Check for updates

Article

Hammill Institute on Disabilities

Assessment for Effective Intervention
38(1) 6–14
© 2012 Hammill Institute on Disabilities
Reprints and permission: http://www.sagepub.com/journalsPermissions.nav
DOI: 10.1177/1534508412451491
http://aei.sagepub.com
SAGE

Early Screening for Risk of Reading Disabilities: Recommendations for a Four-Step Screening System

Jennifer K. Gilbert, PhD[1], Donald L. Compton, PhD[1], Douglas Fuchs, PhD[1], and Lynn S. Fuchs, PhD[1]

Abstract

Response-to-intervention (RTI) models incorporate a screening process to identify students who appear to be at risk for learning disabilities (LDs). The purpose of this position article is to incorporate what is known about screening into a flexible, yet comprehensive screening system to help school psychologists and other school administrators in establishing school-specific screening procedures. The authors begin by discussing past research on screening for reading disabilities (RDs) within the RTI framework. Then, they propose a four-step screening system advocating a short screener (Step 1), progress monitoring (Step 2), follow-up testing (Step 3), and ongoing revision of procedures and cut scores (Step 4). Their goal is to improve screening within RTI systems with practical procedures to permit schools to implement state-of-the-art screening batteries that accurately and efficiently distinguish students who are at high risk for RD.

Keywords

screening, response-to-intervention, reading disabilities

Response-to-intervention (RTI) is a framework in which students who struggle academically are identified early and given supplementary intervention to ameliorate poor academic outcomes (National Joint Committee on Learning Disabilities, 2005). It is also a means of identifying learning disabilities (LDs) and a method for preventing the overidentification of LD due to poor instruction (D. Fuchs & Deshler, 2007; Vaughn & Fuchs, 2003). Various methods have been proposed to operationalize RTI, with current models favoring a three-tier system (see Bradley, Danielson, & Hallahan, 2002; L. S. Fuchs, Fuchs, & Speece, 2002; Vaughn & Fuchs, 2003). In terms of instruction, Tier 1 consists of research-based classroom instructional practices that have been proven effective for the majority of students. In Tier 2, students who do not respond sufficiently to Tier 1 are given more intensive instruction, often in the form of supplementary, small-group intervention. Finally in Tier 3, students who do not respond sufficiently to both previous tiers of instruction are given a higher intensity intervention, typically in the form of frequent one-on-one tutoring. If students respond to successive tiers of instruction, they are considered not at risk for LD and continue to receive intervention with the aim of integrating them back into general education. However, those who repeatedly do not respond sufficiently to instruction are generally recommended for a special education evaluation as mounting evidence suggests they need more (or different) support than what general education can provide to be successful academically (L. S. Fuchs & Fuchs, 1998). Sometimes, special education referral/evaluation occurs following Tier 2 (with Tier 3 constituting special education); sometimes it occurs after Tier 3. In theory, this framework makes sense. In practice, many questions about its implementation remain. One such question is how to accurately determine who is at risk for LD and therefore, who should enter the higher tiers of supplemental intervention. Because schools have limited resources of time, personnel, and money and because early intervention generally produces better outcomes for struggling students (see discussion in Simmons et al., 2008), determining who does and does not need instruction beyond Tier 1 is critical. Thus, all RTI models incorporate some type of screening process to identify students who are at high risk for developing LD. The purpose of this article is to incorporate what is known about screening into a flexible, yet comprehensive four-step screening system that might

[1]Vanderbilt University, Nashville, TN, USA

Corresponding Author:
Jennifer K. Gilbert, PhD, Peabody College of Education and Human Development, Department of Special Education, Vanderbilt University, 230 Appleton Place, Nashville, TN 37203-5721, USA
Email: Jennifer.k.gilbert@vanderbilt.edu

help school psychologists, school personnel, and others to establish school-specific screening procedures.

