Comparing the Cost-Accuracy Ratio of Multiple Approaches to Reading Screening in Elementary Schools

Courtenay A. Barrett

Lindy J. Johnson

Adrea J. Truckenmiller

Michigan State University

Amanda M. VanDerHeyden

Education Research and Consulting

Author Note

Courtenay A. Barrett https://orcid.org/0000-0001-9258-4725

Lindy J. Johnson https://orcid.org/0000-0002-2769-4642

Adrea J. Truckenmiller https://orcid.org/0000-0002-6101-6175

The first three authors contributed equally to the production of this article and therefore share first author status. Correspondence for this article should be addressed to Courtenay Barrett morsicou@msu.edu.

Support was provided by Grant H325H190003 for Lindy J. Johnson, from the Office of Special Education Programs (OSEP). The authors would like to acknowledge the considerable efforts of the participating school and comments on the manuscript from Dr. Margaret Kuklinski.

Abstract

Elementary schools administer reading screening assessments to identify students in need of remedial instruction. However, the administration of additional assessments comes at an opportunity cost and it is currently unclear the extent to which multiple types of reading screeners warrant the increase in resources that could be used for instruction. This study compared the cost-accuracy ratios for three types of reading screeners in Grade 3: curriculum-based measurement (Acadience), computer adaptive assessment (Star), and informal reading inventory (Fountas and Pinnell Benchmark Assessment System). Further, we demonstrated how schools can use classification and regression tree analysis to identify local cut-scores, maximizing classification accuracy. Results suggested that a multivariate approach, which included a computer adaptive assessment and oral reading fluency yielded the greatest accuracy and lowest cost-accuracy ratio. Furthermore, the Classification and Regression Tree analysis signaled two different instructional implications for subgroups of students with specific profiles on the screening measures.

Keywords: reading screening, classification accuracy, cost analysis

Comparing the Cost-Accuracy Ratios of Multiple Approaches to Reading Screening in Elementary Schools

Screening students' reading development can lead to more individualized and effective instruction that improves elementary reading outcomes (Connor, 2019), which can lead to later literacy development, improved graduation rates (Blachman et al., 2014), and fewer negative life outcomes associated with illiteracy (Cree et al., 2012). Many states require that all students in Kindergarten through Grade 3 be identified for risk of not meeting grade-level reading proficiency (Council of Chief State School Officers, 2019). The primary purpose of screening is early identification of students in need of remedial instruction at a time when the effects of reading intervention can have the largest long-term benefits (Blachman et al., 2014). As more schools have adopted reading screenings, questions about the best types, acceptable levels of accuracy, timing and number of measures to administer, and how to use scores to make instructional decisions have become pressing (Compton et al., 2010).

Examining how schools conduct screening is critical because the resulting decisions provide remediation for some students and not others, and the assessments may take the place of instruction that might have yielded other benefits (i.e., opportunity cost; Levin et al., 2018). The purpose of the study was threefold: (a) to develop a metric for applying the concept of cost effectiveness to the context of screening in schools, (b) to illustrate a method for determining the cost-accuracy of using reading screeners to identify groupings of student performance that may be instructionally useful (Kiernan et al., 2001), and (c) to compare the costs and cost-accuracy of common reading screening approaches.

Reading Screening

Three approaches to reading screening predominate: curriculum-based measurement, computer adaptive assessments, and informal reading inventories. Each approach differs in several aspects valued by schools: technical adequacy, instructional utility and administration time and costs (summarized in Table 1). Curriculum-based Measurement in reading (CBM-R; Deno, 2003) has been used extensively as a universal screening (Wayman et al., 2007). CBM has high technical adequacy with multiple equivalent forms, which facilitates measurement of incremental growth (Deno, 2003). For instructional utility, educators use some metrics within CBM-R to adjust their instruction for building fluency, decoding, and phonemic awareness.

In computer adaptive assessment, the computer uses the students' performance on previous items to select easier or more difficult items, which provides more precise information and reduces administration time and errors (Wainer et al., 2000). There are several technically-sound commercial computer adaptive assessments (e.g., FASTBridge, iReady, NWEA MAP Growth, RAPID; National Center on Intensive Intervention, 2021). These assessments vary in instructional utility as some provide scores aligned with broad strands of reading standards (e.g., NWEA MAP Growth) and some provide scores aligned with research-based domains of reading useful for guiding intervention (e.g., iReady, RAPID). Computer adaptive assessments are the most reliable for capturing growth across school years because they do not rely on the grade level of the passages (see Mitchell et al., 2015).

Informal reading inventories, such as the Benchmark Assessment System, are the most widely used by general education teachers (Ford & Opitz, 2008) for making instructional decisions (e.g., forming guided reading groups) as they have easily interpretable scores (e.g., grade level). Utility for screening is limited due to the potential for high rates of administrator error, lack of technical adequacy (Klingbeil et al., 2015), lack of bias analysis, time-intensive

administration, and larger measurement error due to lack of text equivalence across forms (Tortorelli, 2019). However, educators have been using informal reading inventories for much longer than either CBM or computer adaptive assessments have existed, and use of informal inventories for instruction is considered standard practice in most teacher preparation programs.

Multivariate Screening

In multivariate screening, several screeners are administered to all students and then a combination of scores is used for decision making. Research on multivariate screening is mixed, with some finding that it improves accuracy (Compton et al., 2010), some finding that it can decrease accuracy (VanDerHeyden et al., 2018), and others recommending that the very small improvement in accuracy is not worth the opportunity cost required to administer multiple assessments (Klingbeil et al., 2017; VanDerHeyden et al., 2018).

