3pSC17

Acoustical Society of America

San Diego, California. 4 December 2019

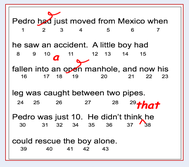
Accurate reading rate: Validations of machine scoring.

Masanori Suzuki, Jared Bernstein, Jian Cheng, & Tomoyo Okuda – ***Analytic Measures Inc.***

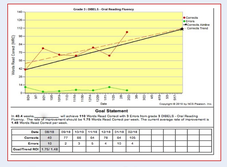
Analytic Measures Inc.

Accurate oral reading rate is a common measure of early reading proficiency (grades K-4). A concurrent validation study compared a fully automated measure of accurate reading rate (words read correctly per minute) to DIBELS, a well-established, human-scored assessment. We report data from 174 students at a parochial school in Delaware and 130 students at a public school in Texas. Each student took three forms of an automatically scored test (yielding 898 test forms automatically scored) and also took the grade-appropriate form of the standard DIBELS test, which was scored by a trained person. Test administration order was counterbalanced. Correlations were computed for Machine-Machine and Human-Machine score pairs, for each grade separately and for the total dataset. When all grades are combined, Pearson r = 0.87, and correlations for single grades are all above 0.85, thus at or above the reliability ceiling of the DIBELS criterion score. Also, the automated tests have higher test-retest reliability (r=0.91) than reported in an independent DIBELS study (Goffreda and DiPerna, 2010). [This research was supported by IES, U.S. Department of Education contract ED-IES-17-C-0030 to Analytic Measures, Inc.]

Oral Reading Fluency (ORF) is a commonly used reading measure that benchmarks and tracks progress in early readers. For most early readers, a simple count of *words read correctly per time* is a strong predictor of reading comprehension. Oral reading performance is traditionally hand scored and tracked against a reading-score growth goal.



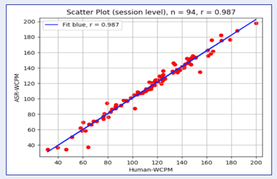
Hand-Scoring



Progress Chart

Automatic ORF Scoring can be accurate and very efficient. Students self-administer a test in 12 minutes with scores returned immediately.

2017 data [1] from 94 students shows that automatic scoring can be as accurate as scoring by a team of skilled human judges who are scoring the same audio files. The students also preferred the self-administered test to a human examiner.



Accurate Reading Rate (WCPM)

Average Human *vs.* **Moby.Read**

r(Moby, Human)=0.987

Human Inter-rater reliability = 0.992

Concurrent Validation was needed. Earlier studies [2,3] have shown that machine scoring matches human scoring of the same performances and that self-administration of an oral reading fluency test is well-accepted by K-5 students.

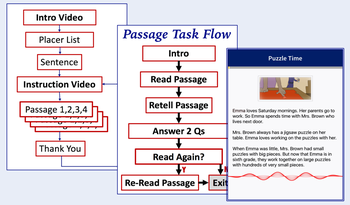
Still unanswered is: ***Are machine scores valid*?**

1. ***Do machine scores match other, more traditional ORF tests?***
2. ***Are scores consistent (reliable) across forms when students are re-tested?***

We compared AMI's ***Moby.Read*** [4], a fully automated ORF assessment, to ***DIBELS*** [5,6], a human-administered, human-scored ORF test.

Instruments

Moby.Read is a self-administered, fully automated ORF assessment for grades K-5 that scores Expression, Accuracy, Level, and Words Correct per Minute (WCPM). It runs in iOS or in Chrome.



Moby.Read presents four grade-appropriate passages and scores the last three readings.

DIBELS (Dynamic Indicators of Basic Early Literacy Skills) is a well-established assessment of oral reading fluency. DIBELS is administered one-to-one to a student by a trained adult. It is also hand-scored by a teacher or another trained adult.

WCPM is the principal score returned by DIBELS and the only score construct truly shared by the two tests, so validation analysis is limited to *Accurate Rate* scores, which are reported in *Words Correct per Minute* (WCPM).

Methods The validation study compared Moby.Read scores as a measure of oral reading fluency with DIBELS scores.

Participants: A total of 304 students in Grades 1-5 were recruited for the study at two locations: 174 students at one parochial school in Delaware and 130 students at a public school in Texas. Student demographics are given in Table 1.

Table 1.	Delaware		Texas	
Demographics	Count	Percent	Count	Percent
European-American	127	73.0%	89	68.5%
African-American	25	14.4%	20	15.4%
Multiple races	8	4.6%	14	10.8%
Asian	7	4.0%	2	1.5%
Indigenous NAm.	1	0.6%	4	3.1%
Pacific Islander	0	0.0%	1	0.8%
No information	6	3.4%	0	0.0%
TOTAL	174	100.0%	130	100.0%

Independent of race-ethnicity, 62.3% of the Texas students were identified as Hispanic/Latino. The female:male ratio was 58:42.

Materials: Both tests score 3 leveled text passages.

