Evaluating Early Literacy Assessments

As a part of efforts to support improvement of literacy outcomes for Iowa's Children, the Iowa Department of Education in collaboration with Iowa's AEAs commissioned a review of commonly used assessments for universal screening and monitoring progress of reading in four year old preschools through grade 6. The initial list of assessments to be reviewed came from a survey of Iowa schools and included all assessments used for universal screening and progress monitoring by at least two schools. Additional assessments known to be commonly used assessments for universal screening and progress monitoring were added to this list. All assessments on this list were subjected to an initial review for technical adequacy by DE and AEA consultants. All of the assessment authors/vendors were invited to provide new data during an RFI/RFP process as a part of the State's selection of assessments for the new MTSS/RTI data system. The results of the full review of assessments were summarized and shared with the field in April 2013 and may be found on the Iowa Department of Education Multi-Tiered System of Supports (MTSS) website (A Summary Report of Iowa's Review of PreK-6 Reading Assessments for Universal Screening and Progress). One outcome of this review process was the purchase of the Formative Assessment System for Teachers (FAST) battery of tests (as well as the Individual Growth and Development Indicators (IGDIs)).

Legislation funded at the end of the 2013 legislative session caused the Department of Education to revisit these reviews for the purposes of setting minimum standards for assessments approved for use in universal screening and progress monitoring of K-3 literacy. Chapter 62 rules required minimum standards for the following statistics: reliability and validity for all assessments, area under the curve and sensitivity/specificity for universal screening assessments, and number of equivalent forms and reliability of slope for progress monitoring assessments. See the appendix of this document for definitions of each of the required statistics.

The scoring rubrics used for the initial review process described above were examined to determine if the existing scoring for these statistics could be used to set the standards for K-3 literacy assessments.

The scoring rubrics used the following scale:

4	3	2	1	0
exceeds standard	desired standard	minimally meets standard	below standard	missing or unacceptable

Based on this scale it was determined that for an assessment to meet the minimum standard, it needed to earn at least a 2 rating on all of the required statistics based on the scoring processes used during the previous reviews. Next, each statistic was examined independently to validate the scaling of the rubric used in the initial review. All of the scales except for sensitivity/specificity were determined to be acceptable based on industry standards and practice. The sensitivity/specificity rubric scaling was adjusted to align with the ratings used for area under the curve (AUC).

Because many assessment systems employ more than one test to cover all grades and because tests are effective at specific grades the team decided to summarize each assessment at each grade, retaining all of the assessments that met the criteria at any grade. Furthermore, although the review

requirements are only for grades K-3, the assessments were also reviewed and summarized at grades 4-6. (Early Childhood assessments will be reviewed and summarized separately and are not included in the current document.) Any assessments that failed to meet the minimum criteria at all applicable grades was removed from the final list.
In order for schools to meet the requirements for using assessments that meet the minimum standards, the battery of assessments used for universal screening must include assessments meeting the standards at each grade from K to 3. The same is true for progress monitoring assessments. Ideally, this battery of assessments should be created to gather universal screening and progress monitoring information as efficiently as possible. Considerations should include the number of assessments, amount of time needed to administer the assessments, the costs of the assessments, both in actual dollars, but also in terms of issues such as training and related technology requirements.

FAST assessments
The FAST assessments are supported by the Iowa Department of Education and available to schools at no cost and are packaged to provide an efficient and reliable set of data using high quality assessments that meet the technical requirements of Chapter 62. If schools choose to use other assessments on the approved list to meet their K-3 literacy assessment requirements they are encouraged to review further information on these assessments to determine their individual merits. Not all assessments on the approved list are of equivalent technical quality.

In the table below are the FAST tests that met all of the assessment standards with a rubric score of at least a 2 (indicated by the check marks). The Iowa Department of Education has adopted the FAST suite of tests and will provide support to all Iowa schools who use these assessments for K-3 (and 4-6) universal screening and progress monitoring.

