A Summary of An Effort to Improve the Practical and Psychometric Properties of Traditional ORF

Joseph F. T. Nese, Ph.D.

University of Oregon

jnese@uoregon.edu

CORE

Computerized Oral Reading Evaluation

CORE is a computerized ORF assessment framework that combines multiple shorter passages, automatic speech recognition (ASR) for scoring, and a latent variable psychometric model to produce model-based WCPM scores

Institute of Education Sciences Grant R305A140203

Three CORE research studies

Purpose

1. Why?

2. How?

3. What?

4. Where?

Traditional ORF Assessment

- a student reads aloud for 1 min (passage ~250 words)

- an assessor marks each word the student reads incorrectly (including omissions)

- **WCPM** - outcome is **w**ords read **c**orrectly **p**er **m**inute (total words - incorrect words)

Advantages

- essential part of reading proficiency [1]

- used in classrooms across the U.S.

- quick for one student

- indicator of reading comprehension & general reading achievement [2]–[7]

Limitations

1. opportunity cost (not quick for many students) [8]

2. administration errors [9]–[12]

3. passage inequivalence [13]

4. score reliability and precision [14], [15]

5. consequential validity [16]

Limitations 1 & 2

1. opportunity cost

2. administration errors

Approach: ASR

- groups (or entire classrooms) to be assessed at one time with one educator

- standardized delivery, setting, & scoring

  - timing the reading for exactly 60 s

  - accurately calculating WCPM

  - recording the correct WCPM score

Study 1

Nese, J. F. T., & Kamata, A. (2021). Evidence for automated scoring and shorter passages of CBM-R in early elementary school. *School Psychology, 36*, 47-59. doi: 10.1037/spq0000415

What we did

902 students across Grades 2-4 (from 4 schools)

Read 13,766 passages

- 1 Traditional ORF passage (easyCBM)

- 14-18 CORE passages

    - 25-85 words in length (short, medium, long)

Scored by:

We wanted to know

1. Do WCPM scores differ by passage *length* ($\times$ 4)?

We found

None of the WCPM pairwise comparisons (across grades and scoring methods) between the traditional ORF passages and the shorter CORE passages were statistically significantly different.

The shorter passages read in their entirety functioned similarly to the traditional ORF passages read for 1 min

We also wanted to know

2. Do **WCPM** scores differ by *scoring method* ($\times$ 3)?

We found

Human criterion vs:

- **ASR**: all 12 pairwise comparisons statistically different
- **Traditional**: 8 of 12 pairwise comparisons statistically different

ASR <u>and</u> traditional scores were different from criterion scores, <u>but</u> most of these differences were below the "acceptable threshold"[17]

We also *also* wanted to know

3. Do **word** scores differ by *scoring method* ($\times$ 3)?

We found

Average agreement rates between the human criterion and:

- **Traditional**: .97 to .99

- **ASR**: .81 to .94 (all but one > .87)

Human-to-ASR degradation from the human-to-human score agreement was generally below the "acceptable threshold" [17]

Conclusions

- Shorter passages can be used for ORF assessment

- ASR can be used to score ORF assessment

  - ASR applied here is now 6+ years old

Model-based WCPM

Kara, Y., Kamata, A., Potgieter, C., & Nese, J. F. T. (2020). Estimating model-based oral reading fluency: A Bayesian approach. *Educational and Psychological Measurement, 80,* 847-869. doi: 10.1177/0013164419900208

Model-based WCPM

- Two-part latent-variable psychometric model

- Passages equated, horizontally scaled, and vertically linked

  - scores on a common scale, especially useful for progress monitoring

- Same metric as traditional ORF scores (i.e., WCPM)

  - immediately usable for teachers and reading specialists

- *SE*s computed for each WCPM score (*CSEM*)

Limitations 3 & 4

3. passage inequivalence

4. score reliability and precision

Approach: Model-based WCPM (Kara et al., 2020)

Two-part latent-variable psychometric model (jointly modeled and estimated)

1. Accuracy - binomial-count factor model (number of passage words read correctly)

2. Speed - log-normal factor model (passage reading time)

Study 2

Nese, J. F. T. & Kamata, A. (2021). Addressing the large standard error of traditional CBM-R: Estimating the conditional standard error of a model-based estimate of CBM-R. *Assessment for Effective Intervention, 47*, 53-58. doi: 10.1177/1534508420937801

What we did

1,021 students across Grades 2-4 (from 4 schools)

Read 4,084 passages

- 10 to 12 CORE passages (fixed sets randomly assigned)
  - 50-85 words in length (medium or long)
- Across 4 occasions (Oct, Nov, Feb, May)

Scored by ASR

Model-based WCPM scores

We wanted to know

1. Is the average conditional standard error of measurement (*CSEM*) for the model-based WCPM estimates smaller than the reported *SEM*s of traditional ORF systems?

