1.0 Validation Data and Reports

1.1 Technical Adequacy of the Assessment

mCLASS:Reading 3D is a synthesis of DIBELS and Text Reading and Comprehension (TRC) assessments. This assessment program is an innovative approach to K-5 literacy that balances the assessment of Foundational Skills with Text, Reading and Comprehension (TRC) diagnostics, giving a complete picture of a student's reading development.

The mCLASS:Reading 3D solution is the only validated, research-based assessment that combines quick indications of early skill development with deep observations of students' interactions with authentic texts. This solution integrates the predictive power of the DIBELS assessment and the strength of TRC.

DIBELS – Literacy screening and progress monitoring

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) are a well-known set of procedures and measures for assessing the acquisition of early literacy skills from kindergarten through Grade 6. They are short (1 minute) fluency measures used to regularly monitor the development of early literacy and early reading skills. Please refer to https://dibels.org/pubs.html for more information about the research establishing the reliability and validity of the DIBELS Next measures.

Reliability

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) are a well-known set of procedures and measures for assessing the acquisition of early literacy skills from kindergarten through sixth grade. They are designed to be short (one minute) fluency measures used to regularly monitor the development of early literacy and early reading skills.

The DIBELS Next assessment has been researched extensively to ensure that it meets strict criteria for reliability and validity. Reading is a complex process that requires the integration of many skills. DIBELS Next results provide teachers a perspective on overall reading performance as well as performance on specific reading skills. Field tests indicate that the DIBELS Next Composite Score predicts future reading proficiency better than similar analysis performed with just information about a student's ability on an individual reading sub-skill. This enables teachers to accurate identify students – as early as kindergarten – who are at risk for reading difficulty. The DIBELS Next benchmark goals and cut scores are empirically derived, based on research that examined the longitudinal predictive validity of a score on a measure at a particular point in time. These goals indicate the probability of achieving the next benchmark goal.

Reliability information for DIBELS Next includes technical data and analyses from five research studies:

- Study A evaluated the reliability and validity of a new kindergarten measure, First Sound Fluency.

- Study B was designed to evaluate the new directions and materials that would become part of DIBELS Next, and to examine the reliability of two new DIBELS Next measures, First Sound Fluency and Daze.
- Study C was designed to obtain the necessary information to set benchmark goals for DIBELS Next, in addition to obtaining data on the reliability and validity of all DIBELS Next measures.
- The goal of Study D was to evaluate DIBELS Next Oral Reading Fluency (DORF) passages for reliability, validity, and passage difficulty. Study E was designed to obtain alternate-form reliability information on Phoneme Segmentation Fluency (PSF) in first grade and all sixth grade measures, and additionally, to evaluate the alternate-form reliability of Retell in second grade.

For more information about the studies, such as demographic information and recruitment procedures, please see Chapter 3 (p. 34) of the DIBELS Next *Technical Manual (*https://dibels.org/pubs.html*)*.

The overall reliability of DIBELS Next is summarized in Table R-1. Reliability coefficients are consistently high across all three forms of reliability. The magnitude of the coefficients suggests that DIBELS Next possesses little test error and that users can have confidence in test results. With repeated assessment across multiple forms, reliability increases substantially, as noted where the estimated three-form reliability is reported.

For more information about the reliability of DIBELS Next, please see the Chapter 5, page 81, of the DIBELS Next *Technical Manual*.

Summary Table of Reliability Estimates for DIBELS Measures

	Type of Reliability			
	Alternate-Form		**Inter-Rater**	
***DIBELS* Measure**	Single-Form	Three-Form	Single-Form	Three-Form
	Kindergarten			
First Sound Fluency	.82	.93	.94	.98
NWF Correct Letter Sounds	.71	.88	.99	1.00
DIBELS Composite Score	.66	--	.97	--
	First Grade			
Phoneme Segmentation Fluency	.54	.78	.95	.98
NWF Correct Letter Sounds	.85	.94	.99	1.00
DORF Words Correct per Minute	.95	.98	--	--
DIBELS Composite Score	.95	--	.99	--
	Second Grade			
DORF Words Correct per Minute	.89	.96	--	.99
DIBELS Composite Score	.92	--	.98	--
	Third Grade			

