Selecting a Universal Screening and Progress Monitoring Tool to Use in a MTSS Framework for On-Going Instructional Decision-Making and Special Education Eligibility Purposes:

A White Paper to Inform Decision Making

Erica Lembke, Ph.D in collaboration with
William Rasplica, MS Ed., Stephanie King, Ph.D., and Susan Ruby, Ph.D.

The Washington AIMS Project
Goodlad Institute for Educational Renewal
University of Washington Bothell

October 16, 2023

This content was produced under U.S. Department of Education, Office of Special Education Programs, Award No. H325L200006. The views expressed herein do not necessarily represent the positions or polices of the U.S. Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service, or enterprise mentioned in this paper is intended or should be inferred.

Purpose: This brief white paper further the AIMS goal of building the capacity of local district administrators to implement a MTSS framework that includes the tools and structures to support eligibility decisions. This information and guidance may also help to provide additional information for local special education administrators as they approach decisions regarding the 2028 target date for using RTI for eligibility decision making.

With a target audience of district and school administrators tasked with the work of selecting a tool, this document includes guidance for elementary through high school, with a particular focus on K-8. Research-based guidance in this white paper will help district leaders to make important, high stakes decisions regarding screening and progress monitoring tools.

- The document references the need to prepare not only for ongoing instructional decisions, but for the 2028 target date for a RTI-based eligibility process.
- Information regarding best practices in selecting screening tools, including sources for identifying vetted tools and a discussion of computer adaptive tools is provided relative to eligibility decision making.

Characteristics of an efficient and effective screening tool

Curriculum-Based Measurement (CBM). CBM is a system of screening and progress monitoring in academic areas that utilizes technically adequate measures to assess performance and progress.

CBM draws upon research support for automaticity and fluency, with a focus on development of measures that serve as indicators of broad constructs, such as mathematics proficiency. In the area of mathematics, Rhymer et al. (2000) cites literature that suggests that computational fluency, defined as responding accurately and rapidly, leads to better long-term outcomes such as longer-term maintenance of skills and better application to novel mathematics tasks. The National Mathematics Advisory Panel suggests that mathematical fluency includes both computational and procedural fluency. Clearly, there is a common theme throughout these reports and manuscripts indicating that rapid naming of facts and the ability to quickly apply procedures are critical to developing further mathematics skill. In their article on computational fluency for high school students, Calhoon et al. (2007) cite work demonstrating the far-reaching influences of fluency. For instance, The National Research Council (2001) provides an analogy suggesting that lack of computational fluency may have negative effects on mathematical comprehension similar to the effects that poor decoding has on reading comprehension (in Calhoon et al. 2007). In addition, Calhoon and her co-authors provide a nice overview of the literature suggesting that higher order mathematics cannot be accessed as efficiently if fluency is not present (Gerber & Semmel, 1994; Johnson & Layng, 1994; Pellegrino & Goldman, 1987 in Calhoon, 2007).

Overview of resources for where to go and what criteria to examine when initially selecting a screening tool. How do we interpret the information provided to make the best decision?

CBM measures embody specific characteristics, including: (a) efficient administration, (b) short duration, (c) technical adequacy, and (d) indicator of academic proficiency. The term *indicator* is used to signify the short duration of the measures as well as their strong relation to other measures of broad academic proficiency in that content area. Utilizing the research base that supports fluency, we can develop brief measures that serve as proxies for overall academic proficiency. Thus, although a common measure of CBM in reading is the number of words read correctly in one minute (oral reading fluency), this score serves as a broader indicator of academic proficiency in reading. In her 2004 article on the use of CBM measures, L.S. Fuchs described three stages of CBM research:

- Stage 1, technical features of the static score—reliability and validity of the screening/benchmarking/performance measures
- Stage 2, technical features of slope—reliability and validity of progress measures and the data collected from these measures on an ongoing basis
- Stage 3, instructional utility—how do teachers use the measures to make timely and important decisions regarding student instruction or intervention

These stages are important because measures need to be researched and then utilized only for the purpose they were intended and the purposes for which they have been validated. A measure that is appropriate for Stage 1 (assessing performance) may not have strong instructional utility. When considering what measure or combination of measures will be utilized for screening decisions for students, one must consider several technical features including: the accuracy of decision-making, predictive validity, and instructional utility of the measures across grades. In certain content areas like early mathematics (see Gersten et al. 2012), a battery of measures might be considered rather than a single measure.