Importance of Screening in RTI

For the RTI approach to succeed in early identification and prevention of LD, screening procedures must result in accurate classification of children who are at risk for LD (true positives) and those who are not at risk for LD (true negatives). Failing to identify all true positives would result in some children (false negatives) failing to receive intervention they require. False negatives occur when truly at-risk students score above the cut point on an indicator of reading risk at the beginning of the year. Because these students score above the cut point, they are deemed not at risk and do not receive intervention. The reality, although, is that their score on the screening measure was not a good indicator of risk because these students are actually at risk for later LD. The result is that these truly at-risk students remain in Tier 1 although Tiers 2 or 3 instruction would better serve their academic needs. Conversely, not identifying all true negatives would result in some children (false positives) receiving unnecessary and costly Tier 2 (or higher) intervention. False positives occur when truly not at-risk children score below the cut point on the screening measure. Students who score below the cut point are deemed at risk and are given more intensive intervention. In the case of false positives, the students' relatively low scores on the screening measure are not good indicators of risk because these students are actually not at risk for later LD. The result is that these children are receiving Tier 2 intervention when Tier 1 would have sufficed or Tier 3 instruction when Tier 2 would have sufficed. Thus, to establish an RTI methodology that is effective and efficient for students and schools, a solid screening process is critical.

Useful Screening Indices

Some useful indices for determining the utility of a screening measure include classification accuracy, sensitivity, and specificity. VanDerHeyden (2011) thoughtfully explained that these indices offer maximum utility when comparing similar outcomes, cut scores, and samples, which is the case for individual schools and/or school districts. Sensitivity is the proportion of truly at-risk students who were correctly identified as at risk (true positives) and is calculated using the equation: true positives/(true positives + false negatives). Specificity is the proportion of truly not at-risk students who were correctly identified as not at risk (true negatives) and is calculated using the equation: true negatives/(true negatives + false positives). Classification accuracy is the proportion of true positives and true negatives to the whole sample ([true positives + true negatives]/*N*).

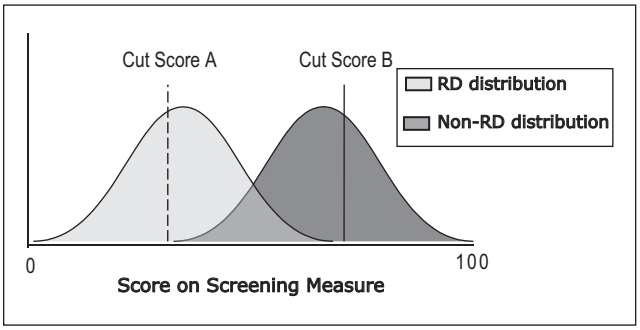

Figure 1. Hypothetical overlapping RD and non-RD distributions with two cut scores.
Note. RD = reading disability.

It would be ideal for sensitivity, specificity, and classification accuracy to all equal 100%, but in reality no screening system achieves 100% classification accuracy, and thus, there are tradeoffs between sensitivity and specificity (Swets, 1988). This occurs because the choice of cut score will inevitably include too many not at-risk children in an attempt to capture all at-risk children (high sensitivity, low specificity), or the opposite will occur where too many at-risk children are excluded in an attempt to also exclude all not at-risk children (low sensitivity, high specificity). Figure 1 depicts this trade-off between sensitivity and specificity. In Figure 1, there are two distributions: one for the population of children with RD and one for the population of children without RD. The figure shows that these two distributions have overlapping scores on the (generic) screening measure. Cut-Point B represents the high sensitivity and low specificity case and Cut-Point A represents the low sensitivity and high specificity case. Unless there is perfect separation between the two distributions on a particular measure, a trade-off will have to be made. With this trade-off in mind, we begin our discussion by using sensitivity, specificity, and classification accuracy to provide an overview of prior work on screening. Although RTI addresses LD in general, the focus for the recommended early screening system in this article is specifically on reading disability (RD).

