The title

Joseph F. T. Nese[1]

[1] University of Oregon

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. Joseph F. T. Nese: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Joseph F. T. Nese, 275 Education, 5262 University of Oregon, Eugene, OR 97403-5262. E-mail: jnese@uoregon.edu

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

The title

Data

Research Questions

The purpose of this study is to compare the consequential validity properties of CORE and a traditional ORF assessment (easyCBM) for students in Grades 2 through 4.

(1) Comparing traditional CBM-R and model-based CORE scores, which has better within-year growth properties, including (a) the standard error (*SE*) of the intercept and slope estimates, and (b) the reliability of each measurement occasion?

(2) Comparing traditional CBM-R WCPM scores and CORE model-based fluency scores, which has better distal (fall) and proximal (spring) predictive accuracy for spring CBM comprehension scores for students in Grades 2 through 4?

(3) Comparing traditional CBM-R WCPM scores and CORE model-based fluency scores, which has better distal (fall) and proximal (spring) predictive accuracy for spring state reading test scores for students in Grades 3 and 4?

Method

This study was conducted in the 2017-18 and 2018-19 school years in Oregon and Washington, with institutional IRB approval. The 2017-18 study was replicated in 2018-19 to increase the student sample size. That is, the sample was the only difference between the two years. The study consisted of a longitudinal design with four repeated measurement occasions (waves) to address the research questions.

Participants

The original sample included 2,519 students from four school districts and seven elementary schools in Oregon and Washington (four schools participated in both years, and three schools only in 2018-19). All students in Grades 2 through 4 at the seven participating schools were invited to participate such that the sample would be representative, to the extent possible, of typically developing students across reading proficiency levels.

According to 2018-2019 NCES school data, the populations of the seven schools ranged from 357 to 759 students, approximately half of whom were students in Grades 2 through 4. Four school locales were classified as Suburb: Midsize, and three as Town: Distant (for more information, see https://nces.ed.gov/ccd/commonfiles/glossary.asp). Six schools received Title I funding, and the percentage of students receiving free or reduced lunch ranged from 49% to 86%. The ethnic/race majority for all schools was White (56% to 76%), followed by Hispanic (16% to 34%), Multi-racial (3% to 9%), American Indian/Native Alaskan (0% to 5%), Asian (0% to 1%), Black (0% to 1%), and Native Hawaiian/Other Pacific Islander (0% to 1%).

We removed extreme WCPM scores that suggested they were not a part of the data generating process, rather an artifact of the audio data collection process. We removed WCPM scores that were based on less than 30 sec of audio because (a) traditional CBM-R scores are intended to be 60 sec, and (b) CORE scores are intended to be based on reading 10 to 12 passages and it appears implausible to do that in 30 sec. We also removed WCPM scores that were based on less than 10 words read, which applied only to traditional CBM-R scores. We acknowledge that other researchers may have made different theoretical data decisions.

The analytic sample varied according to the research question and outcome variable. Table **??** shows the sample demographic characteristics for each research question (RQ).

The analytic sample for longitudinal analysis of WCPM (RQ 1) included 2,108 students (84% of the original sample) who had at least one wave of data for each of the Traditional CBM-R and the CORE WCPM scores; 601 Grade 2, 770 in Grade 3, and ‘r 737 were in Grade 4. Approximately 6% of students were missing demographic data but 27% of students were missing EL data as one state did not provide EL data for 2017-18.

Of the 2,108 students in the longitudinal analysis, only 987 (47%) had scores both fall and spring scores on the traditional CBM-R and CORE assessments, which limited the sample size for subsequent RQs. The analytic sample for RQ 2 were the 427 students (43%) that had a score on the spring CBM comprehension assessment. Note that one school district (District 2, Schools B and E) did not administer the spring CBM comprehension assessment, which further limited the sample. The analytic sample for RQ 3 were the 722 students (73%) that had a score on the SBAC ELA/L test. Note that Grade 2 students do not take the year-end state test.

Measures

[1]Time, in months, between waves; also the latent slope factor loadings.

Table ?? shows the descriptive WCPM data, and Figure 1 shows the WCPM means for each wave.

CORE CBM-R. Each CORE passage is an original work of narrative fiction that follows the story grammar of English language short stories, with a main character and a clear beginning, middle, and end (link blinded for review). To reduce construct-irrelevant variance associated with different authors' voice and style, the author of the CORE passages was part of the team that authored the easyCBM traditional CBM-R passages used in this study. Apart from the passage length requirements, the CORE passages were written to similar specifications as the easyCBM passages. Each CORE passage was written within 5 words of a targeted length: long = 85 words or medium = 50 words. Ultimately, 150 passages were written: 50 at each of Grades 2-4, with 20 long passages and 30 medium passages for each grade.

Administration instructions were to allow students to read the CORE passages in their entirety, but a time limit was set at 90 s. At each wave, sample students read on average 8.40 passages (*SD* = 1.80; range = 1 - 12).