One of the best practical resources to access when examining data for measures and making decisions about which measure to use is data from the National Center on Intensive Intervention tools charts (intensiveintervention.org). Both Progress Monitoring and Screening Tools Charts are assembled by the National Center on Intensive Intervention (NCII) to provide information regarding technical adequacy. These charts are updated annually with new screening measures as well as with evolving information for existing tools. A users guide for teams to use as they access the tools charts can be found here: https://intensiveintervention.org/sites/default/files/Tools_Chart_User_Guide-508.pdf.

While the NCII has experts who review the measures and provide detailed ratings, it is up to school teams to use the information provided to make the best decisions for their building and students. The NCII review process is just that—a review. Value judgements about the potential utility of a given tool are up to the user.

Other features that may impact screening decisions. Once an appropriate measure is selected that maps on to our desired educational decision, other factors must be considered. The importance of classification accuracy is a critical component of any screener (Gersten, 2012; Johnson, Jenkins, & Petscher, 2010; Kovaleski, et al., 2022). Classification accuracy refers to how accurately a measure can be utilized to predict a decision regarding future student performance. For instance, classification accuracy might be calculated to determine how likely a student would be to pass or fail a high-stakes assessment in the spring based on initial performance on a CBM during fall screening. Sensitivity, specificity, and the area under the

Receiver Operating Characteristic (ROC) Curve are some of the statistics used to estimate the accuracy of a given screening measure and is only interpretable for given associated set of cut points, in terms of correctly identify students at one point in time as at at-risk or on track for outcomes measured at a later time. Sensitivity (i.e., the true positive fraction) describes how acutely a particular cut point on a screening measure identifies children as at-risk who end up failing the outcome measure; sensitivity is not interpretable without also knowing the corresponding value of specificity for that same cut point. *Specificity* (i.e., 1 – false positive fraction), on the other hand, refers to the degree to which a given cut point on specific screening measure rules out students who are not at risk for failing the outcome measure; a screener that is specific reduces the number of students who are identified erroneously as needing additional instructional support. There are tradeoffs between sensitivity and specificity (i.e., as one increases the other decreases) and, depending on how accurate the screener is overall, the differences between sensitivity and specificity can be quite large. The second edition of a book published in 2022 by Kovaleski and colleagues, *The RTI Approach to Evaluating Learning Disabilities,* provides a comprehensive overview of how classification accuracy can be improved and utilized for high-stakes decision making and details how schools could use CBM screening data in a multi-tiered system to make classification decisions regarding students who might be in need of additional intervention. The goal for schools would be to maximize the number of students correctly identified. Utilizing a complementary process of screening and then a few weeks of follow-up progress monitoring to confirm or disconfirm the screening decision can be effective in enhancing classification accuracy.

This recent work in classification accuracy highlights the movement towards greater precision in decision making utilizing CBM. Initial development of CBM focused on decisions that an individual teacher might make about a small group of students. A teacher would examine recent data values that had been collected and would apply a decision rule like the "three below, six above rule" (see Deno & Mirkin, 1982) where if three weekly data points were below the goal line (see Example 1), an intervention change would be needed but if six weekly data points were *above* the goal line, it would be time to raise the goal for the student. As the uses for CBM morphed from individual teacher decision making to decision making for larger groups of students (utilizing normative data) the need for greater accuracy emerged. CBM essentially transitioned from serving as a key measure in an individualized, instructionally driven model for special education teachers, to being utilized across general education for universal screening, to serving as a key component in special education *eligibility* decision making as part of an RTI model.

