# Roundup 3: Curriculum-Based Measures

2018-05-22

For this roundup, the third installment, I decided to focus on Curriculum-Based Measurement. I've been fielding a lot of questions about progress monitoring within a multi-tiered system of support, especially on the academic side.

The first two articles here discuss CBMs for all students and the last three are ELL specific. All of the summaries are written in the following manner: an Introduction, a description of the methods, a description of the results, and my takeaway. The last section is where I have included some of my other thoughts on the topic occasionally. I tried to keep research limited to the last 2 years, but had to expand my search to find articles about ELLs and CBMs.

Due to the fact that there are common terms throughout many of these articles that might be confusing, and I found myself tempted to explain multiple times, I have included a glossary up front.

As usual, feel free to contact me using any of the methods on the sidebar. A PDF version of this is also on the sidebar.

Enjoy!

## Glossary of terms:

*Curriculum Based Measure (CBM)*: Quick, easy tool to monitor student progress. CBMs can be electronic or paper based. These are designed to be easy to administer and score, as well as quick. They should be administered weekly or every two weeks.

*Maze*: A CBM in which students are given a reading passage with a missing word every seven or so words. At the missing word the students are presented with a list of 3 choices and fill in the blank. See this pdf for more information.

*Daze*: DIBELS Maze. See Maze

For the following terms consider a sample of 200 people given a test to identify a new disease, schoolitis.

| | Actually have the disease | Do not actually have the disease |
|---|---|---|
| Positive test result | 100 | 30 |
| Negative test result | 20 | 50 |

*Sensitivity*: Probability that a test result will be positive when the disease is present (true positive rate). = True positive / (True positive + False negative) =100/100+20 = .833 = 83.3%

*Specificity*: Probability that a test result will be negative when the disease is not present (true negative rate). = True negative / (False positive + True negative) =50/30+50= .625 = 62.5%

*Positive predictive value (or power)*: Probability that the disease is present when the test is positive. = True positive / (True positive + False positive) =100/100+30= .769 = 76.9%

*Negative predictive value (or power)*: Probability that the disease is not present when the test is negative. = True negative / (False negative + True negative) =50/20+50= .714 = 71.4%

So, for this example, this test has a sensitivity of 83%, meaning that doctors can expect the test to correctly identify someone with the disease 83% of the time. It has a specificity of 62.5%, which means that a doctor could expect it to correctly identify the lack of disease 62% of the time. For us, the patients, the predictive value is more helpful because we know the result of the test, not if we actually have the disease, so for this test we can rely on the positive result 77% of the time and the negative result 71% of the time. I don't know about you, but I want a second test!

*Receiver operating characteristic (ROC) Curves*: For an explanation of ROC curves see this explanation. A basic explanation is this: A ROC curve shows the accuracy and usefulness of a given test. These are often used in medicine but also useful in diagnostic testing in education. Two measures, sensitivity and 100-specificity are used to calculate this. This curve uses the sensitivity as a function of 100-specificity, which is the rate of false-positives (test says you have cancer, but you don't). This curve is constructed and the area under the curve tells you the usefulness. If 90-100% of the area is under the curve it is an excellent test, 80-90 is good, 70-80 is fair, 60-70 is poor, and 50-60 is a fail.

# Articles

**Lembke,E., Allen, A., Cohen, D., Hubbuch, C., Landon, D., Bess, J., & Burns, H. (2017). Progress Monitoring in Social Studies Using Vocabulary Matching Curriculum-Based Measurement.**

***Learning Disabilities Research and Practice, 32*(2), 112-120. DOI: 10.1111/ldrp.12130**

This study looked at the validity and reliability of a Vocabulary matching CBM in Social Studies, as well as the ability to measure growth using the same measure. The literature review of this article has a nice description of the history of different types of CBM and the utility of each. Many studies have confirmed the effectiveness of using CBM in elementary school, but very few have looked at the effectiveness at the secondary level, particularly in a content area. For this study the authors do just this with vocabulary matching, which is a CBM that is designed by a teacher, or team of teachers, using vocabulary terms learned in class during that specific time-period. If doing CMB frequently this would be the one or two week period between measures. Terms are gathered using texts, teacher lectures, notes, or any other source of vocabulary assigned to the class by the teacher. Words are placed on the left side and definitions, sometimes with distractors (incorrect definitions) on the right. Students then match the work with the definition in five minutes.