The CORE scores are model-based estimates of WCPM, based on a recently proposed latent-variable psychometric model of speed and accuracy for CBM-R data (Kara, Kamata, Potgieter, & Nese, 2020). The model-based CBM-R WCPM estimates are based on a two-part model that includes components for reading accuracy and reading speed. The accuracy component is a binomial-count factor model, where accuracy is measured by the number of correctly read words in the passage. The speed component is a log-normal factor model, where speed is measured by passage reading time. Parameters in the accuracy and speed models are jointly modeled and estimated. For a detailed description, please see Kara et al. (2020).

Traditional CBM-R. We administered the easyCBM (Alonzo, Tindal, Ulmer, & Glasgow, 2006) oral reading fluency measures as the traditional CBM-R assessments for the purpose of comparison to CORE passages. easyCBM CBM-R passages range from 200 to 300 words in length and are original works of fiction developed to be of equivalent difficulty for each grade level following word-count, grade-level guidelines (e.g., Flesch-Kincaid readability estimates), and form equivalence empirical testing using repeated measures ANOVA to evaluate comparability of forms (Alonzo & Tindal, 2007). The easyCBM CBM-R measures have demonstrated features of technical adequacy that suggest they are sufficient to meet the needs as the comparative example of an existing traditional CBM-R assessment (Anderson et al., 2014). The reported alternate form reliability across passages ranged from .83 to .98, test-retest reliability ranged from .84 to .96, and G-coefficients ranged from .94 to .98 (Anderson et al., 2014). Predictive (fall, winter) and concurrent (spring) relations between Grade 2 CBM-R and spring SAT-10 reading scale scores were .59 to .62, and .66 respectively (Anderson et al., 2014).

Following standard administration protocols, students were given 60 s to read the traditional CBM-R passages.

ASR Scoring.

The ASR engine scored each audio recording file, scoring each word as read correctly or incorrectly, and recording the time in centi-seconds to read each word and the time between words. Bavieca, an open-source speech recognition toolkit, was the ASR applied in this study (http://www.bavieca.org/). Bavieca uses continuous density hidden Markov models and supports maximum likelihood linear regression, vocal tract length normalization, and discriminative training (maximum mutual information). It uses the general approach of many state-of-the art speech recognition systems: a Viterbi Beam Search used to find the optimal mapping of the speech input onto a sequence of words. The score for a word sequence was calculated by interpolating language model scores and acoustic model scores. The language model assigned probabilities to sequences of words using trigrams (where the probability of the next word is conditioned on the two previous words) and was trained using the CMU-Cambridge LM Toolkit (Clarkson & Rosenfeld, 1997). Acoustic models were clustered triphones based on Hidden Markov Models using Gaussian Mixtures to estimate the probabilities of the acoustic observation vectors. The system used filler models to match the types of disfluencies found in applications.

CBM Comprehension. The easyCBM comprehension assessment contains 12 (Grade 2) or 20 (Grades 3 and 4) multiple-choice items assessing students' comprehension of a 1,500 word fictional narrative. The comprehension items are designed to target students' literal (7 items), inferential (7 items), and evaluative (6 items) comprehension. Split-half reliability ranged from .38 to .87, item reliability from Rasch analyses ranged from .39 to .94, and Cronbach's alpha ranged from .69 to .78 (Saez et al., 2010). Predictive (fall) and concurrent (spring) correlations between Grade 2 CBM comprehension and spring SAT-10 reading scale scores were .62 and .66 respectively (Jamgochian et al., 2010). Predictive (fall) and concurrent (spring) correlations between Grade 3 and 4 CBM comprehension and spring state reading test scores (Oregon Assessment of Knowledge and Skills [OAKS] and Washington Measures of Student Progress [MSP]) were .52 to .70, and .37 to .68 respectively (Anderson et al., 2014). Predictive diagnostic statistics for fall CBM comprehension and spring state reading test scores included sensitivity from .68 to .86, specificity from .57 to .92, and AUC from .74 to .86. Concurrent diagnostic statistics for spring CBM comprehension and spring state reading test scores included sensitivity from .69 to .89, specificity from .63 to .80, and AUC ranged from .76 to .87 (Anderson et al., 2014).

MEAN/SD ETC

SBAC Reading Test. The Smarter Balanced Assessment Consortium (SBAC) English language arts/literacy (ELA/L) summative assessment is administered to students in Grades 3 through 8 and 11 and consists of two parts: a computerized adaptive test (CAT), and a performance task (PT) component. The SBAC ELA/L was developed to align to the Common Core State Standards (CCSS) and measures four broad clams: reading, writing, listening, and research (Consortium, 2020). Within each claim there are a number of assessment targets, and each test item is aligned to a specific claim and target and to a CCSS (CITE). The CAT consisted of selected response items that assess all four claims. The PT consisted of a set of related stimuli presented with two or three research items requiring both short-text responses and a full written response that assess the writing and research claims. The overall SBAC ELA/L performance scaled score is divided into four proficiency categories, Well Below, Below, Proficient, and Advanced, where the third and fourth categories designate meeting grade-level achievement standards.