Thus, screening serves as an important technique to identify students at risk early, while there is still time to intervene. For CBM screening, the higher the stakes of the decision, the more important precision in decision-making becomes. For instance, making a decision about student placement in special education is extremely high stakes and CBM screening data is one piece of data to aid that process. Precise decision making is necessary when utilizing CBM data for this purpose, where a students' placement will be substantially influenced. A lower stakes decision that still requires specificity, but not to the same degree as special education evaluation, might be determination of small-group intervention activities for a low-performing classroom based on

CBM screening data. The good news is that educators can find greater detail and more specificity on these issues in resources such as the book by Kovaleski and colleagues (2022).

Potential issues when utilizing a computer adaptive tool (CAT) (or pros/cons of computer adaptive vs. General Outcome Measures)
A brief literature search resulted in very few peer reviewed publications that specifically compared computer adaptive measures versus general outcome measures. So, in the absence of information about these assessments utilized as tools for identification, the following are some considerations.

Articles that might be useful in helping think through the different test formats and pros/cons can be found at:

- https://ed.lehigh.edu/sites/ed.lehigh.edu/files/Shapiro%20and%20Gibbs%202014.pdf
- https://www.researchgate.net/publication/275040844_The_Predictive_Validity_of_a_Computer-Adaptive_Assessment_of_Kindergarten_and_First-Grade_Reading_Skills/download
- https://scholarworks.umass.edu/pare/vol18/iss1/14/.

In the literature, one of the biggest cons with some CATs is content sampling and precision at the tails. First, the CAT needs to be designed to pull a representative sample of the content domain. Second, if the item bank isn't wide enough, the measurement error can be very high at the extremes (and this is particularly problematic when trying to make accurate recommendations for students experiencing difficulty). So, this is yet another reason why it is important for administrators to seek out information on technical adequacy, such as the information that can be found in the National Center for Intensive Intervention (NCII) tools charts.

In addition, Nathan Clemens (University of Texas at Austin) a prominent reading researcher who has studied CATs in the past suggests "In my opinion, the value of CATs as screeners improves as students get older, when there is more confidence that students are interacting with the questions/answers independently. With younger kids, I have more confidence in paper-based measures of print-related skills. Overall, I think that CATs are decent screeners for risk. I wouldn't recommend them for IDing LD, but if a goal is identifying risk/possible LD then the full score gives you some sense of that and then other assessments can be done from there. Where CATs can be problematic is when they provide scores and interpretations of subskills/subscales. Given their adaptive nature and vertical scale, students may see only a couple items (or in some cases no items) in a subskill area but the program estimates their skills in that area regardless. I'm not aware of data on accuracy for identifying LD specifically, most of the data on classification accuracy is on passing/failing state tests."

In a 2023 workshop at the National Association of School Psychologists conference, Dr. Robin Codding (Northeastern University) shared her recent work in examining the use of assessments in mathematics. She states, "I…. have noted that many of the CAT are now also including

CBMs….and that still the promise of CAT has not been realized for instructional planning or for progress monitoring. The difficulty is that these data have not yet been analyzed and I suspect that there are few studies on CATs for anything other than screening. Our data showed that CBM is fine for screening. My argument was that the reason to choose CBM over CAT is to actually engage in data-based decision making and progress monitoring."

While a district administrator would not be expected to know about how the test developer examined 'content sampling and precision at the tails', there is enough doubt surrounding Computer Adaptive Tests and their use for disability decision making, that school districts should *proceed with caution* as they consider a CAT for this purpose.

Discussion of importance of logistical issues like cost of the tool and amount of time for administration.

Practically speaking, it is important for school teams to consider their resources available and the time and money they would like to dedicate to screening. The National Center on Intensive Intervention has summarized information on cost and time for administration on their tools charts (https://charts.intensiveintervention.org/ascreening). Information can be accessed by clicking on the name of the measure. Teams should consider selecting a measure or system that meets their needs for the subject matter, grade level, and use they are considering. For instance, if a school wants to implement screening in mathematics and already has a system in reading, utilizing the same system might make sense. If a school wants to try screening in an academic area that they have not attempted before, they might consider a free or low cost system to start.