**Design** In this study 202 sixth grade students from a Midwestern city were given this CBM at regular intervals. The demographic breakdown of the participants was: 65% White, 25% African American, 4% Asian, 1.5% Hispanic, .5% Native American, and .5% multiracial. 25% of the students in the sample received special education services. The authors used the year-long curriculum to gather 369 terms and create 35 probes out of these terms. Probes were administered by teachers from September to May once per week.

**Results** Overall the reliability of these measures was moderate (mean reliability coefficient .64) when taken as single measures and strong (mean reliability coefficient .89) when taken as combined adjacent measures (1+2 and 2+3, 2+3 and 3+4, etc...). In the discussion section the authors note that this implies the importance of teachers using averages or combined adjacent probes as the data for making instructional decisions: "When making instructional decisions, it is important to look at averaged scores across weeks to get a clearer and more consistent picture of student performance."

Validity for this measure was mixed, but the authors point out a flaw, which I consider a major gap in this study: the probes consisted of terms that covered the entire year, so students were most likely encountering terms used at different points in the year that they may not have learned.

**What I learned from this article** CBMs are useful and reliable measures for vocabulary knowledge. We as teachers need to use words that we are actually teaching in order to accurately measure growth.

This study serves as reinforcement for the use of CBMs and has a really nice literature review.

**Allen, A.A., Poch, A.L, Lembke, E.S. (2018). An Exploration of Alternative Scoring Methods Using Curriculum-Based Measurement**

**in Early Writing.** ***Learning Disability Quarterly, 41*****(2), 85-99. DOI: 10.1177/0731948717725490**

This article describes two studies that explored the technical adequacy of rubrics used for scoring CBMs in writing. The first study looked at a trait-based rubric in first grade, the second a trait-based rubric in third grade, with the addition of production dependent and independent scores.

**What is CBM in writing and what does it look like?** For those that are new to CBM in writing (CBM-W) I think it is important to describe what these are and how they are administered. This is foundational knowledge that you need to understand the rest of this article. If you know it, feel free to skip this description.

CBM-W is "an objective way to assess student writing and identify those in need of intervention." As with all CBMs the measures are designed to be quick to administer, easy to score, cheap, and standardized tasks that "represent indicators of overall proficiency in an academic area." CBM-W measures are usually a writing prompt, sometimes the beginning of the first sentence of a story (e.g. "Yesterday I went to the store and saw a _____"). Students are asked to complete the sentence and write a story in 3 minutes. Measures are scored in three ways: total words written, correctly spelled words, or correct writing sequences *see* This great guide by Jim Wright.(this is an auto-download pdf). According to this article the research on these tasks has shown that these story completion tasks are most appropriate at 3rd grade and above.

For younger children other options exist. A couple of studies have examined the reliability of copying and dictation measures at the sentence level. The authors of this article state that sentence dictation is more reliable than copying at the second grade level but sentence copying was reliable (though not as much) at the first grade level. Novel sentence writing tasks are reliable and valid for use in K-3.

As stated above, these measures can be scored in multiple ways: total words written, correctly spelled words, or correct writing sequences. This article describes these methods, along with correct minus incorrect words, "production-dependent measures." The authors also describe production-independent scoring methods. These focus on quality over quantity and include percent of words spelled correctly and percent of correct word sequences. The authors state: "These indices capture the differences between the transcription-focused instruction in the primary grades and the higher level skills required in writing at later grade levels." In two studies these measures were found to be more valid in the earlier grades, but less sensitive to growth over time.

The authors also describe qualitative scoring methods, primarily the use of rubrics. The authors describe two different rubric types, holistic, which looks at the quality and proficiency of the writing as a whole, or trait-based, which look at proficiency of certain writing traits. The authors cite Deborah Crusan's work *Dance, ten; looks, three: Why rubrics matter* to argue that holistic rubrics

are not very good measures for identifying areas of struggle and not good for instructional use. According to the authors the purpose of trait-based rubrics is too "increase objectivity, reliability, validity, and instructional utility of scores."