If additional screening measures are detracting from teachers' time for instruction or intervention, schools need to justify what benefit each measure provides (Clemens et al., 2016). Collectively, these tradeoffs include (a) teachers' time spent in training, administering, and scoring assessments instead of instruction; (b) opportunities for error in weighing multiple scores and score types; and (c) financial resources that could be spent on instructional materials rather than assessment materials. These tradeoffs can be collectively quantified as 'opportunity costs' (Levin et al., 2018), and represent how we conceptualized and calculated costs in the current study. Further, experts recommend that schools evaluate the cut-scores used in their screening system because schools may use different outcomes (e.g., different state tests) than the vendor's study and each school may have a different base rate of non-proficiency on the outcome which effects classification accuracy (Schatschneider et al., 2008).

Finally, to begin to understand the potential instructional utility of multiple screeners, we use classification agreement and regression tree (CART) analysis, opposed to prior research which has typically used multiple regression. Unlike regression, CART is a nonparametric procedure that identifies groups of students that are homogenous in their performance on screeners rather than assuming students fall in a rank order on a normal distribution of overall reading performance. In this way, CART has utility for identifying groups of students with similar instructional needs (Kiernan et al., 2001). Next, we describe cost-accuracy ratios, which combines opportunity cost with classification accuracy to evaluate the value of each of the screening approaches.

Combining Cost with Classification Accuracy in a Cost-Accuracy Ratio

Cost-accuracy analysis is conceptually similar to cost-effectiveness analysis, a formal methodology that calculates cost-effectiveness ratios by dividing per-student costs by the effect (e.g., effect size) of an evidence-based program (per student cost / effect) (Levin et al., 2018). In cost-accuracy ratios, we replace effect with two measures of screening accuracy: overall classification accuracy and negative posttest probability. Negative posttest probabilities indicate improvement in probability of correct identification of risk in a given context (base rate). For example, if negative posttest probability were 10% in a context where 59% of students failed the year-end test, then the probability of correct identification of true positives was improved by 49%. On the other hand, if negative posttest probability were 55% in a context where 59% of students failed the year-end test, use of the measure should be questioned because it did not improve the probability of identification of risk in that particular context. In the case of overall classification accuracy, the cost-accuracy ratio reflects the cost to accurately identify one student.

In the case of negative posttest probability, the cost-accuracy ratio reflects the cost to reduce the probability of a false negative error by 1%.

Research Questions

In this study we sought to help schools weigh the value of various aspects of their screening system by asking: (1) What were the costs and cost-accuracy ratios associated with four approaches to reading screening assessments? (2) What were the best cut-scores for each assessment in this particular study and did CART analysis identify homogeneous groups of students with specific instructional implications?

Method

Sample

Archival data were collected from 114 Grade 3 students (110 in the analysis) in six classrooms in one elementary school in Michigan during the 2018-2019 school year. The school served 618 students and was designated a Title I school with 65% of the Grade 3 students qualifying for free or reduced-price lunch. Students' scores on the Grade 3 state achievement test were not significantly different from the mean state score for Grade 3 (t (112) = -0.66, p = .510). Student demographics are further detailed in Table 2.

Screening Measures

Although the school administered reading assessments three times per year, this study used fall screening scores collected in September for the analyses.

Curriculum-Based Measurement

The Acadience Reading assessment is a composite score that weighs four scores using the following formula from the technical manual: Acadience Reading Composite score = Oral

Reading Fluency + (2 × Retell) + (4 × Maze Adjusted Score) + Oral Reading Accuracy (Acadience Learning Inc., 2021). The four scores and composite were entered into analys

Oral Reading Fluency and Accuracy. Students read three passages aloud for 1 minute each. The assessor marked word substitutions, omissions, and pauses of greater than 3 seconds as errors. The oral reading fluency score was calculated by subtracting the number of errors from the total number of words read for each passage. The recorded oral reading fluency score was the median words read correctly per minute from the three passages. Oral reading accuracy for each passage was calculated by dividing the words correct by the number of correct words plus incorrect words (i.e., total words read), then multiplying by 100%. The median percent of correctly read words was recorded as the oral reading accuracy score.

Retell. Following oral reading of each passage, the student was asked to retell the story. The assessor tracked the total number of retell words by marking a line through consecutive numbers in the scoring booklet for each word the student said that related to the passage. The median total number of words recalled was recorded.

Maze. Maze was measured separately in a group format. Students were asked to silently read a passage for 3 minutes in which every seventh word was replaced with a box containing the missing word and two distractor words. The Maze score (adjusted for the composite) was calculated as the correct responses minus incorrect responses divided by 2.

Computer-Adaptive Assessment

The Star Reading assessment (Renaissance Learning, Inc., 2022) is a group-administered test that consisted of 34 multiple choice vocabulary-in-context questions and reading comprehension questions. Student performance resulted in a scaled score ranging from 0-1400.

Informal Reading Inventory

In the Benchmark Assessment System, teachers chose to administer either an informational or literary text. Accuracy was measured by the percentage of words read correctly on the full text and comprehension was rated on a 4-point scale. The assessor used accuracy and comprehension to identify the students' instructional level (i.e., Level A-Z). Level N is designated as the beginning of Grade 3 (Fountas & Pinnell, 2016). For analyses, the letter was translated into a numerical scale (i.e., A = 1, B =2, etc.).

Criterion Measure: Michigan Student Test of Educational Progress (M-STEP)

The Michigan Student Test of Educational Progress (M-STEP) in English Language Arts (ELA) is a summative computer-adaptive assessment given annually to students in Grades 3 through 7 (Michigan Department of Education [MDE], 2019). The 2019 M-STEP was derived from the Smarter Balanced Assessment Consortium (SBAC) assessments (variations used by 12 other states). Students received an overall scaled score.

Cost Data

Cost data were collected retrospectively during the fall of 2021 through semi-structured interviews with school staff to comprehensively capture all of the resources needed to implement the assessments as intended (i.e., the ingredients method; (Levin et al., 2018). The interview protocol was based on Hollands et al. (2021) and asked school staff about the quality and quantity of resources needed for implementation. Additionally, school staff were asked to review school records or budgets to answer some of the questions. The interviews were recorded and then a member checking process was used to ensure the trustworthiness of the data.