DORF Words Correct per Minute	.89	.97	--	.99
Daze	.81	.93	.99	1.00
DIBELS Composite Score	.97	--	--	--
Fourth Grade				
DORF Words Correct per Minute	.88	.95	--	.99
Daze	.74	.90	.98	.99
DIBELS Composite Score	.95	--	--	--
Fifth Grade				
DORF Words Correct per Minute	.92	.96	--	.99
Daze	.66	.85	.99	1.00
DIBELS Composite Score	.91	--	--	--
Sixth Grade				
DORF Words Correct per Minute	.83	.94	--	.99
Daze	.79	.89	.99	1.00
DIBELS Composite Score	.91	--	--	--

Note. Inter-rater reliability calculated from Study C. Alternate-form reliability reported is the median reliability from studies A, B, C, D, and E, for measures where multiple reliability coefficients were available. DORF single-form reliability based on Study D, DORF three-form reliability for first through fifth grades based on Study C, and DORF three-form reliability for sixth grade is based on Study E.
All correlations significant at the *α* < .001 level.

The following table outlines alternate form reliability, and criterion-related validity (concurrent and predictive) for each IDEL measure.

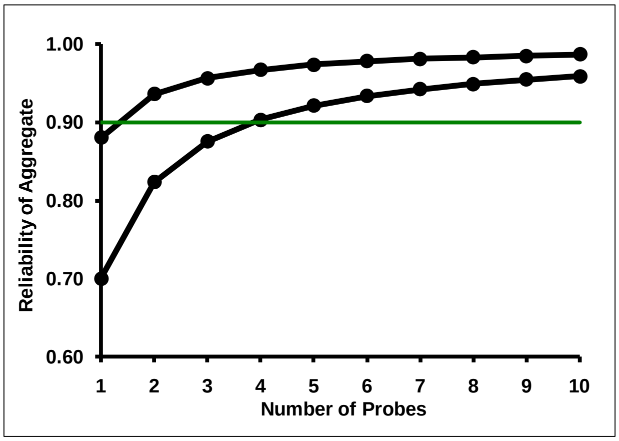
Measure	Criteria	Score	Comparison
IDEL Fluidez en Nombrar Letras (FNL)	Alternate form reliability (3 week)*	.91	N/A
	Criterion validity	.58	Woodcock-Muñoz, Batería-R APR Destrezas Básicas
Fluidez en la Segmentación de Fonemas (FSF)	Alternate form reliability (3 week)*	.87	N/A
Fluidez en las Palabras sin Sentido (FPS)	Alternate form reliability (3 week)*	.76	N/A
	Criterion validity	.72	Woodcock-Muñoz Pruebas de Aprovechamiento subtest of Análisis de Palabras
Fluidez en la Lectura Oral (FLO)	Alternate form reliability (3 week)*	.87-.94	

	Criterion validity	.79	Woodcock-Muñoz Bateria-R Combined Scores of Amplia Lectura

Details on research conducted to establish these data is available at https://dibels.org/pubs_idel.html.

Reliability of multi-probe aggregates

Because each DIBELS probe is essentially an item and there are multiple, brief, repeatable probes, repeated assessments can be used to examine a student's skills over time and educational decisions can be made on the basis of repeated alternate forms. The Spearman-Brown Prophesy Formula (Nunnally, 1978) provides an estimate of the reliability of measurement of student skills based on multiple probes. The following figure illustrates the reliability of a total when combining probes with reliability of .70 and .88. When 4 repeated assessments are aggregated together, the result has a reliability above .90, even when the individual probe (item) has a relatively low reliability of .70. Repeated any of the DIBELS assessments 4 times is still an extremely brief and efficient assessment, taking less than 5 minutes to administer.



