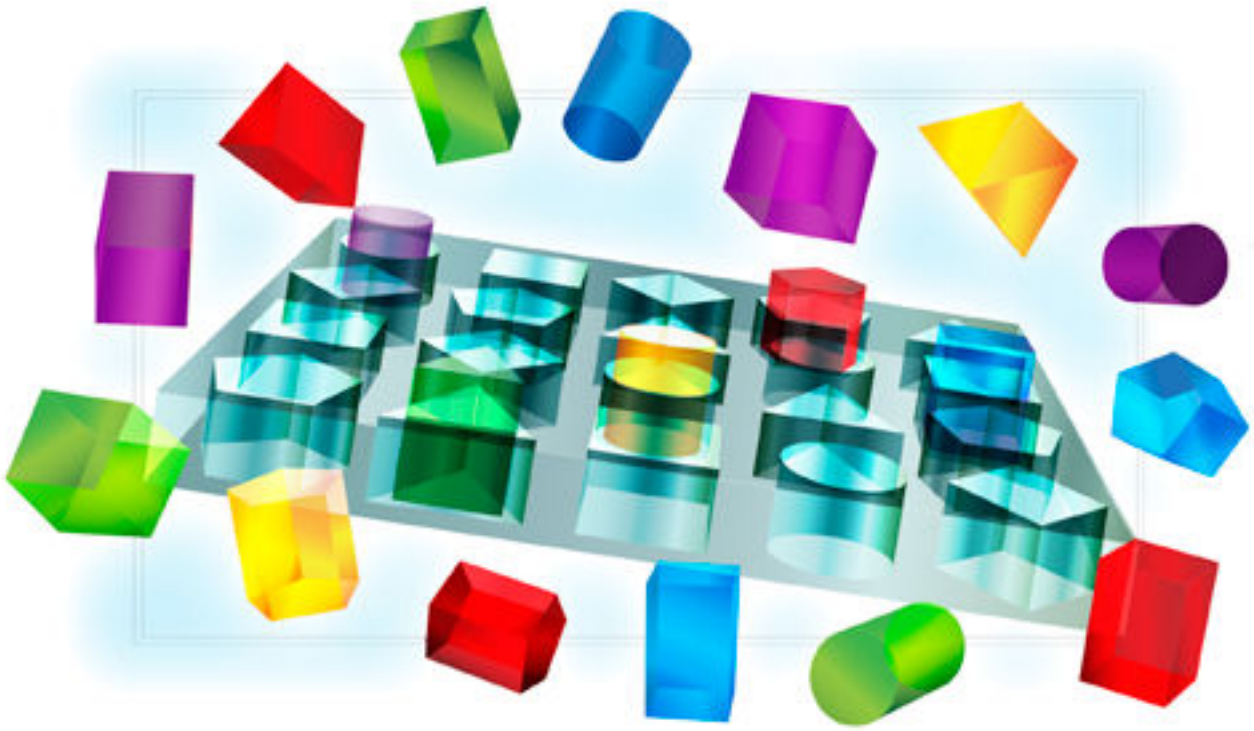
A LITERATE NATION WHITE PAPER

LITERATE NATION SCIENCE CORE GROUP—On the Reading Wars Fall 2013

Selecting Screening Instruments:

Focus on Predictive Validity, Classification Accuracy, and Norm-Referenced Scoring



by Steven P. Dykstra, Ph.D.

Literate Nation Science Core Group and Board of Advisors

Selecting Screening Instruments:

Focus on Predictive Validity, Classification Accuracy, and Norm-Referenced Scoring

by Steve K. Dykstra, Ph.D.

The goal of universal, early reading screening is to identify children at risk of future failure before that failure actually occurs. By doing so, we create the opportunity to intervene early when we are most likely to be more effective and efficient. Therefore, the key to effective screening is maximizing the ability to predict future difficulties.