For each assessment program, costs were incurred for: financial costs to purchase the assessment program, materials, and data management system, personnel costs for assessor training, administration, monitoring, and scoring the assessments, personnel costs for substitute

teachers, and equipment for administering and scoring the assessments (e.g., Chromebooks). Financial values were assigned to units within each of the ingredients based on the school's Enterprise Resource Planning (ERP) software program. To maximize generalizability, some values were estimated using the *CostOut Toolkit* (Hollands et al., 2015), based on the year they were incurred. Fixed costs for training and equipment were annualized over the lifetime of the resource to account for the depreciation of resources over time, as well as the interest accrued for the non-depreciated portion (Levin et al., 2018). Chromebooks were annualized over 5 years; the remaining fixed costs were annualized over 3 years, based on the semi-structured interviews, and both were assumed to have a standard 5% interest rate. All assessments were administered in the general education classroom; therefore, facilities costs were excluded from the cost analysis.

Data Analysis

Classification Accuracy

We conducted three CART analyses (using SAS 9.4) on each screener separately and one multivariate CART analysis that combined the scores from the three screeners. CART models partition each student's performance at every possible cut-score on each assessment to split the sample into mutually exclusive subgroups in incremental steps. By using this nonparametric splitting approach, CART is not limited by collinearity of predictor variables as is logistic regression. By default, SAS uses an entropy-based split criterion, a cost-complexity pruning method and subtree evaluation criterion with 10-fold cross validation procedure, which seeks to reduce the average misclassification rate (SAS Institute Inc., 2015). SAS randomly assigns observations to fold in the cross-validation which can result in slightly different results; therefore, we set the initial seed for random number generation at 123 to facilitate replication of the results. CART analysis generates (a) a decision tree that provides the optimal cut-scores in

the predictor variable(s), (b) the number of students classified in each node, which are homogeneous groups of student performance, and (c) a 2x2 contingency table indicating the number of true and false positives and true and false negatives.

The contingency tables were used to calculate classification accuracy. In this study we used overall classification accuracy because of its ease of interpretation for schools in the cost-accuracy ratio and the negative posttest probability statistic because it accounts for base rates of risk on the screening in a specific population (VanDerHeyden, 2013). We also report sensitivity and specificity values for context because these are typically reported. Formulas for each are provided in supplementary online materials. Sensitivity values above .90 and negative posttest probabilities near or below .10 are considered optimal for screening purposes (Jenkins et al. 2007; VanDerHeyden, 2013).

Cost Analysis

For the cost and cost-accuracy analyses, ingredients identified during the semi-structured interviews were first determined to be fixed (did not vary based on the number of students) or variable (varied based on the number of students). Fixed costs included professional development, teacher kits, and other classroom- or teacher-level resources. Variable costs included student materials and equipment, and other student-level resources. Per student costs were calculated by multiplying the units by the unit prices for each individual ingredient, summing the costs across all ingredients, and then dividing the total costs for each program by the number of students. Costs were calculated for the fall benchmark alone to align with the classification accuracy analyses, as well as for the entire academic year to align with typical educational practice.

Cost-accuracy ratios were calculated for overall classification accuracy by dividing the total costs by the number of accurately identified students (total cost / [true negative + true positive]). Cost-accuracy ratios for negative posttest probability were calculated by first subtracting the negative posttest probability from the base rate, which indicates improvement in probability of correct identification of risk in a given context. Then, the total costs were divided by this difference score, to indicate the costs to obtain gains in probability of correct identification of risk above those that could be obtained by chance alone. Because cost effectiveness (and cost accuracy) analyses usually convert metrics to a per student value to assist in comparability of results across studies, we computed the cost to improve (lower) negative posttest probability by 1% (see online supplementary materials for equations). Finally, supplementary analyses were conducted to understand the extent to which the cost-accuracy ratios were robust to variations in the assumptions.

Results

Descriptive Statistics

Complete data were available for 110 students. The 4 missing data points were missing at random (Little's Missing Completely at Random test χ^2 (5) = 5.391, p = .370). The Benchmark Assessment System was left-skewed and leptokurtic (peaked) at Level N (beginning of Grade 3), with very few students scoring higher than Level N and the left tail of the distribution going down to Level C. Scores on the other screening measures were approximately normally distributed. Scores were statistically significantly lower on the Acadience Composite Score (t(110) = -4.35, p < .001) relative to the norm sample, but statistically similar to the norm sample on Acadience ORF, Star Reading, the Benchmark Assessment System, and the MSTEP.

Classification Accuracy and Multivariate CART Analysis

Classification accuracies for the screenings in predicting non-proficient performance on the MSTEP are provided in Table 3. None of the single measures on their own met recommended values of 90% sensitivity and 80% specificity (Jenkins et al., 2007) using the publisher-recommended cut-scores for calculations. These are provided for reference because the use of publisher-recommended cut-scores is typical practice in schools.

Next, to compare accuracies across the three individual measures and a multivariate combination of measures, classification analyses were conducted using cut-scores derived and tested on the study sample using CART analysis (Table 3). For the multivariate screening, CART identified two groups of at-risk students and 1 group of not-at-risk students and found that two scores best characterized student performance (see online supplementary materials for the CART decision tree). Students in the not at-risk category were characterized by scoring at or above 345 on the Star Reading and reading at or above 81 words correctly on the Acadience ORF (n = 45). For the at-risk categories, there is one group performing below 345 on the Star Reading (n = 51) and a group who scored above 345 on Star Reading but read less than 81 words correctly on Acadience ORF (n = 14). This model fit well with a low misclassification rate of less than 12 students (10.9%) misclassified. Other fit statistics included Average Square Error = 0.09, Entropy = 0.45, Gini = 0.18, Residual Sum of Squares = 19.96.

Overall, the sensitivity and specificity values were stronger for Acadience and the CART-identified multivariate approach relative to STAR and the Benchmark Assessment System. Sensitivities were comparable at roughly 85% (Acadience) and 91% (multivariate) and specificities were identical at roughly 87%. Notably, the derived cut-score for Acadience selected only the ORF score rather than the other components of the composite.