MEAN/SD PERCENT PASSING, ETC

Procedure

Students were assessed online, using classroom or school devices, and wore headphones with an attached noise-cancelling microphone provided by the research team. Students were given a task introduction by their teacher, and then directed to the study website where the first page asked for student assent (if a student declined, their participation ended). The standardized instructions were presented via audio as well as print. *"Get ready! You are about to do some reading! After pressing start, read the story on the screen. When you are finished click done. Do your best reading, and have fun!"*

For each of the four measurement occasions (Oct-Nov 2017, 2018; Nov-Feb 2017-18, 2018-19; Feb-Mar 2018, 2019; May-Jun, 2018, 2019), students read aloud online a randomly assigned, fixed set of 10 to 12 CORE passages (3-5 long and 5-7 medium), and one traditional CBM-R passage from the easyCBM progress monitoring system.

An automatic speech recognition engine scored each reading, scoring each word as read correctly or incorrectly (accuracy), and recording the time duration to read each word (and the silence between) which was aggregated to calculate the time to read the passage (speed).

All WCPM scores were based on these readings and data. The model-based WCPM CORE scores (Kara et al., 2020) were estimated for each measurement occasion based on the number CORE passages read. Traditional CBM-R WCPM scores were calculated by dividing the number of words read correctly (wrc) by the quotient of the total seconds read (sec) and 60 (i.e., $wrc/(sec/60)$).

Analyses

To address RQ 1, we applied a latent growth model (LGM; Meredith and Tisak (1990)) separately for each grade to represent students' within-year oral reading fluency growth. The linear time covariate was specified as the elapsed number of months between the median month at wave t and the median month of t_1 (see Table **??**).

Two results are extracted from the LGMs to compare the growth properties of the traditional CBM-R and model-based CORE scores. One, the fixed intercept and slope estimates and their associated standard errors (SE), as estimated by the linear growth model. Two, the reliability of the CBM-R scores at each wave, as estimated by the proportion of true score variance to observed score variance (Rogosa & Willett, 1983; Singer, Willett, Willett, & others, 2003; Willett, 1988). EDIT

$$\rho_t = \frac{\psi_{00} + \lambda_t^2\psi_{11} + 2\lambda_t\psi_{01}}{\psi_{00} + \lambda_t^2\psi_{11} + 2\lambda_t\psi_{01} + \theta_t} = \frac{var(y_t) - \theta_t}{var(y_t)}$$

Where ρ_t represent the reliability at wave t, ψ represents the covariance structure of the intercept and slope factors, λ_t represents the linear time covariate, and θ_t represents the residual variance at a wave, which is equivalent to the ratio of the true score variance

$(var(y_t) - \theta_t)$ to the observed score variance $(var(y_t))$, and can be calculated for each wave by subtracting the residual variance (measurement error) from the observed score variance. This estimate of reliability provides both the true score variance explained by the longitudinal model and the unique measurement error variance of observed scores at each wave, and has been applied for estimating reliability of CBM data (Yeo, Kim, Branum-Martin, Wayman, & Espin, 2012).

All analyses and figures were conducted and created in the R programming environment (R Core Team, 2020). The LGM analyses were conducted using the lavaan package with maximum likelihood estimation with robust (Huber-White) standard errors and a scaled test statistic that is (asymptotically) equal to the Yuan-Bentler test statistic (Rosseel, 2012). This estimator is robust to non-normality and clustering (McNeish, Stapleton, & Silverman, 2017).

To address RQs 2 and 3, we apply a predictive approach to determine which CBM-R predictor most accurately estimates the outcomes, rather an inferential approach that pursues unbiased estimates of β coefficients. Our predictive model is a linear model, separate for by grade and CBM-R predictor, regressing the outcome (spring CBM comprehension, SBAC ELA/L scores, or SBAC ELA/L proficiency) on the CBM-R predictor (traditional CBM-R scores vs. CORE model-based scores from fall or spring).

For RQ 2, we fit 12 linear models: 2 CBM-R predictors each at 2 seasons (fall and spring) for each of 3 grades: $Comprehension_i = \beta_0 + \beta_1 CBM\text{-}R_{season} + \epsilon_i$.

For RQ 3, we model Grades 3 and 4 together and thus include grade level as a categorical covariate, as well as the state (to account for differences in standards). We fit 8 linear models, applying a logistic regression for the categorical SBAC ELA/L proficiency outcome: $SBAC_i = \beta_0 + \beta_1 CBM\text{-}R_{season} + Grade + State + \epsilon_i$.

To measure the accuracy of the models, our predictive performance measures were the $RMSEA$ and R^2 for the continuous outcomes (spring CBM comprehension and SBAC ELA/L scores), and the Receiver Operating Characteristic (ROC) area under the curve

(AUC) for the categorical outcome (SBAC ELA/L proficiency).

To understand the predictive accuracy of the CBM-R measures, and how their accuracy might generalize to new data, we split the data (by RQ) into two sets: a training set, a random sample of 75% of the data; and a test set, the remaining 25% of the data.

To get a measure of variance for the performance measures, we apply 10 fold cross-validation to the training set. For each fold, 10% of the training set is sampled and serves as an assessment sample, so that each observation serves in one and only one assessment sample. The remaining 90% of the training set serve as the analysis sample for a fold. The predictive model (Eq 2) is fit on the 90% analysis sample of each fold, and the resulting model parameters are used to predict the assessment sample within each fold. The performance measures (*RMSEA* and AUC) are taken from each fold, and the final performance is the mean performance measure across the 10 folds, and the of variance around the mean.

