

Improving Efficiency and Accuracy of Screening Procedures and Decisions in Early Reading

Amanda M. VanDerHeyden, Ph.D.

Education Research & Consulting, Inc.

Prepared in December, 2010

Submitted to State Department of Education in a southeastern state that mandates
“overassessment” of students

What follows is the content of two memos submitted to the state department in a southern state offering specific suggestions for more efficient and accurate screening procedures in reading. These memos were drafted by Amanda VanDerHeyden.

Thank you for the opportunity to make some targeted recommendations for early reading assessment that could improve efficiency for schools.

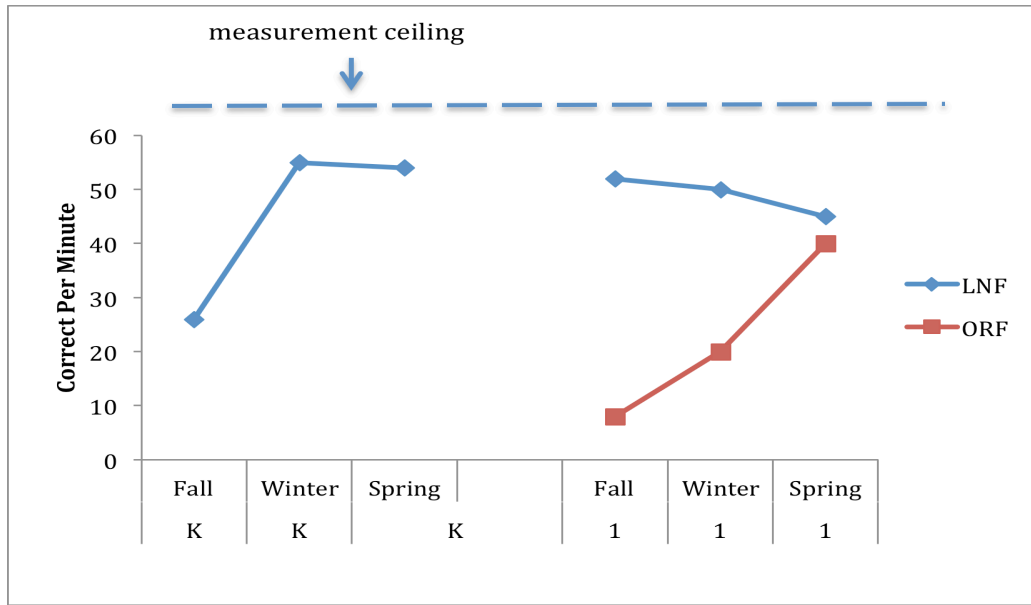
Purpose of Reading Screening (DIBELS):

1. Identify students, classes, grades, schools, and districts who are at risk for reading failure so that intervention can be provided.
2. Evaluate the effects of prevention efforts (identification and intervention) at the student, class, grade, school, and district level. Effective prevention would be reflected by shrinking numbers of students appearing in the risk range on consecutive screenings (Shapiro & Clemens, 2009).

Meaning of the Assessment Data:

The meaningfulness of the data collected depend on many factors (e.g., correct administration and scoring of the measures, quality of assessment passages), however, one is particularly pertinent to our discussion. As students are beginning to learn to read, the early DIBELS measures like Initial Sounds Fluency, Phoneme Segmentation Fluency, and Letter Naming Fluency are powerful indicators of a child's general competence in phonological awareness and fluency. The fluency with which a child can identify beginning sounds in words, separate words into phonemes, and blend phonemes into words forecasts whether a child will learn to read well by third grade or not (Good, Simmons, & Kame'enui, 2001). As children learn to read, however, the value of the early DIBELS measures begins to decrease for two reasons. The first reason the early subtests become less meaningful as children learn to read is that children who are learning to read will reach a measurement ceiling on the early DIBELS assessments. That is, the early measures will not be capable of sensitively detecting further growth once a student has reached a certain level of proficiency. Second, as children learn to read, the earlier DIBELS assessments represent a mismatch between what they are being taught (i.e., to read connected text) and what skills are being assessed. Therefore, the skills become less meaningful as predictors of reading success.

Once children begin to read connected text, oral reading fluency becomes a superior measure to make a risk decision (i.e., for screening) and incidentally also for progress monitoring. In the figure below, you can see that LNF scores reach a measurement ceiling where no further growth can be detected. Oral reading fluency provides a more sensitive and accurate indicator of reading proficiency once students begin to read connected text during instruction.