Something I found very interesting in this whole description of CBM-W measures and scoring is that qualitative and quantitative scores have very weak correlations.

**Method** This study included 40 first grade and 10 third grade students (these students were part of a larger sample from another study who took the criterion measure in that study.) Students were given two CBM-W tasks, picture word, in which students are asked to write a sentence for each picture, and story prompt, in which students construct a story based on a prompt. Both tasks were scored using total words written, words spelled correctly, correct word sequences, and correct minus incorrect word sequences. Students were also given the WIAT-3 Spelling and sentence composition sub-tests to measure validity.

Both of the CBM-W tasks in this study were administered three times, November/December, February, and April; the WIAT-3 was administered in May.

For the first grade part of this study the researchers looked at the picture words task and used a rubric that scored each sentence from 0-3 for sentence type, spelling, and grammar. Mechanics was also measured but was from 0-2.

For the third grade part of this study the researchers used the story prompt task. For this the researchers created a rubric, scored from 0-3 on the following traits: Sentence fluency/structure, spelling and word choice, mechanics and conventions, grammatical structure, relationship to and completeness of prompt, ideas, organization, voice, and development.

**Results** *First grade*: The rubric for this study suffered from weak internal consistency in reliability (the best alpha value was .64 in Spring, far below the .8 desired). Validity results were mixed, the CWS and CIWS measures were strongest. The rubric had a weak relationship with the WIAT-3 sentence composition subtest and a "moderate relationship at best with the. . . spelling subtest in winter and spring." Overall, according to the researchers, there is promise in this tool but it needs to be studied more and further refined. There is also promise as a progress monitoring tool, statistically signifcant growth was shown and no student scored the highest score possible at any point in the year.

*Third grade*: Internal consistency was much better for this rubric (.84 was the lowest alpha value). Overall this rubric showed promise, but growth was not significant and therefore this may not be a good measure for student growth. The researchers discuss that the small sample ($n$=10) probably played a role in the lack of growth and validity relationships. The researchers feel that this is a promising measure and encourage further research.

**What I learned from this study** Alternative measures are new and researchers are attempting to refine them to make them more valid. As they exist in the current moment, CBM-W scoring procedures are good, but could always be improved. I am excited to see where this goes and to see the use of

these rubrics as a way of measuring writing, especially when compared to other rubrics like the 6+1 traits. Having taught in a school that only used rubrics I am unconvinced that this could be completed quickly and is actually doable for every student every week or two, which is the purpose of the CBM, especially when looking at the rubric these researchers created (p. 93 of the article).

**Keller-Margulis, M.A., Payan, A., & Booth, C. (2012). Reading Curriculum-Based Measures in Spanish: An Examination of Validity and Diagnostic Accuracy. *Assessment for Effective Instruction 37*(2), 212-223. DOI: 10.1177/1534508711435721.**

In the abstract of this article the authors state that while there is an abundance of research on CBMs in reading for progress monitoring and screening, almost all of it is in English. This article is 6 years old and this statement is still true today. According to NCES there were 4.6 million English Language Learners in U.S. Schools in 2014-15, 3.7 million of which speak Spanish. Even with these numbers research in this field is still lacking.

The authors begin this article with a description of the validity, predictive validity, and diagnostic accuracy of CBMs in reading (R-CBM), describing the link between these scores and state reading tests in multiple states. These measures also have strong predictive validity and diagnostic scores, which all support their use in schools. Importantly for this study, most of this research is in English.

*For ELLs* "CBMs have been proposed for use to measure the transfer of language skills from one language to another, for academic skills progress monitoring, for adjusting language instruction to improve student language development and achievement outcomes, and to function as an alternative, less biased assessment strategy for bilingual students." As with their monolingual peers, the predictive validity and diagnostic accuracy of R-CBM are strong for ELLs.

*In Spanish* Less evidence exists to support the validity and diagnostic accuracy of R-CBM is Spanish. The authors describe the existing research as only initial evidence, and criticize the construction of the probes in the limited studies that exist. There is only one set of published probes for general use that have technical adequacy data published as well.