Cost Analysis and Cost-Accuracy Analysis

Tables 4, 5, and 6 present the costs per ingredient for Acadience Reading, Star Reading, and the Benchmark Assessment System, respectively. Results indicated that the total costs to implement Acadience Reading Composite for the fall benchmark was $1,921.94 and the per student cost was $17.47. The majority of the costs were related to personnel time for training and coaching (57.48%), and administration and scoring (34.74%), with very few out-of-pocket financial costs for the program and materials (7.78%). For Star Reading, results indicated that the total cost for the fall benchmark was $2,258.11 and the per student cost was $20.34. A smaller percentage of the costs were related to opportunity costs for personnel time for training, purchasing the programs, and administration (29.72%). The majority of the costs were for the Chromebooks (70.28%). Supplementary analyses indicated that excluding the costs for the Chromebooks, which may reflect school contexts in which technology is readily available and not perceived as a significant opportunity cost, resulted in total costs of $671.14 for the fall and $6.05 per student. Finally, for the Benchmark Assessment System, results indicated that the total cost for the fall benchmark was $5,602.50 and the per student cost was $50.93. Most of the costs for were for personnel time to administer and score the assessments (78.35% for the fall).

Table 7 presents the results of the cost-accuracy analyses which calculated the costs to correctly identify one student out of the entire sample (overall classification accuracy), as well as the costs to improve negative posttest probability in this context. Results suggested that Acadience Reading and Star Reading were the most cost-accurate options for correctly identifying students. Out of the total costs that it took to administer the screeners in the fall, it cost the school $19.78 to correctly identify 1 student using Acadience Oral Reading Fluency and $24.28 to correctly identify 1 student using Star Reading as either at risk or not at risk. Supplementary analyses suggested that when excluding the opportunity costs for Chromebooks,

which may reflect resource-rich environments, Star Reading cost $7.22 to correctly identify 1 student. The Benchmark Assessment System cost $64.40 to correctly identify 1 student. Finally, analyses examining the cost-accuracy of the multivariate approach (Star Reading and Acadience ORF) suggested that it cost $42.01 to correctly identify 1 student (overall classification accuracy). Table 7 also presents the cost-accuracy results for negative posttest probability. Acadience Oral Reading Fluency required the least amount of resources to improve negative posttest probability by 1% (or stated another way, lowered the probability that the screening would fail to detect the student who was going to fail the year-end test). The multivariate approach was the second most cost-effective, costing $196.28 to improve negative posttest probability by 1%.

Discussion

This study provides several considerations for schools weighing the opportunity costs of screeners and the value they obtain from them. We provided quantitative information for schools to directly compare the costs associated with several aspects of screening (e.g., training) and provided cost-accuracy ratios to inform decision making. The costs involved in training suggests that schools should be wary of frequently switching assessment programs, particularly within one type of overarching approach (e.g., CBM or computer adaptive tests), as this would require a significant amount of resources with minimal improvements in classification accuracy. Indeed, classification accuracy differences between different CBM systems and different computer adaptive assessments are negligible (NCII, 2021).

For informal reading inventories, our results aligned with prior research suggesting that informal reading inventories can cause additional misclassification (Klingbeil et al., 2015; VanDerHeyden et al., 2018). These data should not be considered in screening decisions

(determining *who* needs intervention and evaluating school-wide progress). However, there are powerful effects of formative assessment when teachers listen to students read (Black & Wiliam, 1998) and we do not recommend that schools eschew opportunities for teachers to listen to all of their students read simply because it is not an efficient component of the screening process. Perhaps schools could consider diverting those resources to professional learning on how to effectively use formative assessment of oral reading to guide instruction (Heritage, 2008).

Our results provide further evidence that schools and researchers should check the classification accuracy of their reading screening assessments (Schatschneider et al., 2008). When using the publisher-recommended cut-scores in this study, sensitivity was remarkably low, detecting only about three to six of every ten children who needed remedial instruction. Negative posttest probability was near chance for the Benchmark Assessment System and only marginally better than chance for Acadience and STAR, and about three to four times the 10% maximum threshold recommended (VanDerHeyden, 2013).

Accuracy improved when sample-dependent cut-scores were derived from CART analysis and then classification accuracies were subsequently reported on the same sample. It was expected that accuracy would be higher because deriving thresholds and testing accuracies on the same sample inflates accuracy estimates (Jenkins et al., 2007). Results indicated that classification accuracy in all categories (except specificity) was highest for the multivariate approach, followed by Acadience Oral Reading Fluency, with Star and BAS demonstrating tradeoffs between the classification accuracy metrics. Interestingly, the CART analysis selected oral reading fluency as the only necessary subscore from the Acadience battery. The multivariate approach correctly identified 4 more students and had a negative posttest probability of 13.2%, which was much closer to the acceptable rate of 10% than the negative posttest probability of

20.4% for oral reading fluency. This replicates past findings that there is added accuracy with multiple measures. Schools may determine whether the additional costs needed to administer multiple screeners is worth it, depending on the resources available in their local context (e.g., technology), as well as local goals, values, and culture related to screening. Results from the cost analysis, which identifies the quality and quantity of each of the resources needed, will be useful in this regard.

The additional value of screening in schools is to inform remedial instruction and intervention. All three assessments measured some predictor of reading comprehension, but each approach offered different information about students' development of domains of reading. The Star is most like the MSTEP; however, the Star score alone was not enough to capture students who needed remediation. Multivariate CART analyses suggested there were two groups of at-risk students in this school: (1) a group who needed remedial instruction in decoding, fluency, vocabulary, and/or reading comprehension, and (2) a group who only needed additional decoding and/or fluency-building instruction. It is possible that Star Reading misidentified a subgroup of students who had sufficient oral language proficiency to compensate for their more limited decoding and oral reading fluency skills to exceed the cut-score on Star but not meet proficiency on MSTEP. This misclassification of false negative errors may be more likely to occur in elementary school when the text is simple enough that students can infer meaning and correctly answer vocabulary and comprehension questions even when they cannot decode all of the words or decode slowly. It is important to identify this group of students in Grades 2 and 3 (Fletcher et al., 2021) before texts become more complex. In later grade levels, these students will be detected on larger-grained assessments like Star. Models of reading fluency within the context of other reading domains support this interpretation that reading fluency is essential to measure for

most students before 4th grade when oral language domains (i.e., vocabulary and syntax) become more predictive for most students (Foorman et al., 2017).