Because of the reasons above, policy groups and researchers generally recommend the use of oral reading fluency for screening and progress monitoring once students begin to read connected text (Fuchs, 2003; Florida Center for Reading Research). To evaluate the success (or non-success) of programs at the district or state level, several options can be used that are more efficient and just as accurate (See the last section of this statement).

Cost of Assessment:

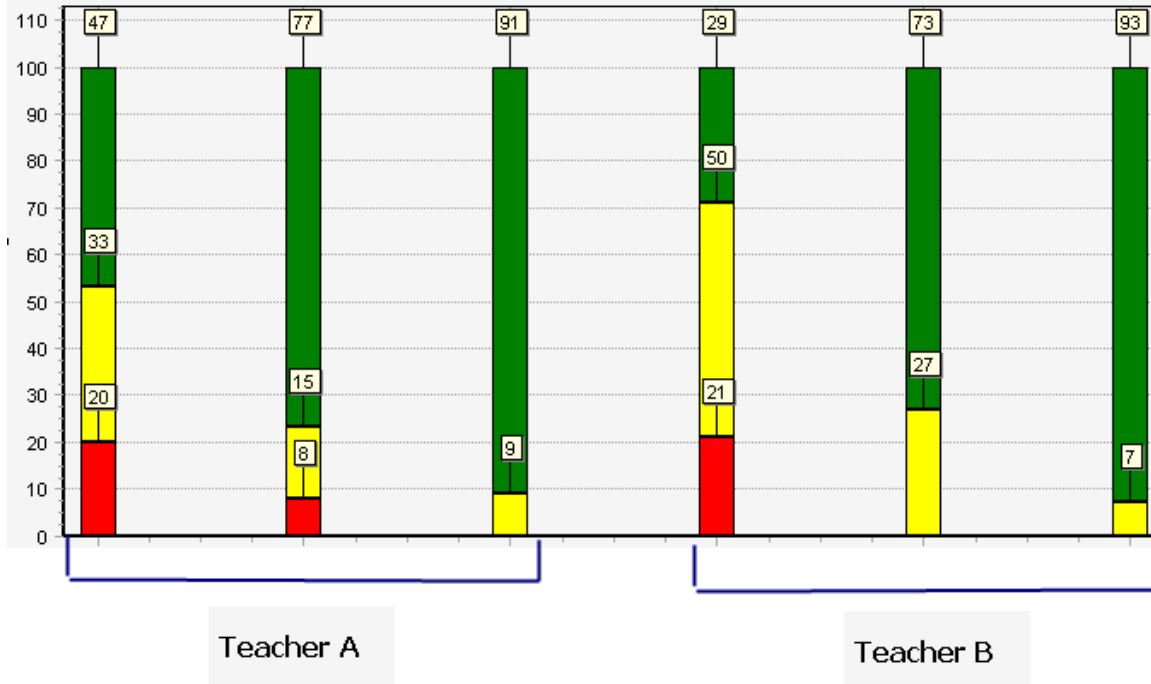
In a class of 25 students, oral reading fluency estimates can be obtained to reach a reliable and accurate screening decision with a single one-minute timed reading (Ardoin et al., 2004). Hence, oral reading fluency scores can be obtained classwide in about 30 minutes. Administering the full set of DIBELS early reading assessments (e.g., ISF, LNF, PSF, and NWF) requires about 10 minutes per student and equates to about 250 minutes of assessment time for the same class. Reducing assessment time is critical to building sustainable data-based decision models and multi-tiered intervention systems. Assessment comes at a direct cost to other activities, namely intervention or instruction. Hence, where the same decision can be reached with the same or better accuracy, implementers should choose the more efficient assessment option. In the case of reading assessment, there is general consensus that early DIBELS measures (or other measures of phonological awareness and phonics) can be used at Kindergarten, oral reading fluency in grades 1-3, and Maze assessment grades 4-9 for maximal accuracy and efficiency of assessment.