**Current Study** This study had two purposes: 1. To find the relationship between Spanish R-CBM and the Texas statewide achievement test and 2. To examine the diagnostic accuracy of Spanish R-CBM 25th percentile cut scores and receiver operator characteristic (ROC) curves for identifying whether students will be successful on statewide reading assessments in Spanish.

**Method** This study was completed in a large school district (more than 21,000 students) in the southeast. This is a diverse, 73% Hispanic, 20% African-American, 6% White, 1% Asian and economically disadvantaged district, 74% are economically disadvantaged. 29% of students in this district are ELLs. The district was already using R-CBM in English and Spanish as part of their

universal screening and benchmarking three times per year. I include this here because it is important, in my opinion, that this was not something new or unexpected for these students, so the effects are more reliable (my opinion, not stated in the article.) The third grade sample included 144 girls and 147 boys. The fourth grade sample included 92 girls and 83 boys.

The researchers used the Spanish AIMSWeb probes for their R-CBMs. These measures are translated versions of the English passages, which have many studies that support the validity and accuracy of the measures. There are no studies of this kind for the Spanish version (according to the article, since this article was published in 2012 there may be others.) Students were also given the Texas Achievement Test (TAKS) in Spanish.

"Students were administered the same three 1-minute reading passages during the fall, winter, and spring and the passages were scored for the WRC" (words read correctly). The researchers used median words read correctly for each time point.

**Results**
Correlations between the R-CBM in Spanish and the Spanish TAKS reading section ranged from .41-.48 for third grade and .37-.44 for fourth grade, moderate to strong correlations.

Diagnostic accuracy using the 25th percentile cut-score was adequate or above for specificity and negative predictive power (NPP). In this study NPP meant a score above the cut score, which implied success on the TAKS. The sensitivity and positive predictive power was low. The same result was found using ROC curves, though the cut scores determined by the curves were better than the arbitrary 25th percentile. This means that, regardless of which cut score procedure you use, the R-CBM in Spanish is adequate or better at predicting when a student will pass the TAKS, but less than adequate at predicting when they will fail. "...of the students who performed successfully on the Spanish TAKS, a large percentage were identified as such using the Spanish R-CBM as the predictor."

**What I learned from this article** Spanish R-CBM have good predictive reliability and diagnostic accuracy, but more research is required. The use of ROC curves made for more accurate tests with better sensitivity and specificity, but these are an advanced statistical procedure that many teachers probably will not be comfortable using. If there was a publisher that created an electronic CBM that used ROC curves that were automatically calculated I could see this idea taking off, but not if teachers had to figure it out on their own. I was relieved to see that using the 25th percentile was adequate, though obviously we want the best, we have to be willing to make a trade, either you take the best and all the learning and challenges that come with it, or you take an easier path and have a loss of quality.

I now feel more comfortable with R-CBM in Spanish, though I want to read more reports like this one.

**Kim, J.S., Vanderwood, M.L., & Lee, C.Y. (2016). Predictive Validity of Curriculum-Based Measures for English Learners at Varying English Proficiency Levels. *Educational Assessment 21*(1), 1-18. DOI**: http://dx.doi.org/10.1080.10627197.2015.1127750.

This study looked at the predictive validity of DIBELS Oral reading fluency and Daze tasks for ELLs. According to the authors there is a body of research showing predictive validity for R-CBM and less for Maze. The research that does exist on Maze has shown mixed results. ELLs are occasionally included in these studies, but they are included as a group and no attention is paid to the fact that they are a diverse group, both in language of origin and language proficiency. Without these details we cannot be sure our students are represented in the sample, and therefor could be going down a path which will turn out to be a wrong turn.

**Method** 522 Spanish-speaking ELL Third grade students from 23 classrooms in six schools in southern California were used in this study. Students were screened in fall using the DIBLES Oral Reading Fluency (DORF) task individually and the Daze task as a group. Teachers were provided with 7 hours of PD specific to DIBELS administration and scoring. Students then took the state ELA in the Spring. The fall CBM screening results were then compared with the Spring ELA results.