	Universal Screening							Progress Monitoring										
State-Supported Assessments	K	1	2	3	4	5	6	K	1	2	3	4	5	6	US time per student (min)	PM time per student (min)	Access to Student Data After Entry	Teacher Training Required (# of Days)
FAST Adaptive Reading (aReading) *	✓	✓	✓	✓	✓	✓									6 to 20		Instant	0.5
FAST Curriculum Based Measurement for Reading (CBM-R) *		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	3	1	Instant	0.5
FAST earlyReading First Grade Composite *		✓													3 to 5		Instant	0.5
FAST earlyReading Kindergarten Composite *	✓														5 to 7		Instant	0.5
FAST earlyReading Decodable Words *		✓							✓						1	1	Instant	0.5
FAST earlyReading Letter Naming *								✓								1 to 3	Instant	0.5
FAST earlyReading Letter Sound *	✓							✓							1 to 3	1	Instant	0.5
FAST earlyReading Nonsense Words *	✓	✓						✓	✓						1	1	Instant	0.5
FAST earlyReading Onset Sounds *	✓							✓							1 to 3	1 to 3	Instant	0.5
FAST earlyReading Sight Words 150 *		✓							✓						1 to 4	1 to 4	Instant	0.5
FAST earlyReading Word Blending *	✓	✓						✓	✓						1 to 3	1 to 3	Instant	0.5
FAST earlyReading Word Segmenting *	✓							✓	✓						1 to 3	1 to 3	Instant	0.5

There are additional FAST tests that did not meet the criteria as individual measures, but are used as a part of a composite that did meet the criteria. Those tests are not represented in this table.

Reading the table: The required grades (K-3) are indicated in light blue. Grades 4-6 are included in the displays to support planning at all grades. At a minimum, the assessment(s) used to meet the assessment requirements stated in Chapter 62 must meet the minimum standards for grades K-3 in Universal Screening and Progress Monitoring as indicated by the check mark. This can be accomplished with one test, or several in combination.

Other reviewed assessments

The table below contains all of the other reviewed assessments that met the minimum standards and follows the same formatting as above.

	Universal Screening							Progress Monitoring										
Other approved assessments	K	1	2	3	4	5	6	K	1	2	3	4	5	6	US time per student (min)	PM time per student (min)	Access to Student Data After Entry	Teacher Training Required (# of Days)
AIMSweb - Letter Sound Fluency **	✓							✓							1	1	Instant	0
AIMSweb - Letter Naming Fluency **	✓							✓							1	1	Instant	0
AIMSweb - Maze **				✓	✓	✓	✓				✓				3	3	Instant	0
AIMSweb - Reading CBM **		✓	✓	✓	✓	✓	✓		✓	✓	✓				3	1	Instant	0
easyCBM *				✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	45 to 50	45 to 50	Same Day	1
Edcheckup Maze Reading Passages **			✓		✓										3		Instant	0.5
Edcheckup Standard Reading Passages **			✓	✓	✓				✓						5	3	Instant	0.5
Gates MacGinite Reading Tests, 4th Edition *				✓	✓										75 to 100		Over 5 days	0
mCLASS : DIBELS Next *		✓	✓	✓	✓	✓		✓	✓	✓					3 to 6	1 to 5	Instant	1 to 2
mCLASS:Reading 3D *	✓	✓	✓	✓											5 to 8		Instant	1 to 2
Observation Survey of Early Literacy Achievement ***		✓													>15		Same Day	2
Phonological Awareness and Literacy Screening (PALS 1-3) *		✓		✓											25		Instant	1
Phonological Awareness and Literacy Screening (PALS-K) *	✓														30		Instant	0.5
STAR Early Literacy *	✓		✓	✓				✓	✓	✓	✓				12.5	12.5	Instant	0
STAR Reading *			✓	✓	✓	✓	✓				✓	✓	✓	✓	11	11	Instant	0
Texas Primary Reading Inventory (TPRI) **	✓	✓	✓	✓											1 to 5		Same Day	0.5

At this time, assessments not included in this list have not been approved by the Iowa Department of Education as meeting the minimum standards for use to comply with Chapter 62 requirements for universal screening and progress monitoring. There will be an opportunity to submit for review additional assessments or new evidence for existing assessments during the 2014-2015 school year. Information about this process will be posted this summer. The presence or absence of any assessment on this list does not imply any endorsement or criticism of the measure for other purposes, nor a specific endorsement for the purpose of universal screening and progress monitoring. There may be other reasonable and valid uses for these assessments. This list simply identifies the assessments that have been reviewed and determined to meet the minimum technical standards for use in K-3 literacy universal screening and progress monitoring.