From Nunnally, J. C. (1978). Psychometric theory. (2nd ed.). New York: McGraw-Hill.

The Outcomes-Driven Model for educational decisions on which the DIBELS are based capitalizes on the increased reliability of multi-probe aggregates. The first step of the model is to *identify need for support*. The first step involves universal screening with a single probe to identify students who may be in need of additional support. The second step of the model is to *validate need for support*. In the second step of the model students are retested with alternate forms of the DIBELS on different days and under different conditions and the pattern of scores is examined. Intervention is indicated when the educator is reasonably confident the student needs additional support to meet important benchmark literacy goals.

Validity

DIBELS Next measures are designed to be general outcome measures. As such, DIBELS Next measures are *indicators* of overall performance in a particular skill domain, meaning, "they measure key skills that are representative of and related to important global outcomes, such as reading competence" (Kaminski & Cummings, 2007, p. 27).

Evidence for the validity of DIBELS Next is summarized in Tables V-1 through V-5. Evidence of content, criterion-related, discriminant validity, and classification consistency are reported.

Overall, the validity of DIBELS Next is well supported by several external criterion measures. For more information about the validity of DIBELS Next, please see the Chapter 6, page 94, of the DIBELS Next *Technical Manual*.

Content Validity

The DIBELS Next measures were designed specifically to be linked to foundational early literacy skills and sensitive to growth and change in response to instruction or intervention in those areas. *DIBELS* measures serve as key indicators of foundational skills in beginning reading (National Reading Panel, 2000; National Research Council, 1998).

Table V-1 illustrates the linkage between the foundational early literacy skills and each DIBELS Next measure.

Table V-1. *Alignment of DIBELS Next Measures with Basic Early Literature Skills*

Basic Early Literacy Skills	***DIBELS Indicators***
Phonemic Awareness	First Sound Fluency Phoneme Segmentation Fluency
Alphabetic Principle & Basic Phonics	Nonsense Word Fluency[1] --Correct Letter Sounds
Advanced Phonics & Word Attack Skills	*DIBELS* Oral Reading Fluency[2] --Words Correct Per Minute
Accuracy & Fluency with Connected Text	*DIBELS* Oral Reading Fluency[2] --Words Correct Per Minute
Comprehension	Daze *DIBELS* Oral Reading Fluency[2] --Words Correct Per Minute

[1] Nonsense Word Fluency is an indicator of early phonics skills or the alphabetic principle, specifically, does the student know the most common sounds for each letter and can he/she correctly blend the sound with the sounds before and after to read an unknown word?

[2] Oral Reading Fluency is a more advanced indicator of decoding skills and the student's application of those skills to reading connected text.

For additional information on the foundation for the DIBELS Next measures, please see Chapter 1 of the *DIBELS Next Technical Manual* as well as Good, Simmons, & Smith (1998); Kaminski (1992; page 23 to 32); Kaminski, Cummings, Powell-Smith & Good (2008); and Kaminski & Good (1996).

Content Validity for Individual Measures

The design specifications for DIBELS Next measures relate directly to their content validity. Each measure was designed according to specific criteria to maximize their utility and sensitivity. For information on design specifications for DIBELS Next measures, see Chapter 2 of the DIBELS Next *Technical Manual*, p 18.