Acadience was the single measure that was most cost effective, but it was not the composite score that had the greatest accuracy. For this grade level at this school, the additional metrics did not improve accuracy. In fact, use of the additional measures and scores worsened accuracy. This is not surprising given the construct validity of retell and MAZE. Although retell and MAZE tasks have higher face validity as reading comprehension metrics, ORF had higher validity coefficients with criterion reading comprehension performance (Good et al., 2019). This may be due to the higher burden of decoding demands compared to language comprehension in Grade 3 text (Tortorelli, 2019).

Experts in evidence-based reading screening practices emphasize the difference between screening and diagnostic purposes (Fletcher et al., 2021). Brief, inexpensive, and highly predictive metrics like ORF should be used in Grades 2 and 3 to *screen* students for risk to provide remedial instruction to these students. Then, additional diagnostic assessment should be used after screening to inform instruction. Or, screening tools should include brief measures of reading constructs that are highly predictive and instructionally useful. These constructs include letter-sound correspondence and phonological awareness in younger grade levels, as well as vocabulary or other oral language skills across grade levels (e.g., Florida Center for Reading Research Reading Assessment; Foorman et al., 2015).

Limitations

The conclusions drawn from this study should be tempered by several limitations. First, the study only included one school with a relatively small sample size. However, the sample size was comparable to other studies providing validity evidence of reading screeners (e.g., $N = 184$

Grade 3 students and 102 Grade 6 students; Good et al., 2019). The generalizability of the cost analyses and the classification accuracy may be limited. For example, other schools may require more or less training, as well as costs to facilitate buy-in, which would alter the cost results (Barrett et al., 2020). Classification accuracy statistics and CART analyses are nonparametric analyses that apply specifically to the base rate of proficiency on the specific outcome used by this school and cannot generalize to a broader population. Furthermore, sources of error in CART analysis are still being explored. The at-risk subgroups of students study need to be replicated and explored in more robust analyses (e.g., Foorman et al., 2017). However, our results generally aligned with several prior studies, suggesting that future studies may not yield substantively different results.

Suggestions for Future Research

Most teachers already know the approximate level of their students' reading development (e.g., VanDerHeyden et al., 2018) and derive more value from assessment that has direct implications for instruction (Ford & Opitz, 2008). Future research is needed to operationalize what effective teachers do with assessment information so that it can be replicated by other teachers. Even when accurate screening assessments are used, many teachers are not provided with the knowledge or support to select evidence-based instruction that meets their students' needs. Future research is needed in the space between screening and the delivery of remedial instruction. Further, none of the assessments provided a reliable score for other important reading domains (e.g., vocabulary, morphology, syntax, text structure) that link directly to the most effective reading interventions (Connor, 2019; Truckenmiller & Brehmer, 2021). Future research is needed to assess these constructs, in tandem with screening, to improve classification accuracy and instructional utility, thereby improving the value of screening systems.

References

Barrett, C. A., Pas, E. T., & Lindstrom Johnson, S. (2020). A cost analysis of the innovation–decision process of an evidence-based practice in schools. *School Mental Health*, *12*(3), 638–649. https://doi.org/10.1007/s12310-020-09372-z

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Murray, M. S., Munger, K. A., & Vaughn, M. G. (2014). Intensive reading remediation in grade 2 or 3: Are there effects a decade later? *Journal of Educational Psychology*, *106*(1), 46–57. https://doi.org/10.1037/a0033663

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Clemens, N. H., Keller-Margulis, M. A., Scholten, T., & Yoon, M. (2016). Screening assessment within a multi-tiered system of support: Current practices, advances, and next steps. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of Response to Intervention* (pp. 187–213). Springer US. https://doi.org/10.1007/978-1-4899-7568-3_12

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology*, *102*(2), 327–340. https://doi.org/10.1037/a0018448

Connor, C. M. (2019). Using technology and assessment to personalize instruction: Preventing reading problems. *Prevention Science*, *20*(1), 89–99. https://doi.org/10.1007/s11121-017-0842-9

Council of Chief State School Officers & Center on Enhancing Early Learning Outcomes. (2019). *Third grade reading laws: Implementation and impact*. Council of Chief State School Officers. http://ceelo.org/wp-content/uploads/2019/09/CCSSO_CEELO_third_grade_reading.pdf

Cree, A., Kay, A., & June Steward. (2012). *The economic and social cost of illiteracy: A snapshot of illiteracy in a global context*. World Literacy Foundation. http://hdl.voced.edu.au/10707/321997

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, *37*(3), 184–192. https://doi.org/10.1177/00224669030370030801

Fletcher, J. M., Francis, D. J., Foorman, B. R., & Schatschneider, C. (2021). Early detection of dyslexia risk: Development of brief, teacher-administered screens. *Learning Disability Quarterly*, *44*(3), 145–157. https://doi.org/10.1177/0731948720931870

Foorman, B., Petscher, Y., & Schatschneider, C. (2015). *Florida Center for Reading Research (FCRR) Reading Assessments (FRA) Kindergarten to Grade 2 [Technical manual]*. http://www.fcrr.org/for-researchers/fra.asp

Foorman, B. R., Petscher, Y., Stanley, C., & Truckenmiller, A. (2017). Latent profiles of reading and language and their association with standardized reading outcomes in kindergarten through tenth grade. *Journal of Research on Educational Effectiveness*, *10*(3), 619–645. https://doi.org/10.1080/19345747.2016.1237597

Ford, M. P., & Opitz, M. F. (2008). A national survey of guided reading practices: What we can learn from primary teachers. *Literacy Research and Instruction*, *47*(4), 309–331. https://doi.org/10.1080/19388070802332895

Fountas, I., & Pinnell, G. S. (2016). *Fountas & Pinnell Benchmark Assessment System* (3rd ed.). Heinemann.