The time for administration is a key consideration. Screening measures serve as indicators of academic proficiency and should be able to be administered in a relatively short amount of time. Each measure should be between 1 and 10 minutes ideally.

Brief discussion of the difference between screening, progress monitoring, and diagnostic tools

The purpose of screening is early identification, identifying students at risk to appropriately place them in intervention, and predicting future performance. Characteristics of high quality screening tools include ease of use (administration, scoring, cost, training), high accuracy in predicting success on the outcome of interest, can be easily linked to instruction, and provide precision in distinguishing students who might develop difficulty in the target area (like reading; Petscher et al. (2011)).

The purpose of high quality progress monitoring tools include providing information that informs teacher's instruction, aiding in determining student response to instruction and/or intervention, aiding in teacher decision making about movement to receive more/less intense support, and in some cases, supporting the decision (as part of data triangulation) to refer to special education. Characteristics of progress monitoring tools include that the tools are reliable and valid, can detect growth in brief periods of time (sensitive), can model growth, accurately inform teachers regarding whether a student is benefitting from instruction, and have low measurement error or measurement error is accounted for.

The purpose of diagnostic skills is to provide information on what skills the student has mastered or what skills are continued needs. Characteristics of high-quality diagnostic skills include that the assessment provides information on specific and important content of interest. The diagnostic tool should be aligned with content that the student is currently learning or content the student will need to know by the end of the unit of study.

Explanation and understanding of data literacy. Data literacy includes both administering measures, but also understanding the data and for what purpose the measures are administered. Typically, we screen all students in a building or grade level to identify students who may not be at the level we would expect in a given academic area at a given time of year. How do we use the data? Following are steps in a data-based model for decision making using CBM.

Step 1--Screening using CBM measures. All students should be screened using CBM measures, ideally, three times per year (fall, winter, spring). Universal screening means that all students in a building are tested. Typical measured used for screening are short-duration tasks that are matched to students' grade levels; the results of those tests are then compared to established normative levels of performance. These norms are developed as a result of national, state, or local data collection, and translate into benchmark levels of performance that are standard criteria where students need to be performing to be deemed ‘not at risk’ at a particular time of year. The criteria that determine risk status are determined statistically after examining data that has been collected for each grade at each time of year. Students who fall below a predetermined benchmark on the CBM are identified as needing additional instruction or interventions and their progress will be monitored more frequently.

Step 2—Setting a goal for the student and label the goal line on the student’s graph. Goals can be determined in one of three ways: 1) according to national norms, which vary by CBM product, 2) grade level-benchmarks, which also vary by CBM product or 3) an intra-individual framework, where a student’s individual data are used to project a reasonable goal in the time allotted using expected rates of growth. For example, using the intra-individual framework, a teacher could specify that a student would gain two words per week on a test of oral reading fluency. The end goal would be determined by the following formula:

$Goal=(2\ words\ per\ week\ expected\ gain)x\ 27\ weeks\ left\ in\ the\ school\ year+CBM\ screening\ score$

If a student had an initial score of 20, the goal score in 27 weeks would be 74.

However the goal is decided, it along with the student’s current level of fluency (baseline; present level of performance) is marked on an individual student graph. The baseline and goal points are connected and a line is drawn between them. This goal line spans the number of instructional weeks between the baseline level of performance and the point by which the goal is desired to be achieved. This goal line determines the most direct route to take when attempting to reach the desired level of performance.

Step 3--Identification of strengths and weaknesses using diagnostic measures. In addition to CBM screening measures, students who have been identified as requiring more intensive interventions may be given diagnostic measures. CBMs tell us *if* there is a problem. Diagnostic assessments tell us specifically *what* skills are at a deficit and what the student is able to do well. Diagnostic information is then used to develop an intervention plan and determine where to focus instruction. An example of a diagnostic fluency assessment is a miscue analysis (Fuchs, Fuchs, Hosp, & Jenkins, 2001), which determines the specific types of errors a student is making in reading. Teachers make notation about student errors as the student is reading aloud and then later go back and categorize the types of errors the student made.