To Monitor Program Effectiveness:

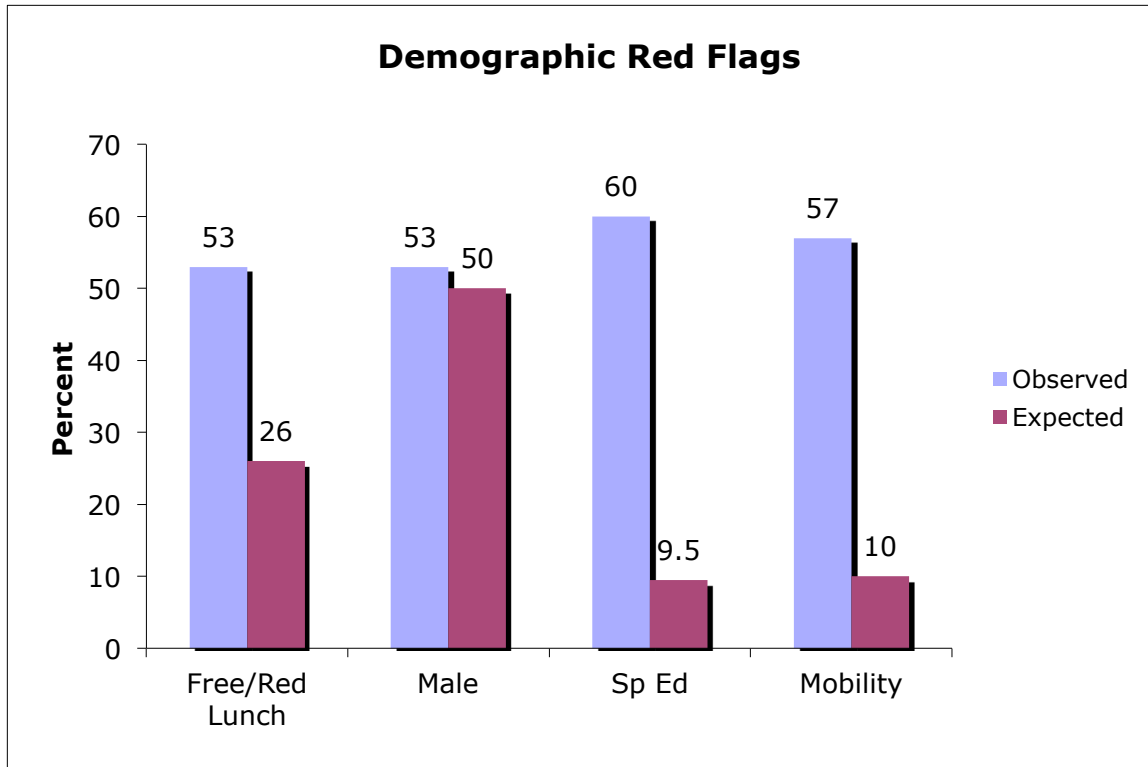
There is an active line of high-quality research specifying challenges associated with monitoring growth in reading performance. Based on this line of research, some targeted conclusions are possible.

1. Differences in the actual materials used from one assessment occasion to the next (e.g., actual reading passage), can affect conclusions about growth and cause error. This finding has led some researchers to recommend the use of the same passage or passage set from one monitoring occasion to the next (Ardoin & Christ, 2008; Griffiths et al., 2009).

2. Measuring slope is problematic (Schatschneider, Wagner, & Crawford, 2008). Our capacity to do this is often overstated and misunderstood. For example, there is often (typically) more error associated with estimating slope than there is expected change in slope due to learning. Growth that occurs during intervention or instruction can be detected using visual analysis and used formatively to know whether an intervention is producing change in the desired direction or not and whether troubleshooting of the intervention should occur. Slope becomes problematic when we try to quantify it, specify expected or typical growth, and use the slope value as part of a decision criterion to evaluate intervention success. The most prudent recommendation is to use slope data formatively to guide intervention implementation, but to use post-intervention benchmarks to evaluate response to intervention and the same is true when evaluating instruction at the class, school, district, or state level (VanDerHeyden & Burns, 2010).
3. In the table on the next page, I have suggested some outcome metrics that could be used to monitor intervention program effects. Improvements over time could be reflected by the number and demographic proportionality of students found to be at risk during screening. Further, prevention effects could be measured by continued risk status of individual students on subsequent screenings and number of students at risk over time and across programs of instruction (Shapiro & Clemens, 2009). If slope or rate of growth is to be used at all, the most defensible approach would be to use the same assessment materials (exact same form) and to constrain growth estimates to a single semester (Ardoin & Christ, 2008). This information, however, does not provide “better” information than would a metric like continued risk status (which is empirically supported). To provide a concrete example (and many other graphs are possible), consider the figure below. In the figure below, the number of students at-risk (scoring below screening criterion) at the fall, winter, and spring screening is shown for two teachers (could be two schools, two districts). In the fall, Teacher A has 20% of her class scoring in the risk range. By spring, no students score in the risk range, 9% score in the instructional range, and 91% are scoring at mastery.