Good, R. H., Kaminski, R. A., Dewey, E. N., Wallin, J., Powell-Smith, K. A., & Latimer, R. (2019). *Acadience Reading K-6 Technical Manual*. Dynamic Measurement Group, Inc. https://acadiencelearning.org/

Heinemann. (2012). *Field study of reliability and validity of the Fountas & Pinnell benchmark assessment systems 1 and 2*. Heinemann. https://www.fountasandpinnell.com/shared/resources/FP_BAS_Research_Field-Study-Full-Report.pdf

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Council of Chief State School Officers.

Hollands, F. M., Hanisch-Cerda, B., Levin, H. M., Belfield, C. R., Menon, A., Shand, R., Pan, Y., Bakir, I., & Cheng, H. (2015). *Costout—The CBCSE Cost Tool Kit*. Center for Benefit-Cost Studies of Education, Teachers College, Columbia University. www.cbcsecosttoolkit.org

Hollands, F. M., Pan, Y., Kieffer, M. J., Holmes, V. R., Wang, Y., Escueta, M., Head, L., & Muroga, A. (2021). Comparing evidence on the effectiveness of reading resources from expert ratings, practitioner judgements, and research repositories. *Evidence & Policy: A Journal of Research, Debate and Practice*. https://doi.org/10.1332/174426421X16366418828079

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review*, *36*(4), 582–600. https://doi.org/10.1080/02796015.2007.12087919

Kiernan, M., Kraemer, H. C., Winkleby, M. A., King, A. C., & Taylor, C. B. (2001). Do logistic regression and signal detection identify different subgroups at risk? Implications for the design of tailored interventions. *Psychological Methods*, *6*(1), 35–48. https://doi.org/10.1037/1082-989X.6.1.35

Klingbeil, D. A., McComas, J. J., Burns, M. K., & Helman, L. (2015). Comparison of predictive validity and diagnostic accuracy of screening measures of reading skills. *Psychology in the Schools*, *52*(5), 500–514. https://doi.org/10.1002/pits.21839

Klingbeil, D. A., Nelson, P. M., Van Norman, E. R., & Birr, C. (2017). Diagnostic accuracy of multivariate universal screening procedures for reading in upper elementary grades. *Remedial and Special Education*, *38*(5), 308–320. https://doi.org/10.1177/0741932517697446

Levin, H. M., McEwan, P. J., Belfield, C., Bowden, A. B., & Shand, R. (2018). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis* (3rd ed.). Sage Publications Ltd.

Michigan Department of Education. (2019). *Spring 2019 Michigan Student Test of Educational Progress (M-STEP) Technical Report*.

Mitchell, A. M., Truckenmiller, A., & Petscher, Y. (2015). Computer-adaptive assessments: Fundamentals and considerations. *Communique*, *43*(8), 1–22.

National Center on Intensive Intervention. (2021). *Academic screening tools chart*. U.S. Department of Education, Office of Special Education Programs. https://charts.intensiveintervention.org/ascreening

Renaissance Learning, Inc. (2022). *Star assessments for reading technical manual*. Renaissance Learning, Inc. https://help.renaissance.com/US/PDF/SR/SRRPTechnicalManual.pdf

SAS Institute Inc. (2015). *SAS/STAT® 14.1 User's Guide*. SAS Institute Inc.

Schatschneider, C., Petscher, Y., & Williams, K. M. (2008). How to evaluate a screening process: The vocabulary of screening and what educators need to know. In L. M. Justice & C. Vukelich (Eds.), *Achieving excellence in preschool literacy instruction* (pp. 304–316). The Guilford Press.

Tortorelli, L. S. (2019). Reading rate in informational text: Norms and implications for theory and practice in the primary grades. *Reading Psychology*, *40*(3), 293–324. https://doi.org/10.1080/02702711.2019.1621011

Truckenmiller, A. J., & Brehmer, J. S. (2021). Making the most of tier 2 intervention: What decisions are made in successful studies? *Reading and Writing Quarterly*, *37*(3), 240–259. https://doi.org/10.1080/10573569.2020.1768612

VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review*, *42*(4), 402–414. https://doi.org/10.1080/02796015.2013.12087462

VanDerHeyden, A. M., Burns, M. K., & Bonifay, W. (2018). Is more screening better? The relationship between frequent screening, accurate decisions, and reading proficiency. *School Psychology Review*, *47*(1), 62–82. https://doi.org/10.17105/SPR-2017-0017.V47-1

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge. https://doi.org/10.4324/9781410605931

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, *41*(2), 85–120. https://doi.org/10.1177/00224669070410020401

Table 1

Characteristics of Three Approaches to Reading Screening

Variable	Acadience Reading Composite	Star Reading	Fountas &Pinnell Benchmark Assessment System
Cutscore for Risk	180	323	Level K [a]
Overall Classification Accuracy	.85	.82	NR
Area Under the Curve	.88	.90	NR
Sensitivity	.51	.81	NR
Specificity	.94	.82	NR
Positive Predictive Power	.69	.53	NR
Negative Predictive Power	.88	.95	NR
Bias Analysis Conducted	Yes	Yes	NR
	Reliability		
Internal Consistency	.85[b]	.94[c]	NR
Test-Retest/ Alternate Forms	.97 - .99	.87	.97
Inter-rater	NR	NR	.26 - .43
Standard Error of Measurement	20.69	47	NR
	Validity		
Discriminant	1.80[d]	NR	NR
Convergent	NR	.84	.42 -.94[e]
Predictive	.78	.78	NR
Concurrent	.75	.79	NR
	Utility		
Administration	Individual & group – paper	Group-computer	Individual – paper
Approximate time to administer & score	30 min./per **student**	30 min./per **class**	40 min./per **student**
Approximate cost (per student/yr)	$30.88	$22.84	$130.82
Instructional utility	Contains student grouping worksheet	No subscores available for guiding instruction.	Levels aligned with books. Teachers observe reading.