We found

ORF Measure	Grade 2	Grade 3	Grade 4
SEM			
aimsweb Plus	7.78	7.46	8.4
DIBELS 8th Edition	7.84	9.59	9.63
easyCBM	--	9.73	--
FastBridge	8.54	8.54	10.41
CSEM			
CORE Model-based WCPM	5.15	5.47	7.63

We also wanted to know

2. What about for students at/below the 25th percentile?

We found

Grade	Unique Students	Number of CBM-R Tests	Percent < 8 WCPM	Mean CSEM
Grade 2	122	251	99.6	3.0
Grade 3	127	284	98.9	3.3
Grade 4	125	253	98.0	4.8

Conclusions

Preliminary but promising evidence that the model-based WCPM scale scores can increase score reliability and precision

- better identify students at risk of poor reading outcomes

- better determine student responsiveness to intervention

Limitation 5

5. consequential validity

Approach: <u>CORE vs. Traditional ORF</u>

Apply ASR, real-world setting, and model-based WCPM scores and compare growth estimates and predictive performance to Traditional ORF assessment

Study 3

Nese, J. F. T. (2022). Comparing the Growth and Predictive Performance of a Traditional Oral Reading Fluency Measure With an Experimental Novel Measure. *AERA Open*, *8*, 1-19. doi: 10.1177/23328584211071112

What we did

2,108 students across Grades 2-4 (7 schools)

- across 4 occasions (Oct, Nov, Feb, May)

- 10 to 12 CORE passages
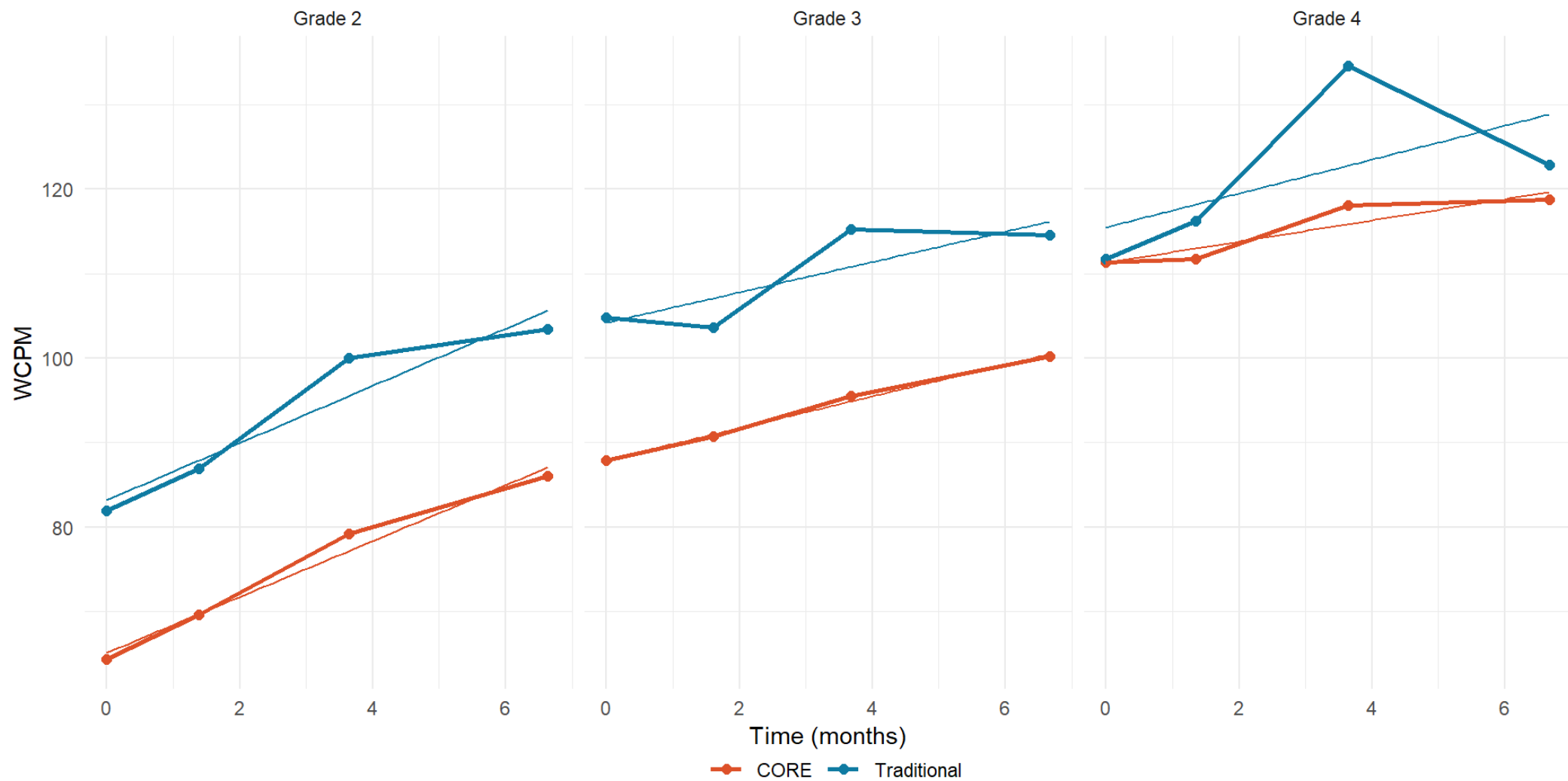
- 1 traditonal ORF (easyCBM)