Step 4--Generating hypothesis about appropriate method to individualize instruction for the student. Using CBM results and diagnostic data as appropriate, educators should come up with logical ideas about what type of intervention program, instructional content, and delivery setting would be appropriate for each student. It is important to consider not only the specific skills the student needs to work on but also the amount and frequency of supplemental instruction and the size and composition of the intervention group.

Step 5--Creating an instructional plan for each student or group of students. Educators will develop an instructional plan with a goal and instructional activities for each student. These activities should be research or evidence based and typically include direct, explicit, and systematic intervention for the deficit area(s) identified during the diagnostic step above.

Step 6--Beginning daily instruction using the instructional plan. The instruction or intervention will be provided for as much time as possible, relevant to the skill needs. The greater the academic needs of the students, the more often the intervention should be implemented and for a greater length of time each session.

Step 7--Weekly progress monitoring, including scoring and graphing, using a CBM measure. To continuously monitor student response to the intervention, regular weekly progress monitoring data using a CBM is necessary. Continue to graph these data on the student's graph to determine if the student's performance is changing and how close his or her data points are to the goal line (see Step 2 and chapter 5 for more details about progress monitoring decision rules and evaluating a student's response to instruction). Weekly progress monitoring is recommended for students who are significantly behind their peers (e.g., someone in a Tier 3-level intervention), whereas monitoring every other week or monthly may be more appropriate for students who are not as far behind (e.g., someone in a Tier 2-level intervention).

Step 8—Making ongoing changes in instruction based on decision-making rules.
Using progress monitoring data, educators can determine if the instructional plan is having the intended effect. The main method for making educational decisions using student progress monitoring data is the trendline rule. There are several methods for determining the trend of student's progress. The National Center for Response to Intervention (NCII, 2014) lists the methods for several of the most common and supported in its glossary of terms available at the following web address: http://www.intensiveintervention.org/ncii-glossary-terms. When the trendline is at or above the aimline, the intervention will be continued and likely faded if progress remains strong. When the student progress (represented by the trendline) is *below* the

aimline, consider making a change to the intervention delivery or, in some cases, content. These changes might include some intensification of intervention, which is described in much more detail and with many resources at Designing Appropriate Academic Interventions Using the Taxonomy of Intervention Intensity | NCII (intensiveintervention.org).

How does this deeper understanding impact important resource allocation issues? What are some common questions about prioritization and allocation of resources to make the best decisions about tools to select? How do administrators provide a rationale for making a change in the type of tool utilized and the cost of a high-quality screening tool?

As administrators learn more about measures and are able to make more informed choices, greater attention can be given to how to allocate resources for screening tools. Some common questions that administrators should consider include the following:

1. How will the screening data be used?
2. Is the screener reliable and valid for the intended use?
3. What is the cost of the screening tool compared to the utility? For instance, if this is the first time implementing math screening, a school may want to pilot a free or low cost screener with a small group of students to get a sense of implementation concerns.
4. How long does it take to complete?
5. How is the data gathered, where is it stored, and how easy is it to access?
6. Who needs to be trained to administer the screening tool and how difficult is it to train administrators?
7. How are the assessment results presented?
8. What level of ongoing support is required?
9. Can the screening tool be used across grades and subject areas?
10. Does the tool correctly classify students into the correct ranges according to their skill levels and are students from a variety of backgrounds and programs accurately represented (for instance, students in special education and English Learners)?

Administrators can use their in-depth knowledge of screening tools and their answers to the questions above to provide a greater rationale for why a change may need to be made in a screening tool and why a more expensive tool might need to be utilized. For instance, if a district has been using a free reading screening tool, but now wants to screen in reading, mathematics, and writing, it will likely make sense to purchase a comprehensive screener that can be utilized for all purposes.

We recognize that making changes in screening and progress monitoring systems at a district level are hard decisions for administrators, so we want these superintendent and cabinet level teams to have the most information possible. Teams might want to consider an 'assessment audit' where they look at each assessment they are requiring at each grade and ask the following questions:

- □ What is the assessment we're using?
- □ In what content area?
- □ How often?