In the next figure, consider year-end test scores and demographic analyses. The proportion of expected non-proficient students can be computed based on the demographic characteristics at the school. In this example, 26% of students receive free or reduced lunch. Hence, we expect that 26% of students who fail the ARMT in the spring will receive free or reduced lunch unless poverty is causing undue risk. The fact that 53% of those students scoring below proficiency in the spring also receive free or reduced lunch tells us that we can do more to prevent failure for this group of students and ought to be a meaningful intervention target. Gender, special education status, and mobility can be examined the same way.



Summary Recommendations:

Thank you again for the opportunity to provide some targeted recommendations. In summary, I respectfully request your consideration of the following actions:

1. Permit Fairhope Elementary School to administer only oral reading fluency probes schoolwide (grades 1-3) at fall, winter, and spring screenings given that oral reading fluency is superior to the early DIBELS measures both in terms of accuracy of the risk judgment and also in terms of efficiency.
2. Permit Kindergarten students who reach a mastery criterion on the early DIBELS measures to exit out of early measures in favor of more challenging measures which will have greater sensitivity for students (and greater value for teachers instructing those students). Specifically, once students reach the low-risk criterion specified by DIBELS and supported through research (Hintze, Ryan, & Stoner, 2003), reduce the assessment requirements for those students as follows. Begin with ISF and LNF in the fall. Continue these throughout kindergarten for all students who do not reach the low-risk criterion. At winter screening, discontinue ISF and LNF for students who met the low-risk criterion in the fall and add PSF and NWF. At spring, discontinue ISF and LNF for all students who met the low-risk criterion at winter. Discontinue PSF and NWF for all students who met the low-risk criterion at winter. Consider a single one-minute oral reading fluency measure using simple sentences at the end of the year for students who are reading (Fuchs, 2003). Bearing in mind that the DIBELS risk criteria are conservative and actually generate a number of

false-positive decision errors, this approach will greatly enhance efficiency of assessment and permit more time for instruction.

3. Use number of students at risk and continued risk status overall, by intervention, and by demographics to evaluate prevention effects at the state, district, school, and class level.

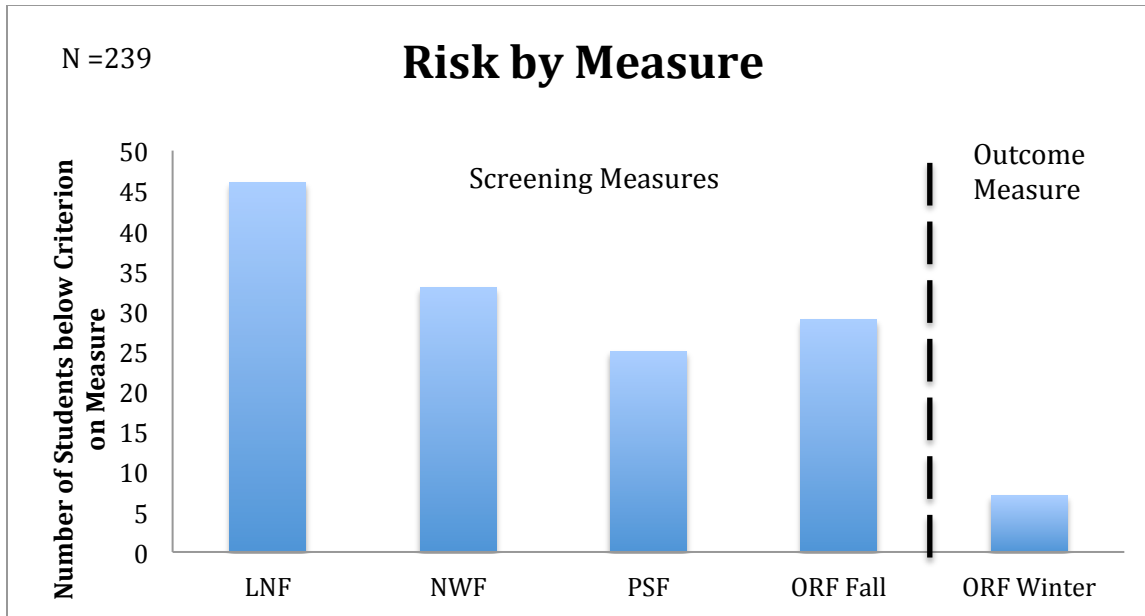
Data in Support of Recommendations:

I would like to submit some data in support of the above recommendations. These data were collected as part of routine schoolwide screening conducted as part of RtI implementation in a school district in south Alabama.

In the figures below, you can see Letter Naming Fluency, Phoneme Segmentation Fluency, and Nonsense Word Fluency, which I have argued previously are not useful for decision making at first grade in reaching screening decisions. Specifically, these are not useful screeners at first grade because these tasks are too easy for entering first grade students (these skills are not well-aligned with instructional expectations and procedures in the classroom by first grade) and therefore may fail to detect students who are at risk for reading failure (i.e., it is feasible that students would perform well on these subtests but still be at risk). Second, these scores are not useful for monitoring further progress because a measurement ceiling has likely been reached. Third, these skills are not well-aligned with instructional expectations and procedures in the classroom by first grade.

For decision making, we can consider classification agreement data to evaluate the utility of each of these measures for predicting reading failure at winter of first grade. Scores on these measures are moderately to strongly correlated with each other and with winter of first grade oral reading fluency. LNF and NWF scores correlated with ORF winter scores at about $r = .7$. ORF fall scores correlated with ORF winter scores at $r = .8$. PSF was the only exception with relatively weak correlation values with ORF fall scores ($r = .3$) and ORF winter scores ($r = .4$).

Rates of risk were similar across screening measures, especially for NWF and ORF Fall.