- all scored by ASR

We wanted to know

CORE vs. Traditional ORF

1. Which has better within-year growth properties?

- the standard error of the slope (*SEb*) estimates
- the reliability of each measurement occasion

Mean Standard Error of the Slope (SEb)

	CORE		Traditional		
Grade	Mean SEb	SD	Mean SEb	SD	d
2	2.82	2.36	3.93	3.04	0.41
3	2.88	2.36	4.32	3.38	0.55
4	3.16	2.46	—	—	—

Reliability Estimates across Occasions

		Reliability		
Occasion	CORE	Traditional		h
Grade 2				
1	.91	.78		0.36
2	.90	.83		0.20
3	.84	.62		0.52
4	.85	.86		-0.03
Grade 3				
1	.93	.79		0.42
2	.86	.70		0.39
3	.86	.74		0.30
4	.84	.79		0.11
Grade 3				
1	.91	—		—
2	.87	—		—
3	.87	—		—
4	.82	—		—

We found

Model-based CORE scores

- more precise individual slope parameter estimates

- higher reliability at measurement occasions (LGM)

- latent slope means measured with less variance

Conclusion

Model-based CORE scores may provide better data with which to make instructional decisions, such as risk status or responsiveness to instruction

We also wanted to know

CORE vs. Traditional ORF

2. Which has better distal (fall) and proximal (spring) predictive performance:

- spring comprehension scores (easyCBM) Grades 2-4
- spring state reading test scores & proficiency Grades 3-4

We found

Predictive framework, training/test data, 10-fold cross-validation, and OLS models

Across all comparisons - grade, outcome, loss function, and distal and proximal predictors

Model-based CORE scores:

- lower *RMSE*

- higher R^2 , sensitivity, specificity, and *AUC*

Model-based CORE scores had a stronger relation with year-end reading comprehension and state reading test scores than traditional ORF scores

Conclusion

The **model-based CORE scores**, with a **stronger relation with important reading outcomes**, can potentially better help with **early identification** of at-risk students and potentially better help **monitor the ORF progress** of those at-risk students because the scores provide a **better estimate** of students' current and prospective reading proficiency

Summary

1. Evidence to support ASR scoring and shorter ORF passages [18]

2. Model-based WCPM scale scores have a lower *CSEM*, better suited for screening and progress monitoring, and thus better educational decisions [19]

3. CORE improved within-year growth properties and predictive performance, providing preliminary evidence for consequential assessment [20]

Where are we going from here?