Distinguishing Features: Predictive Validity, Classification Accuracy, Normative Scoring

Two qualities of a screening tool relate most directly to the ability to make useful and accurate predictions: *predictive validity* and *classification accuracy*. Predictive validity is a measure of how well the prediction of future performance matches actual performance along the entire range of performance from highest to lowest, not just at or near the cut score. It answers the question, "If we used this screener to predict how every child will perform at some point in the future, how good would those predictions be?" Classification accuracy answers the question, "If we used this screener to divide our students into those considered at risk and those considered not to be at risk, how well would we do based on the outcome of their future performance?" Classification accuracy is a measure of predicting into categories while predictive validity measures predictive accuracy over a continuous range of performance. Screeners with good predictive validity will almost always have good classification accuracy, but it is possible to have good classification accuracy with less robust predictive validity.

When comparing levels of predictive validity it is important to understand how the numbers work. Validity almost always is reported as a correlation coefficient, or *r* value. When comparing these values, it is helpful to square the values, yielding an r^2 value, also known as variance. This gives a more direct comparison of the magnitude of the predictive power of the assessment. For example, a predictive validity coefficient of .8 appears to be twice as powerful as a value of .4. In fact, if we square the values (.8 becomes .64, and .4 becomes .16) we see that the first assessment is actually four times more powerful in terms of its ability to predict future performance (64:16 = 4:1).

When comparing predictive validity it is important also to know what the screener predicted. Therefore, some assessment must be used as the benchmark of future performance. A screener that effectively predicts broad reading on a measure like the Woodcock Johnson is meeting a higher standard than a measure that predicts future performance on a brief, less comprehensive assessment. It also is true that valid predictions farther into the future are more difficult and can be evidence of a superior assessment.

Assuming a screener has good predictive validity and classification accuracy, it also is desirable for the assessment to report norm-referenced scores. Norm-referenced scores have been developed on large samples of diverse subjects and allow us to know how common or rare a score is. Norm-referenced scores allow us to compare scores on multiple assessments to properly judge whether we have a consistent picture of performance, or whether some of the scores are aberrant and may need special consideration. *Normative scoring* also gives us better ability to track performance over time. Without normative scoring we only know if a child scored above or below the cut score for being considered at risk. We do not know how far they may be above or below the cut score, how much that performance may have changed over time, or how it compares to other assessment data we may have on that child. Assuming the screener has good predictive validity and classification accuracy, normative scoring always is desirable.

Assuming a screener has good predictive validity and classification accuracy, it also is desirable for the assessment to report norm-referenced scores.

Reliability often is considered an important measure of the quality of an assessment. Reliability is a measure of the likelihood that if we gave the same assessment to the same child twice, under identical conditions, we would get the same results. It is certainly true that reliability is essential, but primarily in how it supports validity. All valid measures are inherently reliable, so we have assured ourselves of adequate reliability by demanding high predictive validity. Surplus reliability beyond what contributes to validity is desirable for progress monitoring, but does not make screening more effective. Predictive validity, including how far the screener predicts into the future as well as the quality of the measure being predicted; classification accuracy; and normative scoring are the major features that distinguish a superior reading screener.

We can review the quality of any screener by examining the features named above: predictive validity, classification accuracy, and normative scoring. Any screener worth considering will clearly report this data in a technical manual and should go into some detail about how the statistics were calculated. Data on many well-known and popular screeners have been collected by the National RTI center (http://www.rti4success.org/screeningTools). Unfortunately, much of the data is reported categorically rather than numerically, but it is possible to use this data to identify potential candidates for a screener then gather more precise data from technical manuals and other sources. Two screeners within the same category in the NRTIC table may still be very different from each other.