**Results** Sample-wide correlation between DORF and ELA was large ($r$=.54), for Daze it was moderate ($r$=.39). Correlations were then run for each proficiency group: Beginning/Early Intermediate DORF: $r$=.59 (large), Daze $r$=.35(moderate), Intermediate: DORF: $r$=.31(moderate), Daze: $r$=.3(moderate); Early advanced/Advanced: DORF: $r$=.36(moderate), Daze: $r$=.15 (small).

For predictive validity only the DORF task was a significant predictor of ELA performance for the entire sample. The authors state that this implies that Daze is not needed once DORF has been administered to third grade Spanish speaking ELLs (the narrow sample this study used).

There was no significant difference in the predictive validity for varying levels of English proficiency "suggesting that there is no difference in the predictive ability of DORF to CST-ELA based on English proficiency." The authors go on to state, "this is promising because it suggests that DORF is able to predict performance on the CST-ELA similarly for Spanish speaking ELLs of all English proficiency levels."

Sensitivity levels of cut-scores established by the DIBELS decreased with English proficiency level, as English levels increased sensitivity decreased, "indicating that the screening measures were able to predict truly at risk students better for Spanish-speaking students with lower English proficiency than for students with higher English proficiency." Overall "the ability of DORF and Daze to identify truly at risk students was not adequate."

**What I learned from this study** While the DORF task had a strong correlation to performance on the ELA test in the Spring, it did not meet the standards for predictive validity. This study used different methods than others in the past, they split the group of ELLs into smaller groups according to proficiency. This may have caused some different results, which they discuss in the discussion and limitations sections of the article. I am not sure how useful the DORF tool is for truly identifying ELLs at risk and would have to advise teachers to do more reading and research on this, not just take what DIBELS says to be true.

As I have believed for a long time, these tools may not be great at predicting at-risk students when used in this manner, but they are more useful when examining growth. This article reinforces one part of this, that they are not good at predicting risk when used in isolation.

**Gutierrez, G., & Vanderwood, L. (2013). A Growth Curve Analysis of Literacy Performance Among Second-Grade, Spanish-Speaking, English-Language Learners. *School Psychology Review 42(1)*,3-21.**

The authors of this article open by discussing the lack of ELL specific growth curves, or, that if they exist, they are meant for a homogeneous group of ELLs and do not acknowledge the variations of language proficiency within this large group. The authors set out to determine if there was a need for proficiency-specific growth curves and if so, to what extent does proficiency in English affect reading level and growth on measures of early literacy.

**Method** 260 second grade students from California participated in this study. Language proficiency was measured on the CELDT (California English Language Development Test). Students were given the DIBELS Oral reading fluency task (DORF), DIBELS Phoneme segmentation fluency (PSF) task, and the DIBELS Nonsense word fluency (NWF) task. All three tasks were administered in the fall, winter, and spring.

**Results** *ORF* English proficiency was significantly related to differences in the initial ORF measure. There were significant differences between every group, except the early advanced and advanced groups, in the number of words read aloud at the beginning of grade 2. Growth rates were also different depending on level. Beginners had a growth rate of .82 words/week, early intermediate had a rate of .95 words/week, intermediate .97, early advanced 1.1, and advanced 1.3. The authors relate the rates for early advanced and advanced to studies that show the growth rate for native English speakers to be similar to this group.

*PSF* Again, initial measures were significantly different, as were the growth curves. Beginner students actually plateaued between winter and spring. The authors state that this suggests that more Phonological awareness instruction should be given to beginners than others.

*NWF* More advanced students were able to read nonsense words more fluently at the beginning of second grade and had steeper growth curves, suggesting they came to the second grade with more word recognition skills.

**What I learned from this study** English proficiency has an affect on early literacy as measured by DIBELS. This was a small study, and it is only one study, so we cannot go making huge claims and changes from this one, but it is worth noting the differences to become more aware of these differences in our classrooms and students. The authors mention that phonemic awareness might be useful to look at for true growth for beginning ELLs in second grade. This study does suggest that growth rates differ, so this reinforces the idea that we have to look at true peers as a group to compare growth and use that group as our baseline. Students outside of that growth pattern and the kids we should be concerned with. This requires us to conduct more frequent screenings and monitoring for those children at risk. This study also shows us that ELLs at more advanced levels could be expected to grow at about the same rate as native English speaking peers.