Research has shown that 10 folds is a sensible value for k-fold cross-validation, and repeating k-fold cross-validation can improve the accuracy of the estimates while maintaining small bias, particularly for smaller sample sizes (Kim, 2009; Molinaro, Simon, & Pfeiffer, 2005). We apply 10-fold cross-validation repeated 5 times for each RQ so that we fit 50 models and record 50 performance measures to the training set (10 folds $\times$ 5 repeats = 50).

Finally, we fit the predictive models to the entire training set, and then use the resulting model parameters to predict the results of the test set. The test set here can be can be conceptualized as "new" (or unseen) data, as it has not been used to this point. The resulting final performance measures serve as estimates of how the two comparison CBM-R measures will generalize in their predictive accuracy.

The predictive modeling process was conducted using the tidymodels package (Kuhn & Wickham, 2020). We also used the following R packages: doParallel (**???**), ggridges(**???**), ggthemes (**???**), gt (**???**), janitor (**???**), lavaan (Rosseel, 2012), papaja

(**???**), patchwork (**???**), tidyverse (**???**).

Results

RQ1. To address RQ 1, we fit LGMs separately for each CBM-R measure and grade.

The fit measures for the Grade 2 CORE LGM were $\chi^2 = 13.70$ with $df = 5$ ($p =$ .018), Tucker–Lewis fit (TLI) = 1, Comparative Fit Index (CFI) = 1, $RMSEA = 0.04$, and BIC = 17,986.3. The fit measures for the Grade 2 Traditional LGM were $\chi^2 = 56.40$ with $df = 5$ ($p < .001$), TLI = 0.93, CFI = 0.94, $RMSEA = 0.13$, and BIC = 13,647.1. The fit measures for the Grade 3 CORE LGM were $\chi^2 = 9.20$ with $df = 5$ ($p = .100$), TLI = 1, CFI = 1, $RMSEA = 0.03$, and BIC = 23,365.1. The fit measures for the Grade 3 Traditional LGM were $\chi^2 = 65.10$ with $df = 5$ ($p < .001$), TLI = 0.96, CFI = 0.96, $RMSEA = 0.11$, and BIC = 19,956.8. The fit measures for the Grade 4 CORE LGM were $\chi^2 = 28.50$ with $df = 5$ ($p < .001$), TLI = 0.99, CFI = 0.99, $RMSEA = 0.08$, and BIC = 21,461.1)'.

The Grade 4 LGM for Traditional CBM-R was not successfully estimated without a negative variance for the slope factor. We tried alternate modeling solutions, including homogeneous residual variances (and zero error covariances), heterogeneous Teoplitz residual structure, first-order autocorrelated residuals (McNeish & Harring, 2019), and transformed slope factor loadings, but all models were unsuccessful due to a negative variance or variance-covariance matrix. Thus, we do not report the results from this model.

Table **??** shows the parameter estimates from the LGMs. The *SE* for the mean intercept estimates across grades are slightly larger for the model-based CORE models (1.27 to 1.39) than the traditional CBM-R models (1.25 to 1.31); however, the *SE* for the mean slope estimates for the model-based CORE models (0.11 to 0.13) are about a third of the size as those of the traditional CBM-R models (0.15 to 0.21).

Table **??** shows the observed variances CBM-Rs at each wave, the estimated residual variances from the LGMs, and reliability estimates by grade and wave. Across grades and waves, the reliability estimates were higher for the model-based CORE scores except for Grade 2, wave 4 (.85 vs. .86). The reliability estimates for the model-based CORE scores ranged from .82 to .93, and for the Traditional CBM-R ranged from .62 to .86.

RQ2

To address RQ 2 we used a predictive approach with resampling and fit linear models separate for by grade and CBM-R predictor, regressing the spring CBM comprehension on the CBM-R predictors.