Past Research on First-Grade Screening

Univariate screening models. Past research has shown that screening for RD using a single measure produces too many false positives to be effective for accurate classification (Johnson, Jenkins, & Petscher, 2010; Johnson, Jenkins, Petscher, & Catts, 2009). Johnson et al. (2009) examined the classification accuracy of four Dynamic Indicators of Basic Early Literacy Skills (DIBELS) indicators (Good & Kaminski, 2002): nonsense word fluency, initial sound fluency, phoneme segmentation fluency, and oral reading

fluency. The authors examined DIBELS indicators because they are popular measures of foundational reading skills, evidenced by the fact that almost 2,000,000 students have been assessed with DIBELS (Samuels, 2007). The authors chose a criterion less than the 40th percentile on the comprehension subtest of the *Stanford Achievement Test-10* (SAT, 10th ed.; Harcourt Brace, 2003) at the end of first grade; this criterion was associated with a 9.90% base rate. End-of-kindergarten and beginning-of-first-grade DIBELS indicators were entered in separate logistic regression models, all predicting end-of-first-grade status on the SAT-10 as either satisfactory or unsatisfactory. The authors reported that none of the DIBELS indicators used as screening tools at the end of kindergarten or beginning of first grade improved classification accuracy over a simple assumption that no students would perform unsatisfactorily on the SAT-10. When the authors held sensitivity of all the indicators at 90%, the resulting specificities (range = 20%–59%) and classification accuracies (range = 41%–77%) were too low to recommend any of the indicators be used as a screening tool. Even entering pairs of indicators in the same model left too many students misclassified; these combinations made only negligible improvements to specificity.

Multivariate screening models. Johnson et al. (2010) found similar results when using a state comprehension achievement test at the end of third grade as an outcome. The utility of a variety of measures was examined to identify their potential to serve as an early RD screener: end-of-second-grade comprehension, end-of-second-grade and beginning-of-third-grade oral reading fluency, end-of-second-grade receptive language, and student status regarding special education and English as a second language. None produced acceptable rates of classification, sensitivity, or specificity. For example, to achieve 90% sensitivity with the indicators, the range of associated specificities was 43% to 58%, meaning that a large number of children who were not at risk were incorrectly classified as being so. The range of classification accuracies revealed similar shortcomings (range = 58%–69%). In an effort to improve classification accuracy, the authors combined all indicators into one model. This resulted in only a 2% increase in accuracy over their most promising single indicator and still left 25% of the sample misclassified.

Other researchers have also attempted to use a multivariate approach to screening. Catts, Fey, Zhang, and Tomblin (2001) relied on several measures of kindergarten reading skill to predict RD at the end of second grade, which was defined as anyone scoring 1 *SD* below the mean on a comprehension composite score. A model that included letter identification, sentence imitation, mother's education, phoneme deletion, and rapid automatized naming of animals produced classification accuracy of 81% and specificity of 79% when the sensitivity was set to 92%. Compton, Fuchs, Fuchs, and Bryant (2006) found similar classification estimates although they used beginning-of-first-grade measures to predict second-grade RD (based on a word reading and comprehension factor score). They reported classification accuracy of 83%, sensitivity of 90%, and specificity of approximately 83% from a model including phonemic awareness, oral vocabulary, rapid digit naming, and slope and intercept of word identification fluency. They justified adding information about response to instruction because an assumption of the RTI approach is that students who do not respond to generally effective instruction have an underlying cognitive deficit that manifests in a RD (Vaughn & Fuchs, 2003). On a related note, Gilbert et al. (2012) found that 6 weeks of progress monitoring eliminated 46% of the sample that was deemed at risk for RD by the screening measures. This represents a tremendous savings on resources that would have been required to conduct further standardized testing or to implement supplemental tutoring for these children who were, in effect, false positives.