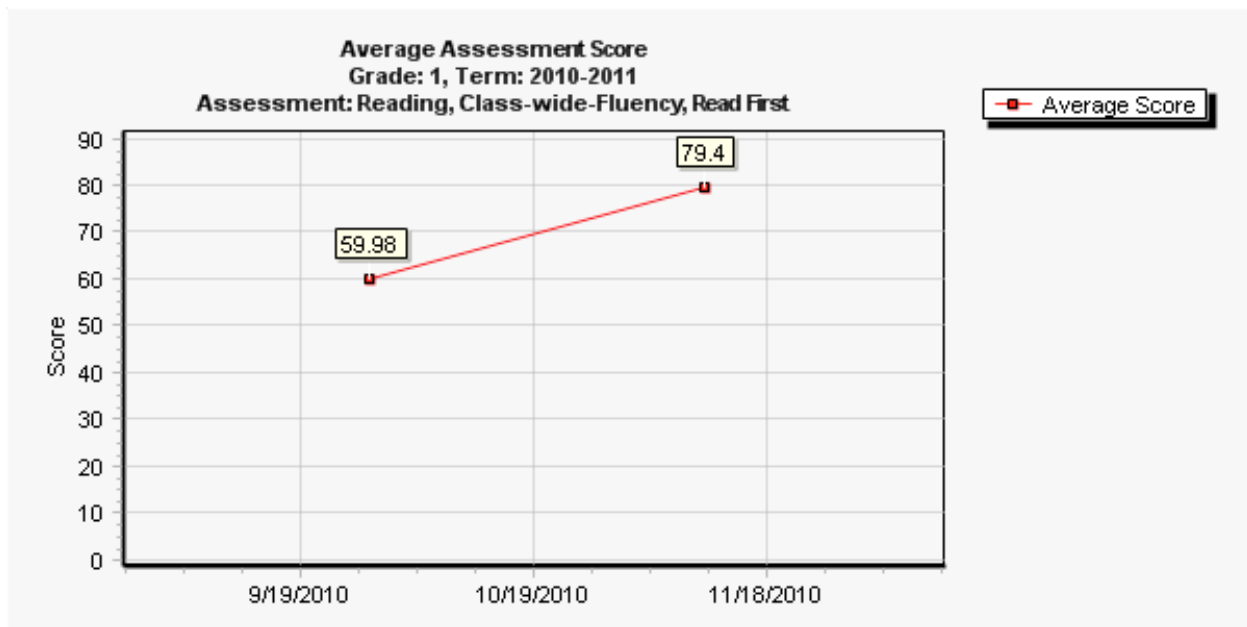


Sensitivity is the power of the test to detect true positives. Specificity is the power of the test to detect true negatives. When cutpoints for risk decisions are set on continuous test scores (like NWF, ORF, etc.), there is always a trade-off between sensitivity and false positive error rates. That is, as the score for judging whether a student is at risk or not at risk is raised to become more stringent, fewer true negatives will be detected (so we will fail to detect students) but we will also have fewer false positive errors. As we lower the cutscore to become more liberal, we will make fewer false negative errors (missing students who should have been detected) but it comes at a cost to efficiency (increasing false positive errors). So the goal in selecting a test measure is to select the one with the greatest sensitivity that comes at the lowest cost in terms of false positive errors. DIBELS subtests (and many others) have cutscores that carry very high false positive error rates (Hintze, Stoner, & Ryan, 2003). So any district using DIBELS must pay close attention to their own decision accuracy analyses to ensure their value for decision making. In the case of the data presented here, I am making an argument that the LNF, PSF, and even the NWF are not useful for screening decisions or progress monitoring at the first grade level. The context in which I am making this argument is one in which students have learned to read connected text by the end of the first month of first grade. In settings where students are not yet reading, the conclusion might be different. Using winter of first grade oral reading fluency as the criterion for comparison (greater than 19 words read correctly per minute at winter of first grade is judged not at risk), the following calculations can be made.

	LNF	PSF	NWF	ORF Fall
Sensitivity	86%	57%	86%	100%
Specificity	83%	91%	90%	91%
False Positive Rate	17%	9%	10%	9%

Based on these data, the measure with the greatest sensitivity and the lowest false positive rate is oral reading fluency scores at entry of first grade.

Finally, for progress monitoring, oral reading fluency at first grade can readily be modeled in a linear fashion throughout the year to evaluate growth in response to instruction (see figure below showing average scores from fall to early winter). The risk criterion for oral reading fluency can readily be adjusted to reflect the typical learning rate so that children at risk for reading failure can be detected earlier and with greater accuracy (e.g., adjusting the criterion to reflect 1.5 words correct per minute growth per week of instruction). This rate of growth maps onto a benchmark criterion score of 20 wc/min by winter and 40 wc/min by end of first grade. For students who are reading connected text, this method of screening and progress monitoring is psychometrically superior to other DIBELS subtests and reduces the assessment burden for schools, permitting more time for instruction.



	HOW MEASURED?	EXPECTED OUTCOMES
EFFICACY	Number of Evaluations	Number of evaluations should decrease at first and remain stable.
	(Number of Students who Qualify for Services/ Number of Students Evaluated) * 100%	Percentage of students evaluated who qualify should increase.
	Number of Classwide Learning Problems per Semester and Year	Number of classwide learning problems per semester and year should decrease.
	Number of Students in Risk Range at Tier 1 Screening across Years	Number of students in the risk range should decrease over time.
	<ol style="list-style-type: none"> 1. (Number of students with failed RtI/Number of students in school) * 100% 2. (Number of students with failed RtI/Number of students participating in Tier 3 intervention) *100% 	The percentage of screened students (all students) who have a failed RtI should approximate 2%. The percentage of students who receive Tier 3 intervention and have a failed RtI should approximate 10%.
	(Number of Students Not Proficient on Year-End Test/Number of Students in School) * 100%	Percentage of students meeting the proficiency criterion on the year-end test should improve.
	Number of Students with Successful RtI who are Detected on Subsequent Screenings.	The number of students repeatedly detected in the risk range should shrink over time.
EQUITY	<ol style="list-style-type: none"> 1. (Number of Boys who Qualify/Number of students Evaluated) * 100% 2. (Number of Students of a Certain Ethnicity who Qualify/Number of Students Evaluated) * 100% 3. (Number of Students who Receive Free or Reduced Lunch and Qualify/ Number of Students who Receive Free or Reduced Lunch in the school) * 100% 4. (Number of Boys who Qualify/Number of Boys in School) * 100% 5. (Number in Certain Ethnic Minority Category who 	<ol style="list-style-type: none"> 1. Percentage of evaluations that are boys should be about 50% 2. Percentage of evaluations for certain ethnicity should be about the same as the percentage of students in the school who belong to the same ethnic category. 3. Percentage of evaluations that occur for students living in poverty should be about the same as the percentage of students living in poverty in the school.