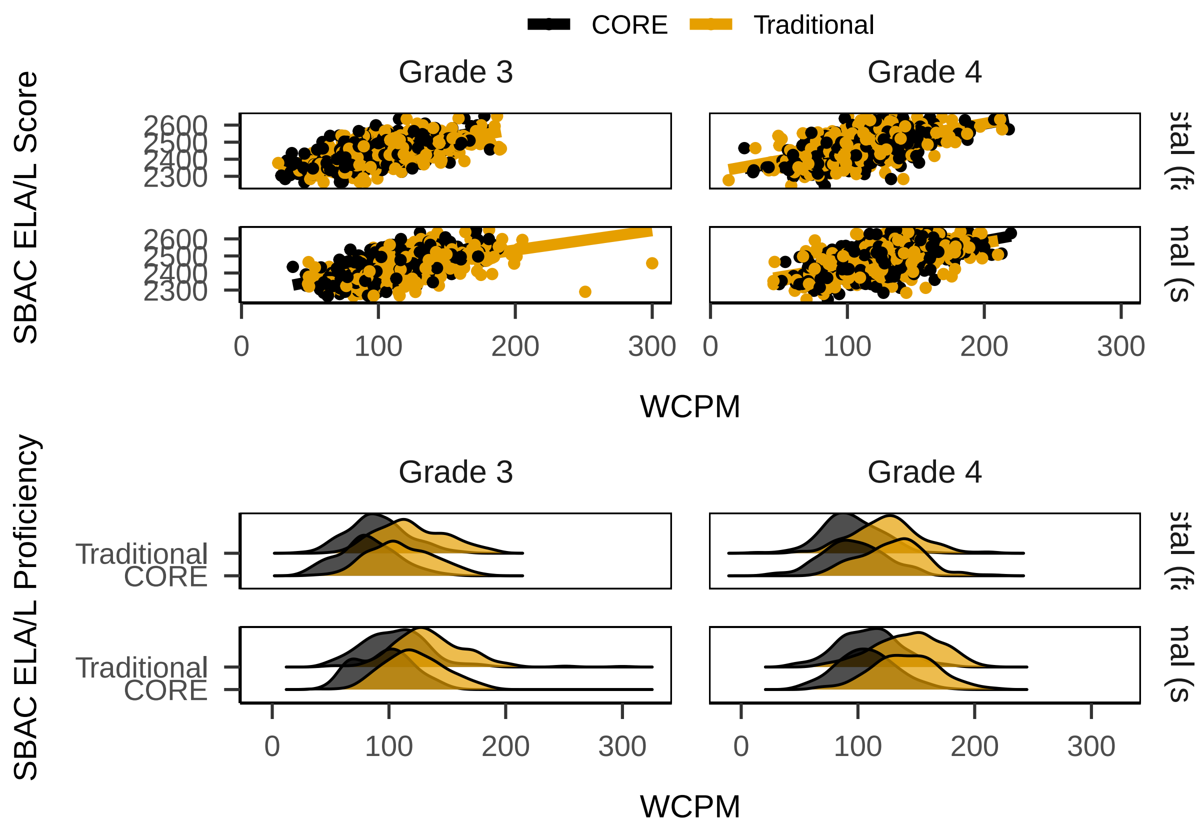
```
## 76.76 sec elapsed
```

For RQ 2 we compared the predictive accuracy of traditional CBM-R and CORE for distal (fall) and proximal (spring) assessments predicting spring CBM comprehension scores for students in Grades 2 through 4. Table **??** shows the mean *RMSE* and R^2 values across the 50 models fit the 10-fold cross-validation samples, as well as the final *RMSE* and R^2 values for the train/test sample. For the distal (fall) CBM-R predictors, the results generally favor CORE, which has better (lower) mean *RMSE* values across grades compared to Traditional CBM-R, and better (higher) mean R^2 values for Grades 3 and 4 (but not Grade 2). For the proximal (spring) CBM-R predictors, the results generally favor traditional CBM-R, which has lower *RMSE* values for Grades 2 and 4 (but not Grade 3), and higher R^2 values across grades. To give context to the *RMSE* values, note that the CBM Comprehension assessment has 12 items for Grade 2 and 20 items for Grades 3 and 4, with *SD*s of 1.69, 4.06, and 3.80, respectively, so the *RMSE* values are generally smaller than the sample *SD*s.

The final *RMSE* and R^2 values in Table **??** represent the parameters of the predictive models applied to the training set (75% of sample) used to predict the test set (25% of sample). For both the distal (fall) and proximal (spring) CBM-R predictors, the results favor CORE, which had lower *RMSE* and higher R^2 values across all comparisons (except Grade 2, distal *RMSE*). The *RMSE* values represent differences of 2% to 7% of a *SD* favoring CORE, and 4% of a *SD* favoring Traditional CBM-R for the Grade 2 distal model. The R^2 values represent increases in explained variance for CORE above Traditional CBM-R of 5% to 82%.

RQ3

To address RQ 3 we used again used a predictive approach with resampling and fit linear models separate for by grade and CBM-R predictor, regressing SBAC ELA/L (score or proficiency) on the CBM-R predictors, grade, and state.




```
## 70.66 sec elapsed
```

For RQ 3 we compared the predictive accuracy of traditional CBM-R and CORE for distal (fall) and proximal (spring) assessments predicting spring SBAC ELA/L (scores and profiency classification) for students in Grades 3 and 4.

Table **??** shows the mean *RMSE* and R^2 values across the 50 models fit the 10-fold cross-validation samples, as well as the final *RMSE* and R^2 values for the train/test sample. For the SBAC ELA/L score (continuous) outcome, the distal results favored CORE which had lower mean and final *RMSE* and higher mean and final R^2 values across grades compared to Traditional CBM-R. To give context to the *RMSE* values, the *SD* of SBAC ELA/L was 79 for Grades 3 and 4 combined, so the *RMSE* values were approximately three-quarters of a *SD*.

For the SBAC ELA/L proficiency (classification) outcome with distal predictors, CORE had higher Accuracy and AUC values across grades compared to Traditional CBM-R. For the proximal predictors, the results were generally comparable. CORE had a slightly higher Mean AUC (0.80 vs. 0.78), Traditional CBM-R had a slightly higher final accuracy (0.69 vs. 0.73), and they had the equivalent Mean Accuracy and Final AUC values.

Discussion

THESE REPRESENT THE VALUES EXPECTED IN A NEW SAMPLE. . . These final performance measures serve as estimates of how the two comparison CBM-R measures may generalize in their predictive accuracy.