Note. NR = Not Reported.
Acadience Reading Composite and Star Reading technical adequacy data retrieved from National Center on Intensive Intervention, unless otherwise noted. Fountas & Pinnell Benchmark Assessment System data retrieved from Heinemann, 2012, unless otherwise noted.
[a] School determined this cutscore for risk based on Heinemann's grade level range. [b] Good et al., 2019. [c] Renaissance Learning, Inc., 2022. [d] Reported as effect size (Cohen's d)
[e] Convergent validity ranged from .42 to .44 (nonfiction and fiction books, respectively) with Degrees of Reading Power texts to .93 to .94 (fiction and nonfiction books, respectively) with Reading Recovery assessment texts (Heinemann, 2012)

Table 2

Sample Characteristics (N = 110 Grade 3 Students)

Variable	*n*	% of Sample
Sex		
Female	54	49.1
Male	56	50.9
Race/Ethnicity		
American Indian	1	0.9
Asian	1	0.9
Black	1	0.9
Hispanic	6	5.5
White	101	91.8
Socioeconomic Status Proxy		
Free Lunch Eligible	62	54.6
Reduced-Price Lunch Eligible	3	2.7
Special Education Eligible	19	17.3
Limited English Proficiency	1	0.9

Table 3

Classification Accuracy for Four Approaches to Screening Using Cut-scores Optimized for the Current Sample

	Acadience Fluency		Star Reading		Benchmark Assessment System		Multivariate Approach
	Publisher (Composite)	CART-derived (ORF)	Publisher	CART-derived	School-chosen	CART-derived	CART-derived
Cut-score	180	81	323	345	14 (Level N)	13 (Level M)	346 (Star), 81 (ORF)*
Overall Correct Classification	70.9%	85.5%	77.3%	83.3%	58.2%	79.1%	89.1%
Sensitivity	52.3%	84.6%	63.1%	75.8%	30.8%	87.7%	90.8%
Specificity	97.8%	86.7%	97.8%	95.6%	97.8%	66.7%	86.7%
Negative Posttest Probability	41.2%	20.4%	35.1%	26.7%	50.4%	21.0%	13.2%

Note. *See Figure 1 in the online supplementary materials for the decision tree illustrating the CART-derived cut-scores. The base rate of non-proficiency for this sample was 59%. Classification accuracies should be compared across measures using cut-scores derived in the same way (CART-derived). Classification accuracies are provided for the publisher-recommended cut-scores as this approach reflects the typical practice of screening in schools and is a useful reference point demonstrating that the publisher-recommended cut-scores generate fewer correct decisions than do cut-scores derived on the actual sample.

Table 4

Costs per Ingredient for Acadience Reading Composite Score

Ingredients	Units and Unit Prices	Total Cost for Fall	Total Cost for Year
Professional Development and Coaching			
Professional Development for MTSS Coordinator (3 full-day workshops, 7.5 hours/day, annualized over 15 years)	22.5 hours x $51.04/hour x .0963	$110.59	$110.59
Professional Development for School-wide Assessment Team (5.5 hours of Training Videos)	5.5 hours x $22.00/hour x 6 team members	$726.00	$726.00
30-Day Access to Training Videos	$199.00	$199.00	$199.00
Coaching Provided by MTSS Coordinator (20 min./benchmark)	0.33 hours x $51.04/hour	$16.84	$51.04
Coaching Received by School-wide Assessment Team (20 min./benchmark)	0.33 hours x $22.00/hour x 6 team members	$43.56	$132.00
Monitoring by MTSS Coordinator (10 min./benchmark)	0.17 hours x $51.04/hour	$8.68	$25.52
Administration and Scoring			
Individual Administration by School-wide Assessment Team (10 min./student/benchmark x 110 students)	18.33 hours x $22.00/hour	$403.33	$1,210.00
Scoring by School-wide Assessment Team (5 min./student/benchmark x 110 students)	9.17 hours x $22.00/hour	$201.67	$605.00
Group Administration for Maze by Teachers (15 min./benchmark x 6 teachers)	0.25 hours/teacher x $41.86/hour x 6 teachers	$62.79	$188.37
Materials and Equipment			
Annual Access to Online Reporting System	$1.00/student x 110 students	$110.00	$110.00
Teacher Binders (6 binders, annualized over 5 years)	$14/binder x 6 binders x 0.231	$19.40	$19.40
Student Booklets	110 booklets	$15.00	$15.00
Timers (6 timers, annualized over 5 years)	$3.67/timer x 6 timers x 0.231	$5.08	$5.08
Total Costs		$1,921.94	$3,397.00
Per Student Cost		$17.47	$30.88

Table 5

Costs per Ingredient for Star Reading

Ingredients	Units and Unit Prices	Total Cost for Fall	Total Cost for Year
Training and Coaching			
Coaching and Monitoring by MTSS Coordinator (15 min./benchmark)	0.25 hours x $51.04/hour	$12.76	$38.28
Administration and Scoring			
Group Administration by Teachers (30 min./teacher/benchmark)	0.5 hours/teacher x $41.86/hour x 6 teachers	$125.58	$376.74
Materials and Equipment			
Program	$4.80/student x 111 students	$532.80	$532.80
Chromebooks (annualized over 5 years)	$229/Chromebook x 30 Chromebooks x 0.231	$1,586.97	$1,586.97
Total Costs		$2,258.11	$2,534.79
Per Student Cost		$20.34	$22.84