Table V-2. *Content Validity for Individual DIBELS Measures*

***DIBELS* Measure**	**Content Validity**
FSF	First Sound Fluency (FSF) is a brief, direct measure of a student's fluency in identifying the initial sounds in words. The ability to isolate the first sound in a word is an important phonemic awareness skill that is highly related to reading acquisition and reading achievement (Yopp, 1988). The ability to isolate and identify the first phoneme in a word is an easier skill than segmenting all the sounds in words or manipulating phonemes in words, thus FSF is used as a measure of developing phonemic awareness at the beginning and middle of kindergarten.
PSF	Phoneme Segmentation Fluency (PSF) is a brief, direct measure of phonemic awareness. PSF assesses the student's fluency in segmenting a spoken word into its component parts or sound segments.
NWF Correct Letter Sounds	Nonsense Word Fluency (NWF) is a brief, direct measure of the alphabetic principle and basic phonics. It assesses knowledge of basic letter-sound correspondences and the ability to blend letter sounds into consonant-vowel-consonant (CVC) and vowel-consonant (VC) words. The test items used for NWF are phonetically regular make-believe (nonsense or pseudo) words. One reason that nonsense word measures are considered to be a good indicator of the alphabetic principle is that "pseudowords have no lexical entry, [and thus] pseudo-word reading provides a relatively pure assessment of students' ability to apply grapheme-phoneme knowledge in decoding" (Rathvon, 2004, p. 138).
DORF Words Correct / Min	*DIBELS* Oral Reading Fluency (DORF) is a measure of advanced phonics and word attack skills, accurate and fluent reading of connected text, and reading comprehension. Oral reading fluency has been shown to strongly correlate to measures of reading comprehension and overall reading skill (Fuchs, Fuchs, Hops, and Jenkins (2001)). The DORF passages and procedures are based on the program of research and development of Curriculum-Based Measurement of reading by Stan Deno and colleagues at the University of Minnesota (Deno, 1989).
Daze	Daze is the standardized, *DIBELS* version of a maze testing procedure for measuring reading comprehension. The purpose of a maze assessment is to measure the reasoning processes that constitute reading comprehension. Specifically, Daze assesses the student's ability to construct meaning from text using comprehension strategies, word recognition skills, background information and prior knowledge, familiarity with linguistic properties such as syntax and morphology, and reasoning skills.

Criterion-Related Validity

Validity of the DIBELS Next measures was examined using a variety of criterion measures including the Group Reading Assessment and Diagnostic Evaluation (GRADE), the Standard 4th Grade Reading Passage used in the National Assessment of Education Progress (NAEP) 2002 Special Study of Oral Reading (Daane, et al., 2005), and the *DIBELS* Composite Score.

Evidence of validity is presented as a correlation between the assessment and the criterion in Tables V-3 and V-4. Table V-3 presents correlations between DIBELS measures and the GRADE Total Test based on beginning- to end-of-year data, except for kindergarten NWF Correct Letter Sounds, which is based on middle to end-of-year data. Correlations between DIBELS measures and the DIBELS Composite Score are based on beginning- to end-of-year, except for first grade DORF, which is based on middle- to end-of-year. Table V-4 presents correlations between student performance on DIBELS Oral Reading Fluency passages and performance on the National Assessment of Educational Progress Oral Reading Study, 4th grade passage.

Table V-3. Summary Table of Criterion-Related Validity for DIBELS Measures

	General Outcome Measure	
***DIBELS* Measure**	**GRADE Total Test**	**DIBELS Composite Score**
Kindergarten		
First Sound Fluency	.52	.57
NWF Correct Letter Sounds	.47	.65
DIBELS Composite Score	.50	.67
First Grade		
Phoneme Segmentation Fluency	.33	.25
NWF Correct Letter Sounds	.56[c]	.82
DORF Words Correct per Minute	.75[c]	.83
DIBELS Composite Score	.77[c]	.79
Second Grade		
DORF Words Correct per Minute	.73	.85
DIBELS Composite Score	.75[c]	.79
Third Grade		
DORF Words Correct per Minute	.66[c]	.88
Daze	.67[c]	.79
DIBELS Composite Score	.75[c]	.86
Fourth Grade		
DORF Words Correct per Minute	.74[c]	.90
Daze	.68[c]	.76
DIBELS Composite Score	.80[c]	.92

	Fifth Grade	
DORF Words Correct per Minute	.65[c]	.89
Daze	.66[c]	.74
DIBELS Composite Score	.77[c]	.89
	Sixth Grade	
DORF Words Correct per Minute	.61[c]	.87
Daze	.64[c]	.78
DIBELS Composite Score	.73[c]	.91

Note. **Based on Study C data. Pair-wise sample sizes range ≈ 440-570. GRADE total test given at end of year benchmark assessment.**

[c] Correlations represent concurrent validity and are based on end of year *DIBELS* scores correlated with end-of-year GRADE Total Test. All correlations not marked represent predictive validity.