Note: This approved list of assessments is shared provisionally, pending a second review of the Administrative Rules in Ch. 62 by the Administrative Rules Review Committee. Ch. 62 rules have received final approval from the State Board of Education.

Appendix

The following definitions are intended to provide a general explanation of the meaning and application for each of the required statistics and/or reporting elements for early literacy assessments.

Reliability

Reliability is a common measure of an important quality of an assessment: consistency. There are several ways to measure reliability, but mostly it comes down to measures of internal consistency (do the parts of the test work together to measure the same thing, or are there contradictions among the items?), consistency over time (can we trust that the test will measure consistently over time?), and consistency across testers (can the test be administered and scored to get consistent results?). Using a scale from 0.0 to 1.0, a reliability value of at least 0.70 is required, and a value above 0.80 is desired.

Validity

Validity statistics are used to help understand if the test results will allow users to make appropriate decisions. Many things can go into this understanding. For example, we ask how well the test results compare to another known measure of reading (this is called criterion validity). If a universal screening or progress monitoring test compares favorably with another measure of reading we can feel more comfortable that the results we get from the screening test are related to the student's reading ability and that our decisions about that student's skills are valid. Using a scale from 0.0 to 1.0, a validity coefficient of at least 0.3 was required, and at least .50 is desired.

Area Under The Curve (AUC)

Area under the curve (or AUC) is shorthand for area under the receiver operating characteristic curve, which is a statistical calculation that represents the relative value of a test for accurately classifying outcomes. The closer to 1.0 the AUC value, the better the test at predicting student success. A test with an AUC value of 0.5 predicts at the same rate as chance – in other words, the test is no better than flipping a coin. AUC values of .90 and better are considered excellent, .80 to .90 good, .70 to .80 fair, and .70 and lower are poor to worthless for predicting success.

Sensitivity/Specificity

Sensitivity and Specificity are statistics that represent the ability of the test to correctly identify students. Sensitivity represents the ability of the test to correctly identify the positive cases (students predicted on track for success). Specificity represents the ability of the test to correctly identify the negative cases (students predicted not on track for success). In the case of universal screening, the aim is for high sensitivity for a prediction of students on track to be successful readers. A test with a high value for sensitivity (approaching 1.0) will rarely miss identifying students who are on track to be successful readers. For these reviews, tests needed a minimum sensitivity value of 0.7, and a value of at least 0.8 was desired.

Number of forms of demonstrated equivalence

When using an assessment to monitor progress weekly it is important to make sure that there are enough forms to avoid a practice effect. It is also important to reduce any variation in test results over time caused by forms that are not of similar difficulty. At a minimum, at least ten forms are desirable,

along with some evidence from the test developer that they used a process to ensure that the forms are equivalent. At least 15-20 forms are preferred, as well as the use of more than one “industry standard” approach to determine the equivalence of forms.

Reliability of Slope

Reliability of slope is a statistic that represents the ability of the test to produce a consistent measure of student growth over time. A test with a less-reliable slope will do a poor job of accurately reflecting student improvement. A test with a very reliable slope will show results that best represent the student’s improvement over time. Using a scale from 0.0 to 1.0, a reliability of slope value of at least 0.60 is required, and a value above 0.70 is desired.

Administration time

It is important to find tests that are efficient. Since testing takes away from instructional time, it is a good idea to minimize the amount of time spent testing. If two tests are otherwise similar (reliable, valid, etc.), the test that takes less time may be preferred.

Accessibility of student data

For the purposes of universal screening and monitoring progress it is important for teachers to gain access to results quickly in order to begin using the data. A lag between testing and availability of data will cause the system to be less responsive to student needs. It is preferred to be able to receive and use results immediately after testing.

Teacher training

The amount of training needed to reliably administer the tests and use the results is important for planning and resource allocation.