In a follow-up to the 2006 study, Compton et al. (2010) found that the 2006 model replicated well on a new sample of 355 first graders, with similar classification accuracy, sensitivity, and specificity as the original sample. The authors extended the original model by showing that adding dynamic assessment (DA) of decoding to a set of core predictors produced the same classification indices as adding progress-monitoring information measured by word identification fluency slope and intercept. Progress-monitoring data and DA data purport to measure student learning potential (L. S. Fuchs et al., 2008). DA differs from more conventional, static assessments in that the former has an assess–instruct/prompt–assess format, whereas the latter has a one-time assessment protocol. Measures of DA are typically scored by indicating the number of instructional prompts the test taker requires to give a correct answer. The higher the number of prompts, the less potential the test taker has to learn from instruction. Progress monitoring also provides an index of response to instruction, but the advantage of DA is that it requires only a one-time assessment, whereas gathering slope and intercept information takes many weeks. O'Connor and Jenkins (1999) also found DA to be a useful screening tool. Their DA measured learning of onset-rime segmentation. They compared a model that included a static measure of onset-rime segmentation with one that included a dynamic measure of the same skill; the other variables remained constant across the two models. The model with the DA increased the number of false negatives by one student, but it reduced the number of false positives from 26 to 9, thereby increasing the specificity from 87% to 96%. It appears that DA is useful in detecting students who may have low static screening scores but have high enough learning potential that their low screening scores might be overcome by their ability to learn from instruction.

Table 1. Procedures, Measures, Advantages, and Disadvantages for Each Step of a Four-Step Screening System

Step	Procedures	Possible Measures	Advantages	Disadvantages
1	Using static screening instrument	Letter identification, oral reading fluency, phoneme segmentation, and word identification	Efficiency, feasibility, and objectivity	Time and money required to administer screener
2	Level 1 and progress monitoring	Fluency in letter identification, passage reading, and word identification	Elimination of false positives	No information on false negatives, more time than administering a one-time screener
3	Level 2 and follow-up testing	Standardized, nationally normed test, and state achievement test	Identification of false negatives and use of logistic regression models	Time and money required to administer criterion measure
4	Level 3 and upgrading procedures for subsequent years	NA	Increased classification accuracy	Requires at least 1 year of data collection

Although using multivariate approaches to screening, including the addition of DA or progress-monitoring information, better classifies students into risk or no-risk classes than any single screening measure alone, obtaining multiple measures on all children is expensive and time-consuming. One alternate way of increasing classification accuracy is to follow Catts et al.'s (2001) recommendation and use a two-stage gated screening procedure. The first step (Gate 1) is to use a screening cut score that will maximize the number of true positives. To ameliorate the unsatisfactorily high number of false positives that comes with using this liberal cut score, the second step (Gate 2) is to conduct further testing on the initial risk sample to gather information for eliminating false positives. Compton et al. used this procedure as a post hoc analysis in the 2010 replication. The authors chose the highest score obtained by a student with later RD on each of the four screening measures. Then, they calculated a 99% confidence interval around that score and eliminated all students who scored above the cut score with the assumption that all true positives have a 99% chance of scoring at or below the cut score. The screening measure that eliminated the highest number of true negatives was a measure of timed decoding. Relying on a liberal cut score on decoding efficiency as the first step in the screening process resulted in a 43% reduction in the number of students who required additional testing. It is important to note that this reduction in testing information did not change the classification accuracy (88%–91%) for the prediction models.

Based on the research just described, it is clear that a one-measure, one-time screen is not sufficient for accurately identifying children who will and will not develop RD. Possible options for improving screening classification include adding static measures, DA, and progress-monitoring data. Again, the purpose of this article is to incorporate what is known about screening into a flexible, yet comprehensive four-step screening system that might help school psychologists, school personnel, and others to establish school-specific screening procedures to identify RD. Our recommendations incorporate advice given by Jenkins, Hudson, and Johnson (2007) to minimize false negatives, work backward from criterion to make cut points, identify most appropriate screening measures to use in a multivariate screening model, and cross-validate procedures with additional samples of students.