References

Alonzo, J., & Tindal, G. (2007). The development of word and passage reading fluency measures for use in a progress monitoring assessment system. Technical report# 40. *Behavioral Research and Teaching.*

Anderson, D., Alonzo, J., Tindal, G., Farley, D., Irvin, P. S., Lai, C.-F., … Wray, K. A. (2014). Technical manual: EasyCBM. Technical report# 1408. *Behavioral Research and Teaching.*

Consortium, S. B. A. (2020). Smarter balanced 2018-19 summative technical report. Retrieved February 24, 2021, from https://technicalreports.smarterbalanced.org/2018-19_summative-report/_book/index.html

Jamgochian, E., Park, B. J., Nese, J. F., Lai, C.-F., Sáez, L., Anderson, D., … Tindal, G. (2010). Technical adequacy of the easyCBM grade 2 reading measures. Technical report# 1004. *Behavioral Research and Teaching.*

Kara, Y., Kamata, A., Potgieter, C., & Nese, J. F. (2020). Estimating model-based oral reading fluency: A bayesian approach. *Educational and Psychological Measurement*, *80*(5), 847–869.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*(11), 3735–3745.

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles.* Retrieved from https://www.tidymodels.org

McNeish, D., & Harring, J. (2019). Covariance pattern mixture models: Eliminating random effects to improve convergence and performance. *Behavior Research Methods*, 1–33.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107–122.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, *21*(15), 3301–3307.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, 335–343.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Saez, L., Park, B., Nese, J. F., Jamgochian, E., Lai, C.-F., Anderson, D., … Tindal, G. (2010). Technical adequacy of the easyCBM reading measures (grades 3-7), 2009-2010 version. Technical report# 1005. *Behavioral Research and Teaching*.

Singer, J. D., Willett, J. B., Willett, J. B., & others. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Willett, J. B. (1988). Chapter 9: Questions and answers in the measurement of change. *Review of Research in Education*, *15*(1), 345–422.

Yeo, S., Kim, D.-I., Branum-Martin, L., Wayman, M. M., & Espin, C. A. (2012). Assessing the reliability of curriculum-based measurement: An application of latent growth modeling. *Journal of School Psychology*, *50*(2), 275–292.

*Table 2.

<i>Mean (SD) WCPM for CBM-R Measures, and Assessment Dates, by Grade and Wave </i>*

	CORE		Traditional			
Wave	Mean	(*SD*)	Mean	(*SD*)	MedianDate	Time (t)[1]
Grade 2						
Wave 1	64.3	(34.4)	81.9	(28.3)	Oct-24	0.00
Wave 2	69.6	(34.3)	86.9	(31.2)	Dec-5	1.38
Wave 3	79.1	(34.8)	100.0	(31.8)	Feb-12	3.65
Wave 4	86.0	(33.2)	103.4	(34.2)	May-14	6.64
Grade 3						
Wave 1	87.9	(35.2)	104.8	(31.8)	Oct-23	0.00
Wave 2	90.7	(35)	103.7	(34.1)	Dec-11	1.61
Wave 3	95.5	(35)	115.3	(35.2)	Feb-12	3.68
Wave 4	100.2	(32.4)	114.5	(34.5)	May-14	6.67
Grade 4						
Wave 1	111.3	(34.6)	111.7	(31.6)	Oct-24	0.00
Wave 2	111.7	(35.8)	116.2	(36)	Dec-4	1.35
Wave 3	118.1	(34.3)	134.5	(34.4)	Feb-12	3.65
Wave 4	118.7	(33.9)	122.8	(33.7)	May-15	6.67

*Table 3.

<i>Latent Growth Model Parameter Estimates by Grade </i>*

	CORE			Traditional		
	Parameter	*SE*	*z*-value	Parameter	*SE*	*z*-value
Grade 2						
Mean Intercept	63.75	1.39	45.86	74.79	1.31	56.89
Mean Slope	3.59	0.13	27.40	4.30	0.21	20.55
Variance Intercept	1070.46	56.82	18.84	694.73	54.94	12.65
Variance Slope	3.04	1.03	2.95	5.25	2.06	2.55
Correlation Intercept-Slope	-0.35	–	–	0.05	–	–
Residual Variance Wave 1	108.15	21.60	5.01	174.89	39.26	4.46
Residual Variance Wave 2	123.28	30.80	4.00	170.13	21.54	7.90
Residual Variance Wave 3	188.05	33.71	5.58	383.15	108.25	3.54
Residual Variance Wave 4	166.29	43.15	3.85	164.71	56.55	2.91
Grade 3						
Mean Intercept	86.86	1.27	68.56	98.34	1.25	78.41
Mean Slope	2.00	0.11	17.69	2.33	0.15	15.06
Variance Intercept	1154.59	61.11	18.89	861.74	72.83	11.83
Variance Slope	2.96	1.20	2.46	0.87	2.57	0.34
Correlation Intercept-Slope	-0.51	–	–	0.25	–	–
Residual Variance Wave 1	86.29	17.68	4.88	211.07	57.28	3.68
Residual Variance Wave 2	170.98	22.35	7.65	345.25	88.15	3.92
Residual Variance Wave 3	175.85	25.57	6.88	325.07	42.81	7.59
Residual Variance Wave 4	173.13	35.41	4.89	245.04	75.52	3.24
Grade 4						