Table 3. Text Material Test Form	Passages Presented	Passages Scored		Passage Length in Words
		Narrative	Informational	
Moby.Read (Grades 1-5)	4	2	1	40-150
DIBELS (Grades 1-3)	3	2	1	220-340
DIBELS (Grades 4-5)	3	1	2	

Procedure: Each participant took three grade-level forms of Moby.Read (Fall, Winter, and Spring) and the grade-level Fall form of DIBELS. Moby.Read forms were administered in a small group setting (6-12 students at a time) and all students were fitted with a headset with a noise-canceling microphone. DIBELS was administered in a one-to-one setting with a trained administrator. The order of the test administrations was counterbalanced.

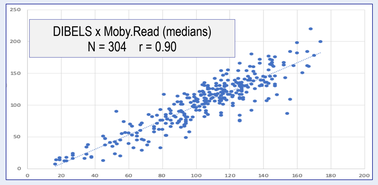
Scored Performance: DIBELS scores the first 60 seconds of the reading after the administrator says "begin", and scores by counting words read minus a count of word reading errors. Moby.Read locates the text span (from 1st to last word read correctly in order) in the text and in the audio file, then counts words read correctly in that time span.

$$Accurate\ Rate\ (WCPM) = \frac{Number\ of\ Words\ Correct\ in\ Text\ Span}{Span\ Duration} * 60$$

This difference in performance scoring logic might make a difference in some cases, but overall, it did not.

Results Both Moby.Read and DIBELS report WCPM scores (Accurate Rate) as the median of the WCPM scores from the three passage readings.

Pearson correlations were computed for score pairs, within each grade separately and also for the total dataset. Almost all of the study's 304 students completed three forms of Moby.Read. As a group the 304 students produced 898 Moby.Read (WCPM) scores. Among 304 students, one student completed only one Moby.Read form and 11 students had only two scores. All the remaining students yielded three Moby.Read scores.



The correlations between the WCPM scores from Moby.Read and DIBELS were high across all grades, ranging from 0.85 to 0.91. When data from all grades are combined, the overall correlation is r = 0.87.

Table 2. Score match.	Moby.Read	DIBELS	Correlation
Grade	Scores (N)	Scores (N)	r (Moby.Read,DIBELS)
1	167	58	0.85
2	209	70	0.89
3	156	52	0.88
4	192	65	0.88
5	174	59	0.91
All Scores	898	304	**0.87**

One independent study [7] of DIBELS reports a test-retest reliability of 0.82 and an inter-rater reliability of 0.85. The reliability of an instrument limits the strength of the correlation between that instrument and others measuring the same construct.

Machine scoring of self-administered Moby.Read tests is relatively consistent across forms.

Table 4. Moby.Read Test-Retest Reliability				
Grade	N	Fall-Winter	Winter-Spring	Fall-Spring
1	60	0.90	0.93	0.93
2	71	0.91	0.95	0.90
3	53	0.94	0.88	0.91
4	63	0.83	0.91	0.87
5	63	0.93	0.91	0.93
Total	**310**	**0.90**	**0.94**	**0.90**

Conclusion

The correlation between DIBELS and Moby.Read is at the ceiling of what would be expected given the reliability of DIBELS. These data demonstrate the close agreement of Moby.Read's machine-scores of Accurate Rate (WCPM) with the WCPM scores from DIBELS, a commonly used measure of oral reading fluency.

The reliability of the machine-generated WCPM scores from Moby.Read is acceptably high (0.91) and higher than that of the criterion test, DIBELS.

These two results lend more support for the assertion that machine-scores for oral reading fluency, ORF, can be valid for use in applications where DIBELS scores can be justifiably applied.

Next Steps

1. Perform a validation study of machine-scored WCPM directly with silent-read comprehension tests – over time, with a wider range of students.
2. Perform validation on other machine oral fluency scores, such as *expression* and *comprehension*.
3. Explore similar machine scoring methods for diagnostic use with clinical populations.

We gratefully acknowledge our collaboration with John Sabatini at ETS (now at U. Memphis), and we thank Matt Mares, Ariel Sabatini and Laura Cook for assistance with test administration and scoring.

References

[1] J. Bernstein, J.Cheng, J. Balogh, and R. Downey, (2018) Artificial intelligence for scoring oral reading fluency. in Applications of artificial intelligence to assessment, H. Jiao and R. W. Lissitz, Eds. Charlotte, NC: Info Age.

[2] J. Cheng, "Real-Time Scoring of an Oral Reading Assessment on Mobile Devices". Proc. Interspeech 2018, 1621-1625, DOI: 10.21437/Interspeech.2018-34.

[3] J. Bernstein, J. Sabatini, J. Balogh, J. Cheng (2017) "Oral Reading Assessment: Leveled Fluency, Self-Administered and Automatically Scored". California Educational Research Association (CERA) 96th Annual Conference, Anaheim.

[4] Moby.Read: https://youtu.be/_V6_7agY5tc

[5] Good, R., Kaminski, R., et alia. (2011). DIBELS Next assessment manual. Eugene, OR: Dynamic Measurement Group.

[6] Measurement Group (2018). DIBELS: https://dibels.org/dibels.html Retrieved February 15, 2018.

[7] Goffreda, C. & DiPerna, J. (2010). An empirical review of psychometric evidence for the Dynamic Indicators of Basic Early Literacy Skills. School Psychology Review, 39(3), 463-483.

Acknowledgment: The research described here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through contract ED-IES-16-C-0004 to Analytic Measures Inc. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Contact: Masanori.Suzuki@AnalyticMeasures.com