In the remaining sections of the article, each step in a four-step screening system is described. First, however, we acknowledge Jenkins et al.'s (2007) point: Each school or school district will have unique challenges for developing its screening system, including a cost–benefit analysis of the types of measures chosen and the number of students entering Tier 2 intervention and beyond. Thus, the system described below should be viewed as a heuristic, which is flexible to accommodate differences in school contexts. Points of flexibility are indicated where appropriate. Furthermore, it is likely that some schools will not be able to implement all steps of the recommended system. To assist school-based personnel strike a balance between the feasibility and value of the components of this system, we will discuss advantages and disadvantages of each step.

Four-Step Screening System

The proposed four-step screening system requires multiple years of data collection, with the 1st year requiring the most intensive data collection. Table 1 contains descriptive information for each step, including procedures, possible measures, advantages, and disadvantages.

Year 1

Step 1: Universal screening. Step 1 of the 1st year of screening takes place while students are in Tier 1 instruction. All students in kindergarten or first grade (whichever the school deems more appropriate) are screened with a variety of relatively quick-to-administer assessments. Suggestions for initial screening measures in kindergarten and

first grade are decoding fluency, letter naming fluency, letter sound fluency, oral vocabulary, phoneme segmentation, nonsense word fluency, and word identification fluency (see Compton et al., 2006; Compton et al., 2010; Jenkins et al., 2007). The National Center for RTI has a more extensive list of screening measures posted as a table on the www.rti4success.org website, and we refer readers to Jenkins et al. (2007) for a more detailed discussion on possible screening measures. We recommend administering a variety of these measures so that schools have options for choosing the most informative screeners based on follow-up testing.

If Step 1 is the only part of the screening system that schools can implement, personnel (e.g., school psychologists) will have to apply a cut score to determine risk for RD at this step. When deciding on a cut score, practitioners must concern themselves with the number of students to whom they are able to provide Tier 2 intervention. This number, the school's base rate of RD risk, is determined partly by school resources. That is, if a school has the capacity to provide 8% of its students with supplemental tutoring, then 8% becomes the maximum proportion of students who can be identified as at risk (i.e., base rate). Using just Step 1 is akin to the direct RTI route, based on a one-time universal screening, in which students who score poorly on a universal screener are placed directly into Tier 2 tutoring (Jenkins et al., 2007).

The advantages of using only Step 1 are efficiency, feasibility, and objectivity. This step is efficient because it requires only a brief, one-time assessment of all students. Because many brief screeners are free of charge, Step 1 is also feasible. Finally, implementing short screeners in Step 1 provides school staff with more objective information with which to make decisions about student placement than simply relying on teacher judgment. Research has shown that teacher ratings of student reading ability are only moderately correlated with objective measures of reading skill and that some teachers are much more accurate at objectively ranking their students than others (r = .32–.99; Madelaine & Wheldall, 2005). Although teacher ratings may be important in the initial identification of students at risk for RD, Step 1 helps eliminate much of the subjectivity in teacher rankings and ensures that the poorest readers are identified for remediation or further testing.

Despite the advantages of Step 1, its disadvantage is a lack of information about false positives and false negatives. That is, students incorrectly identified as at risk by the screening measure in Step 1 (false positives) would receive Tier 2 tutoring that would place unnecessary strain on the school's resources. Although students who qualify as false positives are likely to be near the at-risk cutoff and would certainly benefit from minor accommodations in the classroom, providing Tier 2 tutoring to false positives in addition to true positives would put a strain on school time and monetary resources. However, students who were incorrectly identified as not at risk by the screener (false negatives) would not receive the early remediation they need in the form of Tier 2 tutoring. Most practitioners would see this as unacceptable. More data than those provided in Step 1 are required to correctly classify students who were misidentified (false positives and negatives) at this step. Step 2 provides some of these data.