All correlations significant at the *α* = .001 level.

Table V-4. *Criterion-Related Validity for DIBELS Oral Reading Fluency Single Passage with NAEP*

DORF Single Passage Words Correct per Minute by Grade	NAEP 4th Grade Passage
First	.97
Second	.91
Third	.96
Fourth	.89
Fifth	.96
Sixth	.83

Note. **Based on end-of-year data from Study D. Sample sizes ≈ 23 for all grades.**

All correlations are significant at the *α* = .001 level.

Discriminant Validity

Discriminant validity was examined for the *DIBELS* Composite Score relative to two levels of performance on the GRADE Total Test, below the 40th percentile on the GRADE's national norms and at or above the 40th percentile. *DIBELS* Composite Score descriptive statistics were calculated for each group and compared. Differences in means were examined using a between-groups *t*-test for each grade; the *t*-statistics are reported to illustrate the magnitude of the differences in means. Cohen's *d* is also reported to emphasize the magnitude of the effect size. Results are reported in Table V-5, referenced from the *DIBELS Next Technical Manual*, Chapter 6, page 107.

Table V-5. *Discriminant Validity for the DIBELS Composite Score Based on Percentile Rank of GRADE Total Test Raw Score*

	DCS Descriptive Statistics by GRADE Total Test %-Rank		
Grade by Time	Below 40th %-ile	Above 40th %-ile	Difference Statistics

of Year	N	*Mean*	*SD*	N	*Mean*	*SD*	*t*-stat	Cohen's *d*
Kindergarten								
Beginning	54	22.31	19.65	112	45.42	23.76	6.41	1.03
Middle	55	111.90	54.96	113	156.10	43.16	5.45	0.94
End	53	132.10	40.78	113	156.50	39.09	3.67	0.62
First Grade								
Beginning	54	105.00	29.68	139	145.90	39.54	7.33	1.11
Middle	55	96.51	48.69	140	220.50	88.12	11.17	1.58
End	54	115.10	65.34	139	228.00	59.81	11.26	1.85
Second Grade								
Beginning	61	111.20	61.32	153	219.80	60.88	11.74	1.79
Middle	61	136.70	83.90	158	282.10	60.87	13.26	2.15
End	60	194.00	82.49	157	309.90	67.27	10.19	1.62
Third Grade								
Beginning	49	168.80	96.65	135	327.60	85.88	10.43	1.80
Middle	51	221.50	94.03	136	390.30	83.82	11.56	1.96
End	51	279.80	99.64	136	442.00	79.58	11.00	1.91
Fourth Grade								
Beginning	64	200.00	110.10	119	360.60	82.62	10.68	1.73
Middle	65	250.20	102.00	120	400.90	73.24	11.06	1.79
End	66	316.30	106.30	120	467.60	76.42	10.70	1.73
Fifth Grade								
Beginning	93	311.70	95.38	101	454.30	77.61	11.41	1.66

Middle	92	346.70	82.12	102	477.50	73.50	11.67	1.69
End	92	377.80	92.00	101	527.80	80.28	12.06	1.75
Sixth Grade								
Beginning	19	292.30	98.61	84	442.30	77.95	6.70	1.85
Middle	17	330.90	112.70	85	483.80	93.18	5.61	1.60
End	19	334.90	101.40	86	502.40	84.85	7.11	1.92

Note. Based on Study C data. *T*-tests were performed for difference in means, and Folded *F*-tests were performed for difference in variance; all tests yielded highly significant results. All *t*-tests were performed under both equal and unequal variance assumptions; the reported *t*-statistic is the average between the two tests under different assumptions.