Table 6

Costs Analysis for Benchmark Assessment System

Ingredients	Units and Unit Prices	Total Cost for Fall	Total Cost for Year
Training and Coaching			
Professional Development for MTSS Coordinator	2 hours x $51.04/hour	$102.08	$102.08
Professional Development for Teachers	2 hours x $41.86/hour x 6 teachers	$502.32	$502.32
Monitoring by MTSS Coordinator (5 min./benchmark)	0.08 hours x $51.04/hour	$4.32	$12.96
Administration and Scoring			
Individual Administration and Scoring by Teachers (40 min./student/benchmark x 110 students)	73.33 hours x $41.86/hour	$3,069.73	$9,209.20
Substitute Teachers (2 days of substitutes/teacher/benchmark)	12 substitute teachers x $110/day	$1,320.00	$3,960.00
Materials and Equipment			
Scoring Booklets	1-3 booklets/child	$15.00	$15.00
Teacher Kits (annualized over 5 years)	$425/kit x 6 kits x 0.231	$589.05	$589.05
Total Costs		$5,602.50	$14,390.61
Per Student Cost		$50.93	$130.82

Table 7

Cost-Accuracy Results for Each Approach Using Cut-scores Optimized for the Current Sample

	Acadience Oral Reading Fluency	Star Reading	Benchmark Assessment System	Multivariate Approach (Star Reading + Acadience ORF)
Overall Correct Classification				
(TP + TN) / (TP + TN + FP + FN)	94 / 110	93 / 110	87 / 110	98 / 110
Total Cost for Fall for TP + TN + FP + FN	$1,859.15	$2,258.11	$5,602.50	$4,117.26
Cost to Correctly Identify 1 Student	$19.78	$24.28	$64.40	$42.01
Negative Posttest Probability	20.40%	26.70%	21.00%	13.20%
Additional percentage of students correctly identified beyond the base rate	38.60%	32.30%	38.00%	45.80%
Cost to correctly identify the percentage of students beyond the base rate who score above the cut-point on screener but below the cut-point on the outcome	$4,816.45	$6,991.05	$14,743.42	$8,989.65
Cost to correctly identify additional 1% of students who score above the cut-point on screener but below the cut-point on the outcome	$124.78	$216.44	$387.98	$196.28

Note. TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative. Base rate of non-proficiency for these analyses was 59%.

Supplementary Online Materials for Comparing the Cost-Accuracy Ratios of Multiple Approaches to Reading Screening in Elementary Schools

Formulas Used for Calculating Classification Accuracy and Cost-Accuracy Ratios

Classification accuracy is based on a 2x2 matrix of the number of students scoring above and below the risk cut-score on the screening and the number of students scoring above and below proficiency on the state achievement test. This results in four quadrants, with the number of students who were identified as (a) at-risk on the screening and then below proficiency on the outcome (true positive), (b) at-risk on the screening and then scored proficiently on the outcome (false positive), (c) not-at-risk on the screening and then below proficiency on the outcome (false negative), and (d) not-at-risk on the screening and then scored proficiently on the outcome (true negative)[1].

	MSTEP Below Proficient	MSTEP Above Proficient
Screening At-Risk	a (true positive)	b (false positive)
Screening Not At-Risk	c (false negative)	d (true negative)

Overall classification accuracy was calculated as

$$Overall\ classification\ accuracy = \frac{True\ Positive + True\ Negative}{Total\ n} \quad (1)$$

Sensitivity was computed as the number of individuals performing below the cut-score on both the screening and the MSTEP divided by the total number of students scoring below proficient MSTEP:

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

Specificity was computed as the number of students performing above the cutscore on the screening and the MSTEP divided by the total number of students scoring proficient on the MSTEP:

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (3)$$

Negative posttest probability is the probability of a student performing above the cut-score on the screening, but then scoring below proficiency on the year-end test. This is usually the type of error that schools want to avoid. Using the base rate of risk in the sample (59%), the base rate odds were computed and multiplied by the negative likelihood ratio and then converted from odds back to probability as follows to calculate negative posttest probability:

$$Negative\ Posttest\ Probability = \frac{\left[\frac{Base\ Rate\ of\ Risk}{1 - Base\ Rate\ of\ Risk}\right] \times \left[\frac{1 - Sensitivity}{Specificity}\right]}{\left[1 + \left[\left[\frac{Base\ Rate\ of\ Risk}{1 - Base\ Rate\ of\ Risk}\right] \times \left[\frac{1 - Sensitivity}{Specificity}\right]\right]\right]} \quad (4)$$

[1] Throughout this paper, we categorize performance on screening as at-risk of meeting proficiency on an outcome and we define non-proficiency on an outcome measure as the goal of screening classification accuracy. Because classification accuracy is so confusing, we use these terms to enhance clarity. However, we acknowledge that these terms have been historically used inappropriately as deficit language to blame children and families for 'failure' and to deny children access to high quality instruction. We define non-proficiency on state achievement tests as the failure of the education system to accelerate student development. Rather, we promote the interpretation of screening information as a way to effectively facilitate the positive life outcomes that can be achieved through early high-quality instruction tailored to students' level of development (e.g., Blachman et al., 2014; Connor, 2019).

Cost-accuracy ratios for negative posttest probability were calculated by first subtracting the negative posttest probability from the base rate, which indicates improvement in probability of correct identification of risk in a given context. Then, the total costs were divided by this difference score, to indicate the costs to obtain gains in probability of correct identification of risk above those that could be obtained by chance alone.

$$Cost\ per\ negative\ posttest\ probability = \frac{Total\ cost}{[Base\ rate - negative\ posttest\ probability]} \quad (5)$$

Because cost effectiveness (and cost accuracy) analyses usually convert metrics to a per student value to assist in comparability of results across studies, we computed the cost to improve (lower) negative posttest probability by 1% (see online supplementary materials for equations).

$$Cost\ per\ 1\%\ improvement\ in\ negative\ posttest\ probability = \frac{Cost\ per\ negative\ posttest\ probability}{[base\ rate - negative\ posttest\ probability]} \quad (6)$$

Figure 1

Classification and Regression Tree decision tree for using multiple reading screening assessment cut points that maximizes sensitivity and specificity for a local sample