CORE + Prosody

- Model-based ORF with accuracy, speed, **&** prosody

- Prosody estimate with machine learning

- Grant R305A200018

Computational Tools for Model-Based ORF

- R package for model-based WCPM scores

- "testlet" model (sentences within passages)

- Grant R305D200038

Acknowledgements

Dr. Aki Kamata

Dr. Cornelis Potgieter

Dr. Yusuf Kara

Questions?

jnese@uoregon.edu

[slides will be posted here]

References

[1]   N. R. P. (US), N. I. of Child Health, and H. D. (US), *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health; Human Development, National …, 2000.

[2]   L. S. Fuchs, D. Fuchs, M. K. Hosp, and J. R. Jenkins, "Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis," *Scientific studies of reading*, vol. 5, no. 3, pp. 239–256, 2001.

[3]   R. H. Good, K. A. Powell-Smith, M. Abbott, E. N. Dewey, A. N. Warnock, and D. VanLoo, "Examining the association between DIBELS next and the SBAC ELA achievement standard," *Contemporary School Psychology*, vol. 23, no. 3, pp. 258–269, 2019.

[4]   J. R. Jenkins, L. S. Fuchs, P. Van Den Broek, C. Espin, and S. L. Deno, "Sources of individual differences in reading comprehension and reading fluency." *Journal of educational psychology*, vol. 95, no. 4, p. 719, 2003.

[5]   J. F. T. Nese, B. J. Park, J. Alonzo, and G. Tindal, "Applied curriculum-based measurement as a predictor of high-stakes assessment: Implications for

researchers and teachers," *The Elementary School Journal*, vol. 111, no. 4, pp. 608–624, 2011.

[6]   S. G. Schilling, J. F. Carlisle, S. E. Scott, and J. Zeng, "Are fluency measures accurate predictors of reading achievement?" *The Elementary School Journal*, vol. 107, no. 5, pp. 429–448, 2007.

[7]   J. Shin and K. McMaster, "Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis," *Journal of school psychology*, vol. 73, pp. 131–149, 2019.

[8]   A. R. Hoffman, J. E. Jenkins, and S. K. Dunlap, "Using DIBELS: A survey of purposes and practices," *Reading Psychology*, vol. 30, no. 1, pp. 1–16, 2009.

[9]   K. D. Cummings, G. Biancarosa, A. Schaper, and D. K. Reed, "Examiner error in curriculum-based measurement of oral reading," *Journal of School Psychology*, vol. 52, no. 4, pp. 361–375, 2014.

[10]   S. Munir-McHill, T. Bousselot, K. Cummings, and J. Smith, "Profiles in school-level data-based decision making," 2012.

[11]   D. K. Reed, K. D. Cummings, A. Schaper, and G. Biancarosa, "Assessment fidelity in reading intervention research: A synthesis of the literature," *Review of Educational Research*, vol. 84, no. 2, pp. 275–321, 2014.

[12]   D. K. Reed and K. M. Sturges, "An examination of assessment fidelity in the administration and interpretation of reading tests," *Remedial and Special Education*, vol. 34, no. 5, pp. 259–268, 2013.

[13] D. J. Francis, K. L. Santi, C. Barr, J. M. Fletcher, A. Varisco, and B. R. Foorman, "Form effects on the estimation of students' oral reading fluency using DIBELS," *Journal of School Psychology*, vol. 46, no. 3, pp. 315–342, 2008.

[14] T. J. Christ and B. Silberglitt, "Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency," *School Psychology Review*, vol. 36, no. 1, pp. 130–146, 2007.

[15] B. C. Poncy, C. H. Skinner, and P. K. Axtell, "An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement," *Journal of Psychoeducational Assessment*, vol. 23, no. 4, pp. 326–338, 2005.

[16] E. S. Shapiro, "Commentary on progress monitoring with CBM-r and decision making: Problems found and looking for solutions." *Journal of School Psychology*, vol. 51, no. 1, pp. 59–66, 2012.

[17] D. M. Williamson, X. Xi, and F. J. Breyer, "A framework for evaluation and use of automated scoring," *Educational measurement: issues and practice*, vol. 31, no. 1, pp. 2–13, 2012.

[18] J. F. Nese and A. Kamata, "Evidence for automated scoring and shorter passages of CBM-r in early elementary school." *School Psychology*, vol. 36, no. 1, p. 47, 2021.

[19] J. F. Nese and A. Kamata, "Addressing the large standard error of traditional CBM-r: Estimating the conditional standard error of a model-based estimate

of CBM-r," *Assessment for Effective Intervention*, vol. 47, no. 1, pp. 53–58, 2021.