Classification Consistency

Using Receiver Operating Characteristic (ROC) curve analysis at each grade and time of year, sensitivity, specificity, and accuracy of the DIBELS Next Composite Score was calculated. As an indicator of benchmark performance, these results indicate the degree to which the DIBELS Composite Score accurately predicts subsequent performance at or above the benchmark goal for the next time of year. Additionally, end-of-year results were used to predict student performance on the GRADE, providing validity evidence on an external measure. Across all times of year, the DIBELS Composite Score is a highly sensitive and highly specific metric of reading skill, and provides accurate classification resulting from criterion-referenced outcome measures.

Table V-6 summarizes the results of the ROC analysis.

Table V-6. *Sensitivity, Specificity, and Accuracy for the DIBELS Next Composite Score*

Prediction Time of Year	Outcome Time of Year (or Measure)	Sensitivity	Specificity	Accuracy
KBOY	KMOY	0.65	0.83	0.78
KMOY	KEOY	0.63	0.83	0.77
KEOY	GRADE	0.38	0.85	0.70
1BOY	1MOY	0.68	0.86	0.80
1MOY	1EOY	0.79	0.89	0.86
1EOY	GRADE	0.74	0.88	0.84
2BOY	2MOY	0.79	0.91	0.88
2MOY	2EOY	0.81	0.92	0.89
2EOY	GRADE	0.71	0.89	0.84
3BOY	3MOY	0.82	0.93	0.90

3MOY	3EOY	0.78	0.91	0.87
3EOY	GRADE	0.73	0.90	0.86
4BOY	4MOY	0.84	0.91	0.89
4MOY	4EOY	0.80	0.90	0.87
4EOY	GRADE	0.71	0.84	0.80
5BOY	5MOY	0.83	0.90	0.88
5MOY	5EOY	0.79	0.89	0.85
5EOY	GRADE	0.73	0.88	0.83
6BOY	6MOY	0.77	0.93	0.90
6MOY	6EOY	0.76	0.93	0.90
6EOY	GRADE	0.68	0.93	0.89

Note: Based on Study C data. N = 3,816. BOY = Beginning of year; MOY = middle of year; EOY = end of year. GRADE = Group Reading Assessment and Diagnostic Evaluation.

Technical Adequacy of TRC

We have worked to ensure that our TRC measure provide reliable, valid assessment data for educators.

Reliability

In a field study in a Midwestern state, student data was collected from the 2009-2010 school year from a midwestern state. Students in this sample were identified as 49% white, 12% African American, and 8% Hispanic-Latino. Using this data **Internal consistency** is calculated as marginal reliability (Sireci, Thissen, & Wainer, 1991), an Item Response Theory indicator of overall precision; marginal reliability estimates for each of Kindergarten through Grade 3 are presented in the table below as well as according to the largest reported ethnicity groups. Marginal reliability is shown to be high across grades (median = 0.86) and comparable across student sub-groups, suggesting that TRC is precise when applied to various student populations.

Ethnicity	Grade	N	Marginal Reliability
Overall	K	50,822	0.79
	1st	60,134	0.86
	2nd	56,301	0.88
	3rd	1774	0.86
White	K-3	1076 - 30,057	0.93
Black or African-American (Non-Hispanic)	K-3	269 - 7945	0.92
Hispanic or Latino	K-3	63 - 5157	0.92

Data collected from twenty seven (27) students in a South Eastern state during the 2007-2008 school year was used to evaluate **inter-rater reliability** of TRC. The students in this study attended public elementary schools in a suburban Mid-Atlantic school district, were in kindergarten through Grade 3, and represented a range of reading abilities with respect to their grade levels (text levels PC/RB through

P). Audio recordings were made while students completed the TRC assessment face-to-face with a local test administrator. Twenty-one raters (21) from two East Coast cities with an average of 5 years' experience administering Reading Records were trained to administer and score TRC. These raters followed along with each student recording and scored the Reading Record component of TRC. The average Spearman rank-order correlation between scores recorded by the original local test administrator and each of the trained raters is high ($\rho = 0.73$), indicating a high degree of consistency.