- At what grades?
- What type of information is the assessment providing us? (screening, diagnostic, progress monitoring)
- What is the technical adequacy of the assessment?
- How are we using the data? (i.e., national or state comparison, diagnostically)
- Is this the way the data was intended to be used? (was the test researched to be used this way?)
- Can we reduce the number of times this assessment is administered or eliminate it?

What should screening look like in grades 9-12 and how does it differ from K-8? In the elementary grades, students are being exposed to a great deal of content and academic skills are increasing at a rapid rate. As students get older, their assimilation of new academic skills slows. So while in the elementary grades, it is important to continue to screen basic academic skills in reading, math, and writing 3 times a year, as students move onto high school, we have more existing data to support screening decisions. Because we have lots of historical data from classroom assessments, state assessments, district tests, and other standardized tests, we can use that data for initial screening decisions. For instance, rather than using CBM screening for all students in high school, school teams could use state test data as a first 'gate' to determine students who have scored in the lowest percentiles or who did not reach proficient levels according to state norms. For the lower number of students who are deemed at risk from this initial gated screening, a next step is to provide a more proximal screening measure or group of measures to learn more about present academic needs (second gate). In this way the number of students who need to be screened using CBM is reduced, and more data sources are utilized for decision making from students' historical performance.

Conclusion

To summarize, administrators have many questions to ask and hard decisions to make as they consider the best assessment system to utilize in their district. Perhaps Clemens et al. (2023), in an upcoming book revision, best captures the difficulties in navigating CAT assessments for screening and progress monitoring: "In summary, CATs are an interesting advancement in academic assessment. However, educators must exercise considerable caution in how they are used. CATs are well suited for occasions in which educators want a global estimate of student performance within or across academic domains at a single point in time, such as summative assessment situations. Universal screening is another situation in which CATs are better suited, and there is evidence that CATs can serve as good universal screening tools beginning in middle elementary grades (January & Ardoin, 2015; Klingbeil, Nelson, et al., 2017). However, CATs are not well suited for progress monitoring roles, as revealed by research on their technical properties, characteristics that would be necessary for frequent and repeated administration, and data that can be interpreted for making instructional decisions. This may change with subsequent development and testing with CATs."

References

Calhoon, M. B, Emerson, R. W., Flores, M., & Houchins, D. E. (2007). Computational fluency performance profile of high school students with mathematics disabilities. *Remedial and Special Education, 28*(5), 292–303.

Clemens, N. H., Hsiao, Y.-Y., Simmons, L. E., Kwok, O., Greene, E. A., Soohoo, M. M., Henri, M. A., Luo, W., Prickett, C., Rivas, B., & Otaiba, S. A. (2019). Predictive Validity of Kindergarten Progress Monitoring Measures Across the School Year: Application of Dominance Analysis. Assessment for Effective Intervention, 44(4), 241–255. https://doi.org/10.1177/1534508418775805

Clemens, N. H., Hagan-Burke, S., Luo, W., Cerda, C., Blakely, A., Frosch, J., ... & Jones, M. (2015). The predictive validity of a computer-adaptive assessment of kindergarten and first-grade reading skills. *School Psychology Review*, *44*(1), 76-97.

Kovaleski, J. F., VanDerHeyden, A. M., Runge, T. J., Zirkel, P. A., & Shapiro, E. S. (2022). *The RTI approach to evaluating learning disabilities*. Guilford Publications.

Lembke, E.S., Carlisle, A., & Poch, A. (2016). Using Curriculum-Based Measurement fluency data for initial screening decisions. In K.D. Cummings and Y. Petscher (Eds.). *Fluency Metrics in Education: Implications for Test Developers, Researchers, and Practitioners*. New York: Springer.

Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., ... & Hwang, J. (2023). Review of curriculum-based measurement in mathematics: An update and extension of the literature. *Journal of School Psychology*, *97*, 1-42.

Rhymer, K. N., Dittmer, K. I., Skinner, C. H., & Jackson, B. (2000). Effectiveness of a multi-component treatment for improving mathematics fluency. *School Psychology Quarterly, 15*(1), 40.