Mean Intercept	109.71	1.30	84.62	–	–	–
Mean Slope	1.67	0.11	15.06	–	–	–
Variance Intercept	1125.18	63.04	17.85	–	–	–
Variance Slope	0.74	1.15	0.64	–	–	–
Correlation Intercept-Slope	-0.44	–	–	–	–	–
Residual Variance Wave 1	103.88	20.96	4.96	–	–	–
Residual Variance Wave 2	167.61	33.84	4.95	–	–	–
Residual Variance Wave 3	149.52	21.61	6.92	–	–	–
Residual Variance Wave 4	207.36	46.01	4.51	–	–	–

Tab

(RM

by G

	CORE					
	Mean *RMSE*	*SE*	Mean R²	*SE*	**Final *RMSE***	<b
Distal						
Grade 2	1.41	(0.07)	0.21	(0.03)	1.96	
Grade 3	3.46	(0.09)	0.24	(0.02)	3.96	
Grade 4	3.06	(0.08)	0.38	(0.03)	2.73	
Proximal						
Grade 2	1.41	(0.07)	0.25	(0.03)	1.89	
Grade 3	3.49	(0.08)	0.23	(0.02)	4.08	
Grade 4	3.21	(0.10)	0.34	(0.03)	2.71	

*Table 6.

<i>SBAC ELA/L Predictive Measures (RMSE and R²) For Distal and Proximal CBM-R Predictors by Grade</i>*

	CORE	Traditional
Distal - SBAC Score		
Mean *RMSE* (*SE*)	60.19 (0.70)	60.54 (0.64)
Mean *R2* (*SE*)	0.40 (0.01)	0.40 (0.01)
Final *RMSE*	62.60	68.03
Final *R2*	0.42	0.31
Proximal - SBAC Score		
Mean *RMSE* (*SE*)	60.56 (0.69)	64.14 (0.96)
Mean *R2* (*SE*)	0.40 (0.01)	0.34 (0.02)
Final *RMSE*	62.44	66.70
Final *R2*	0.42	0.34
Distal - SBAC Proficiency		
Mean Accuracy (*SE*)	0.73 (0.01)	0.72 (0.01)
Mean AUC (*SE*)	0.80 (0.01)	0.79 (0.01)
Final Accuracy	0.73	0.68
Final AUC	0.79	0.75
Proximal - SBAC Proficiency		
Mean Accuracy (*SE*)	0.74 (0.01)	0.74 (0.01)
Mean AUC (*SE*)	0.80 (0.01)	0.80 (0.01)
Final Accuracy	0.73	0.69
Final AUC	0.80	0.78