We are currently conducting field testing for performance specific to achievement as measured by the CCSS. Please note that research regarding cut points in Grade K-5 will be available in April but data for Grade 6 will be collected during the first year of implementation. We will be happy to share results once they are available.

Validity

Beginning in 2004, we collaborated with Montgomery County Public Schools(MCPS) as part of the Assessment Program in Primary Reading (AP-PR) to develop the TRC measure. The goal was to develop an assessment instrument that was pedagogically balanced (addressing phonics and comprehension) and vertically integrated – providing information about all children across a reading spectrum, whether they were barely sounding out letters in second grade, or reading third-grade books as a Kindergartner. Years of research established the initial proficiency level cut points by correlating performance levels to performance on external measures of reading performance, such as the Comprehensive Test of Basic Skills (CTBS), TerraNova 2nd Edition, and the Grade 3 Maryland State Assessment.

From 2006-2008, Montgomery County Public Schools conducted a longitudinal study looking at the predictive power of the TRC subtest. The sample consisted of continuously enrolled Kindergarten, first grade, and second grade students (Grade K: 7,691; Grade 1: 8,426; Grade 2: 8,392). Test results were analyzed to determine the extent to which: Kindergarten benchmark status was predictive of first grade benchmark status; first grade benchmark status was predictive of second grade benchmark status; and second grade benchmark status was predictive of third grade Maryland State Assessment (MSA) reading proficiency. Findings were as follows:

Benchmark Status Comparison	Predictive Accuracy (%)
Grade K (2007) TRC to Grade 1 (2008) TRC	86.5
Grade 1 (2007) TRC to Grade 2 (2008) TRC	75.7
Grade 2 (2007) TRC to Grade 3 (2008) MSA	67.8*

**The primary reason for this low prediction accuracy was that 29.7% of students scored proficient or higher on the2008 MSA Reading even though they were reading below text level M (27) at the end of Grade 2(performed better than expected). Only 2.4% of students performed worse than expected. This finding provides further evidence that Grade 2 students who meet the MCPS end-of-year reading benchmark are likely to attain a score of proficient or higher on the MSA Reading after they receive Grade 3 reading instruction.*

Evidence for the **content validity** of TRC is derived from an understanding of the test development process. TRC is based on an assessment approach developed by Marie Clay, author of *An Observation Survey of Early Literacy Achievement*, and built around the idea of leveled texts. Organized similar to an informal reading inventory combined with an explicit focus on comprehension, TRC is a reading record assessment (alternately known as a running record) for analysis of student reading performance which allows teachers to evaluate students' performance on the foundational skills necessary to become fluent readers and their ability to correctly apply those skills to increasingly complex text.

By administering a series of benchmark books, the teacher can efficiently determine each student's instructional reading level. The instructional reading level is further validated through a set of oral comprehension questions and, starting at text levels typically appropriate for students in grades 2 or 3, written comprehension questions. The instructional reading level determined within TRC helps educators select texts for students to read that are neither frustratingly difficult nor ineffectively easy with regards to promoting reading development (Clay, 2002, 2005).

Grade 2 TRC data from the previously described Great Lakes state was matched to subsequent Grade 3 student performance data on the statewide, high-stakes assessment of English Language Arts in 2009-2010. The majority of students in this matched data set identified as White, Non-Hispanic (57.0%), Black/African American (15.6%) and Hispanic (9.5%).

Using this matched data set, evidence for the **predictive validity** of TRC is calculated as the disattenuated correlation between students' instructional reading levels and subsequent scores on the high-stakes assessment, accounting for the reported internal consistency of the high-stakes instrument. The results suggest a moderate-to-strong relationship between student performance on TRC and the high-stakes assessment. Further, the correlations are seen to be similar across student sub-groups.