Review Options & Compare Screeners Without Emotion or Prejudice

Any group or individual choosing a screener is urged to make a complete review of their options and compare different screeners without emotion or prejudice. That process often is confounded by pre-existing notions of what a screener should look like or what it should include. Options are often rejected for no better reason than they do not look like what we are accustomed to or do not include some feature we may think is vital, even though the screener is measurably superior in every important way. For instance, some may prefer a screener that is timed while others may prefer a screener that is not timed. These are arbitrary preferences based on our personal impressions of what works best. We should rely on objective measures of what works best and make a careful comparison of the statistical details and qualities of our screening options, not our natural human biases and desire to use something familiar.

The most comprehensive evaluation of screening tools will consider the independence of their various subscales. It takes time to administer 5 different subscales that all yield different scores. That only is worth doing if the scales assess different skills. Ideally, all the subscales would have high correlations (e.g., around .5 or higher) with broad reading ability but relatively low correlations (e.g., .3 or lower) with other subscales. Regardless of the actual values, any comparison of different screening options should favor higher correlations with broad reading, and lower correlations between subscales. That would show that they measure vital but relatively unique aspects of reading, meaning each subscale tells you something important you did not know from the other subscales. That level of independence between subscales is very difficult to achieve and generally leads to very high predictive validity when it is accomplished.

Predictive Assessment of Reading (PAR): An Excellent Screener

As mentioned, any group or individual choosing a screener is strongly urged to investigate all of their options before making any final choice. They should consider the available data, as well as the robustness of the reported predictive validity: What does the screener predict and how far into the future can it make that prediction? Applying those principles, the science team at Literate Nation has been unable to identify a screener as good as or superior to the Predictive Assessment of Reading (PAR). PAR has superior predictive validity and classification accuracy, is norm referenced, and predicts performance farther into the future than is reported for any other screener.

Unlike other screeners, PAR uses a complex algorithm to make superior predictions of future performance. Composite scores made up of multiple subskills should have greater predictive power than individual scores. Most screeners form composite scores simply by adding subscores together, giving each subscore equal weight in the final calculation, if they form composite scores at all. PAR gives different weight to each subscore in their algorithm and changes the weights in the algorithm depending on age and level of reading development. This allows PAR to predict 1st grade performance from a kindergarten screening by giving different weight to the various subskills than would be used to predict performance in 3rd grade or 8th grade.

PAR also uses the same data used to produce the flexible algorithm to make instructional recommendations. PAR can accurately identify which of several deficient skills is most important right now, and give guidance on the intensity and duration of intervention that will be necessary to remediate it.

As a norm referenced assessment, data from PAR can be usefully compared to other assessments, and student performance can be tracked along the entire continuum of scores. This allows PAR to accurately identify gifted students and make instructional recommendations for them as well.

Other Screeners Also Are Worth Considering

Other screeners that should be considered include DIBELS (DIBLESNext), AIMSweb, and PALS, and the RAN/RAS, the classic naming speed tests. AIMsweb and DIBELS include the ability to progress monitor with very frequent probes of specific skills that assist teachers to direct instruction toward targeted areas of weakness in a student's profile. PALS also includes a progress monitoring tool known as a quick check. Progress monitoring is a critical function when implementing multitier system supports in general