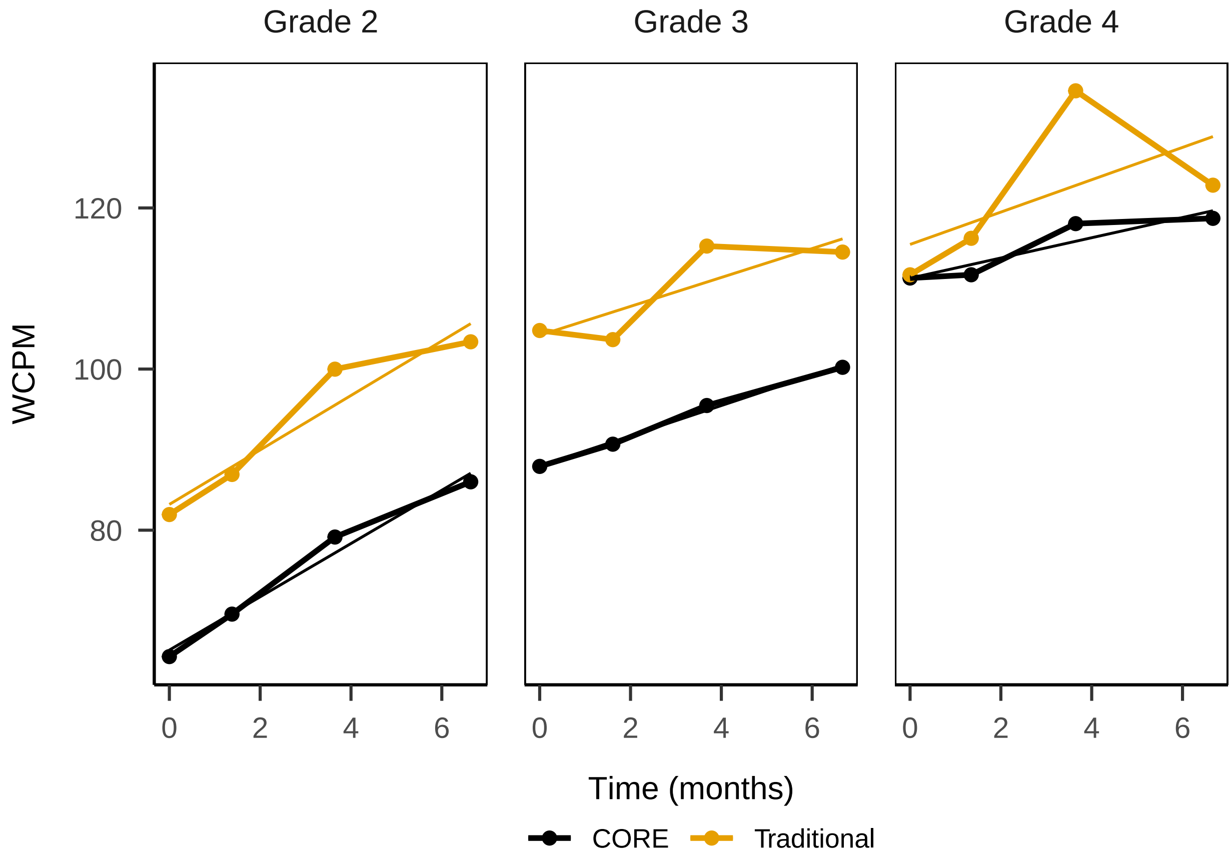


Figure 1. Mean words correct per minute (WCPM) score across waves by grade and CBM-R measure.

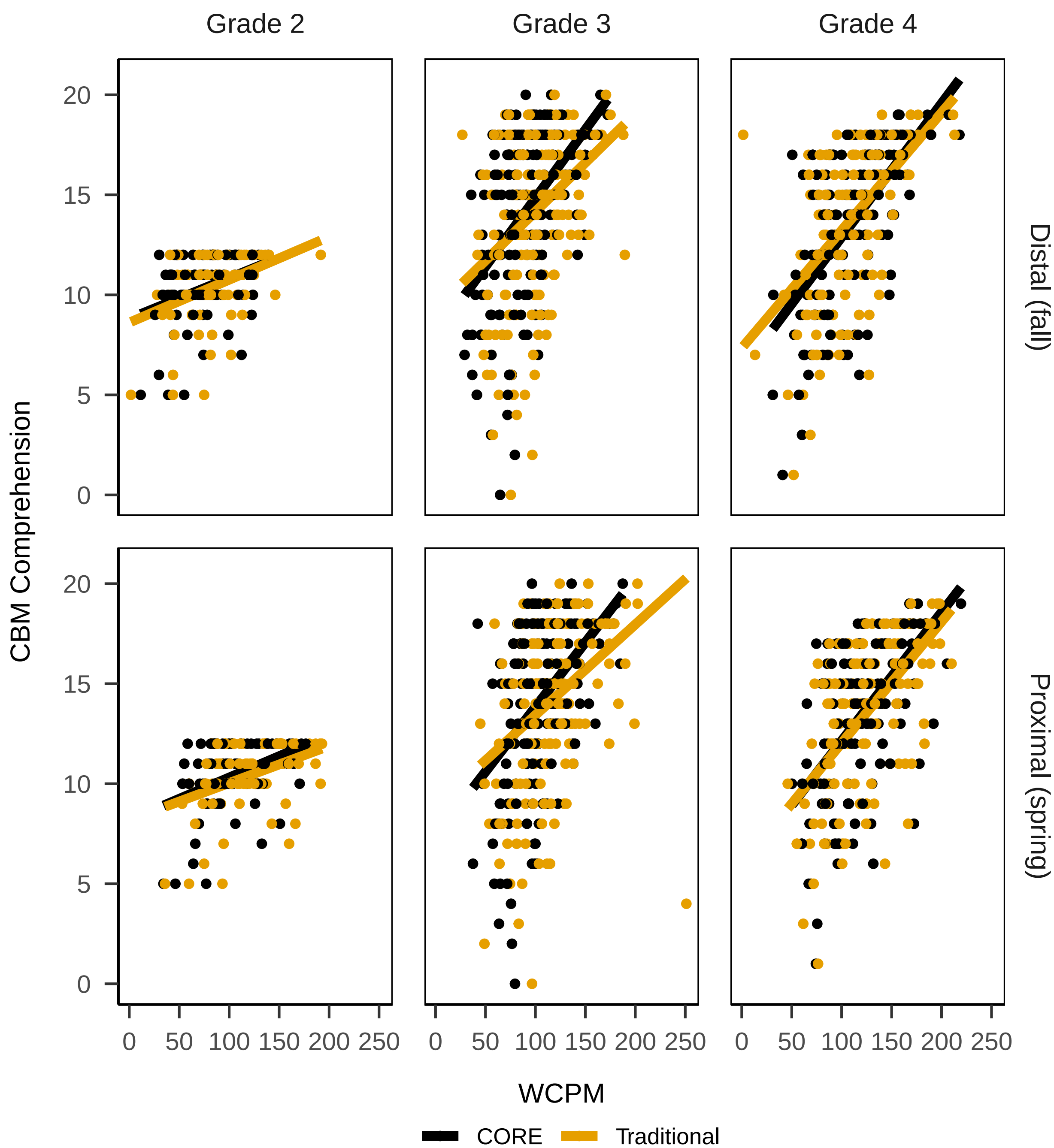


Figure 2. Words correct per minute (WCPM) and CBM Comprehension scores by grade and season, distal (fall) and proximal (spring).