Ethnicity	N	Disattenuated Correlation
Overall	21,066	0.70
White	12,010	0.68
Black or African-American (Non-Hispanic)	3280	0.65
Hispanic or Latino	2007	0.69

Classification Consistency

Receiver Operating Characteristic (ROC) analysis was conducted on the matched dataset described previously to determine **Sensitivity** (the ratio of students correctly predicted by TRC as Not Proficient on the statewide assessment to all students identified as Not Proficient), **Specificity** (the ratio of students correctly predicted by TRC as proficient on the statewide assessment to all proficient students), and **Accuracy** (the overall proportion of correct predictions by TRC). The results show that TRC demonstrates high Accuracy and Specificity and moderate Sensitivity. Performance for the three student sub-groups generally approximates the overall rates suggesting that the assessment does not function differently across student sub- groups.

Ethnicity	N	Sensitivity	Specificity	Accuracy
Overall	21,066	0.55	0.89	0.81
White	12,010	0.49	0.92	0.84
Black or African-American (Non-Hispanic)	3280	0.62	0.80	0.73
Hispanic or Latino	2007	0.55	0.86	0.74

DIBELS and TRC Combined

In two studies conducted internally, the combination of DIBELS and TRC has been found to be predictive of high-stakes assessment performance on two state summative reading comprehension tests.

Our results provide compelling evidence that both DIBELS and TRC scores in second grade help educators gain insight into how their students will perform on high stakes tests in grade three. If problems are detected, instruction can be adjusted to prevent high stakes failure before it is too late. Additionally, the analyses reported here suggest that while ORF and TRC are effective predictors on their own, the predictive utility of the mCLASS:3D assessment is maximized when both measures are considered.

In the first analysis, robust correlations were found between ORF, TRC, and state summative reading comprehension tests, ranging from 0.69 to 0.72 for the EOG summative test and from 0.66 to 0.67 for the ISTEP+ summative test. These correlations echo those found in other studies of the validity of ORF. Roehrig et al (2008), for example, found that DIBELS ORF given in third grade has linear correlations in the range (r = 0.67-0.70) with the Florida Comprehensive Assessment Test (FCAT-SSS), a third grade reading comprehension assessment used in Florida for purposes of accountability. OLS regression of EOG score onto ORF score and TRC level showed that the mCLASS:3D measures each make independent and significant contributions to the prediction of the EOG score outcome, yielding a high linear correlation of r = 0.77. Likewise, OLS regression of ISTEP+ score onto ORF score and TRC level yielded a similarly high linear correlation of r = 0.73.

Our second analysis investigated the degree to which second grade ORF and TRC correctly predict whether or not students reach reading proficiency at the end of third grade on the EOG four-point achievement scale and the ISAT+ three-point achievement scale. Overall, mCLASS:3D correctly classified 84.81% of the students taking EOG and 84.22% of the students taking ISTEP+, a strong result that again compares well with those reported by Roehrig et al. (overall correct classification in the range 78% to 86%). Interestingly, the sensitivity (0.91 for EOG; 0.97 for ISTEP+) and specificity scores (0.70 for EOG; 0.46 for ISTEP+) found in analysis 2 parallel those found by Roehrig et al for ORF but contrast somewhat with the scores reported by Goffreda and DiPerna (2009), for whom, at the optimum cut-off, the sensitivity and specificity of ORF was 0.88. The difference is perhaps explained by differences in the outcome measures and grade level of the students, as Goffreda and DiPerna investigated the predictive utility of ORF in first grade against the TerraNova assessment, administered in second grade. While the FCAT, EOG, and ISTEP+ tests are focused on reading comprehension, the TerraNova also includes items designed to assess earlier reading skills.

It is important to recognize that the significant overlap in predictive validity between ORF and TRC does not entail that the two measures are broadly equivalent. This is because the classroom utility of the two measures is quite different.

ORF is intended as a screening measure of reading fluency and accuracy, which it accomplishes by calculating an oral reading fluency score based on the number of words a student reads correctly in one minute. TRC is also intended to measure reading accuracy, based on the percentage of words in a text that student reads correctly. (Optionally, TRC can be timed to assess reading fluency as well.)