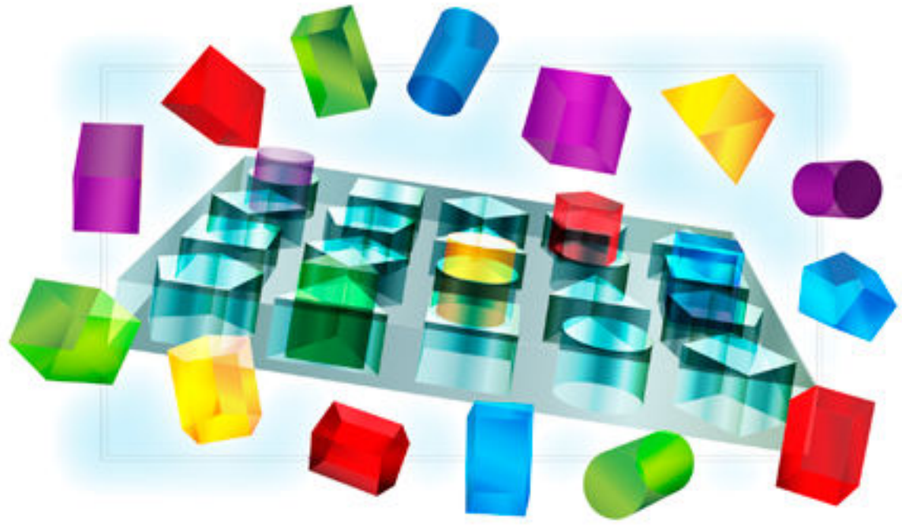
education classrooms. Only certain types of assessments can be given as often as progress monitoring sometimes requires. Any assessment plan must include progress monitoring and screeners that include a progress-monitoring component have that advantage. Users of PAR or other screeners may still opt for DIBELS or AIMSweb as a probe for progress monitoring. PALS may be a more comprehensive set of assessments, while DIBELS and AIMSweb are norm referenced and PALS is not.

The RAN/RAS tests represent one of the most important predictors of reading ability across every writing system tested in the last three decades. Naming speed tests provide a quick, easily administered measure of the brain's underlying ability to connect visual and verbal processes. As such, they give a very basic index of present and future issues related to word-retrieval processes and the development of fluency in reading. RAN/RAS is also an excellent example of a skill that both predicts broad reading and is independent of each other subskills. It contributes unique information to the screening data, not available through any other assessment. Many screeners use some version of the original RAN, including PAR, but often differ on: the nature and number of stimuli to name; the administrative procedures with which the norms were collected; or, the added dimension of retrieving names from different categories in the RAS. The extensive data collected on the 2005 version of the classic RAN/RAS, which now includes genetic and brain imaging studies, assures that these three dimensions are incorporated in this screener

PALS, DIBELS, RAN/RAS, and AIMSweb all have a longer track record than PAR. They have been used in more schools for more years and all of them generate useful data. How that data compares to PAR is a question that deserves careful consideration. Depending on past practices, some teachers or schools may be better prepared to make use of some data while requiring additional training to make full use of other assessments. Schools and districts with an established relationship to another screener may consider adding the RAN/RAS to other measure of phonologic processing and decoding in order to improve the range of critical skills including in screening. They could also add a simple picture vocabulary screening. The balance will always be between more information and the time it takes to gather it. However, it is good to note that in practice many schools are currently conducting assessments of multiple subskills that have very large correlations. They might do better to drop one or more of these assessments in favor of RAN/RAS, or picture vocabulary, which contribute important, unique data. At the very least, comparing all these screening assessment options puts each in useful context, and other assessments should not be dismissed simply because they are not mentioned here. Some may require more time or training to administer. Others may be statistically superior or less expensive. Some publishers may be better equipped to provide support and help plan implementation and new assessments worthy of consideration could appear on the market at any time. There are many issues to consider. However the basic advice to gather broad, useful information which improves our ability to identify who will struggle and why, and to do so as efficiently as possible, avoiding repetitive assessment which don't improve on what we already have, is rock solid.

Any individual, group or state education authority choosing a screener must gather their own data and make their own decision. They should ask hard questions of publishers and demand the best answers. An answer or marketing pitch that relies on emotion or suggests that some less significant feature of the test makes up for inferior statistical quality should be duly noted. Every claim for an assessment should be carefully investigated. Initial issues of training, support and familiarity may be solvable over time, but a statistically inferior assessment plan always will be so. While pragmatic concerns are real and must be considered, the first and greatest concern should be the quality of the screener in terms of predictive validity, classification accuracy, and norm referenced scoring.

Prepared for Literate Nation's State Coalitions
Primary Author: Steven P. Dykstra, Ph.D.
Secondary Authors: Maryanne Wolf, Ed.D., Susan Smartt, Ph.D.
Reviewed and approved by Literate Nation, Science Core Group



LITERATE NATION