	<p>Qualify/Number of Students of Same Ethnicity in School) * 100%</p> <p>6. (Number of Students who Receive Free or Reduced Lunch and Qualify/ Number of Students who Receive Free or Reduced Lunch) * 100%</p>	<p>4. Percentage of boys who are evaluated should be about the same as the overall percentage of students evaluated.</p> <p>5. Percentage of students of a certain ethnicity who are evaluated should be about the same as the overall percentage of students evaluated.</p> <p>6. Percentage of students living poverty who are evaluated should be about the same as the overall percentage of students evaluated.</p>
	<p>Number of Students in Risk Range at Tier 1 Screening by Demographics</p>	<p>Number of students in the risk range should shrink over time for all students and reductions should be apparent by ethnicity, gender, poverty, mobility, and ELL status.</p>
	<p>Percent Failed Rtl by Demographics</p>	<p>The percentage of students with a failed Rtl should be about the same across ethnicity, gender, poverty, mobility, and ELL status categories.</p>
	<p>1. (Number of Boys Not Proficient on Year-End Test/ Number of Students Not Proficient on Year-End Test) * 100%</p> <p>2. (Number of Students of a Certain Ethnicity Not Proficient on Year-End Test/Number of Students Not Proficient on Year-End Test) * 100%</p> <p>3. (Number of Students of a Certain Ethnicity Not Proficient on Year-End Test/Number of Students of the Same Ethnicity) * 100%</p> <p>4. (Number of Students who Receive Free and Reduced Lunch and Are Not Proficient on Year-End Test/Number of Students Not Proficient on Year-End Test) * 100%</p> <p>5. (Number of Students Who Receive Free and Reduced Lunch and Are Not Proficient on Year-End Test/Number of Students who Receive Free and</p>	<p>1. Percentage of students not proficient on the year-end test who are also boys should be about 50%.</p> <p>2. Percentage of students who belong to a certain ethnic category and are not proficient on year-end test should be about the same as the percentage of students in the school who list the same ethnicity.</p> <p>3. Percentage of students of a certain ethnicity who are not proficient should be about the same as the overall percentage of students who are not proficient.</p> <p>4. Percentage of students living in poverty who are not proficient should be about the same as the percentage of students who are not proficient overall.</p>

	Reduced Lunch) * 100%	5. Percentage of students living in poverty who are not proficient should be about the same as the percentage of students living in poverty enrolled in the school.
EFFICIENCY	Number of Evaluations	Number of evaluations should decrease at first and remain stable.
	(Number of Students who Qualify for Services/ Number of Students Evaluated) * 100%	Percentage of students evaluated who qualify should increase.
	Number of Students Receiving Tier 2 and 3 Interventions	Percentage of students receiving Tier 3 intervention should shrink below 10% over time. Percentage of students receiving Tier 2 interventions should shrink below 20% over time.

References

- Ardoin, S. P., & Christ, T. J. (2008). Evaluating curriculum-based measurement slope estimates using data from triannual universal screenings. *School Psychology Review, 37*, 109–125.
- Ardoin, S. P., Witt, J. C., Suldo, S. M., Koenig, J., McDonald, E., & Smith, L. (2004). The response in response to intervention: Evaluating the sensitivity of curriculum-based measurement to intervention effects. *School Psychology Review, 33*, 218–233.
- Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities: Research & Practice, 18*, 172–186.
- Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance of decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*, 257–288.
- Griffiths, A. J., VanDerHeyden, A. M., Skokut, M., & Lilles, E. (2009). Progress monitoring in oral reading fluency within the context of RTI. *School Psychology Quarterly, 24*, 13–23.
- Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the dynamic indicators of basic early literacy skills and the comprehensive test of phonological processing. *School Psychology Review, 32*, 541-556.
- Schatschneider, C., Wagner, R. K., & Crawford, E. C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences, 18*, 308–315.
- Shapiro, E. S., & Clemens, N. H. (2009). A conceptual model for evaluating system effects of response to intervention. *Assessment for Effective Intervention, 35*, 3-16.
- VanDerHeyden, A. M., & Burns, M. K. (2010). *Essentials of response to intervention*. New York: Wiley.