Step 2: Progress monitoring. Step 2 of the screening system includes universal screening and the monitoring of student progress for the entire year. Implementing this step is similar to using the progress-monitoring route to RTI discussed by Jenkins et al. (2007), whereby students who score poorly on a screening measure are monitored in their current educational setting before entering Tier 2 tutoring. For the 1st year of the screening system, all students (in kindergarten or first grade) are progress monitored. Obtaining progress-monitoring data for all students allows schools to (a) establish norms for growth rates and final intercepts and (b) detect false negatives not detected as at risk by the screening measures. Possible measures for monitoring student progress in the early grades are letter name fluency, initial sound fluency, phoneme segmentation fluency, oral (passage) reading fluency, nonsense word fluency, and word identification fluency (Compton et al., 2006; Compton et al., 2010; Johnson et al., 2009; Johnson et al., 2010). To reduce the amount of time schools must spend on progress monitoring during the 1st year, students could be placed on a biweekly rotation such that half the students are tested in any given week. After the 1st year, these progress measures are used after initial screening for only a subset of students to identify and eliminate false positives from the risk pool. An alternate to collecting progress-monitoring data for several weeks is to administer a one-time, reliable DA. Ideally, progress monitoring and DA are used in a coordinated fashion to capture a student's learning potential. Should a school not have the time or resources to obtain progress-monitoring data, which would involve weekly or biweekly assessments of all at-risk students, DA might be a viable alternative (Compton et al., 2010; O'Connor & Jenkins, 1999). Current options for DA include decoding (D. Fuchs, Compton, Fuchs, Bouton, & Caffrey, 2011), phoneme segmentation (Spector, 1992), and onset-rime segmentation (O'Connor & Jenkins, 1999). However, DAs in phonological awareness (Sittner-Bridges & Catts, 2008) and comprehension (Elleman, Compton, Fuchs, & Fuchs, 2008) are under development.

If schools are only able to implement Step 2, a decision about response to Tier 1 instruction will need to be made to determine who will enter Tier 2 tutoring. We refer readers to L. S. Fuchs and Fuchs (1998) for a discussion on the benefits of using level of performance and rate of growth (i.e., dual discrepancy) rather than either alone to determine response.

The major advantage of using Step 2 over sole reliance on Step 1 is that gathering progress-monitoring data holds the potential to identify and eliminate false positives. Some children who score low on the Step 1 screening measure may fare well once they experience high-quality classroom instruction. Children who show a high rate of growth despite a low screening score are considered to have responded sufficiently to classroom instruction and are therefore not at risk for later RD. Without progress-monitoring data, although, these children would likely be recommended for unnecessary Tier 2 tutoring. Several studies have demonstrated the utility of using progress monitoring or DA to eliminate false positives (Compton et al., 2006; Compton et al., 2010; Elleman et al., 2008; Gilbert et al., 2012; Sittner-Bridges & Catts, 2008), and this elimination of false positives represents a huge savings to schools because it means that only the children who are truly at risk will receive costly supplemental tutoring. The disadvantage of using Step 2 is that it requires more time than using Step 1 alone. Although progress-monitoring measures are fairly quick to administer, several weeks or months of data collection are required to properly assess growth (see discussion by Hintze & Christ, 2004; see www.rti4success.org for resources about gathering, graphing, and analyzing progress-monitoring data.) Furthermore, Step 2 provides no information about false negatives.

Step 3: Follow-up testing. Step 3 of the screening system, which includes Steps 1 and 2, requires follow-up testing. Before administering follow-up test(s), schools must choose a criterion of reading success. The criterion can be based on standardized tests (Catts et al., 2001; Compton et al., 2006; Compton et al., 2010; Johnson et al., 2010) or state achievement tests (Johnson et al., 2009). For the first cohort of students participating in the screening system, it is ideal to have a criterion measure for every student. Thus, it may be most feasible for schools to use state achievement tests as the criterion to avoid the costs associated with purchasing, administering, scoring, and interpreting standardized tests for every student. The results of most state achievement tests indicate whether a student has met proficiency in particular academic areas, and the proficiency status can be used to decide which students require academic remediation. If schools use standardized tests instead, they have to decide on a criterion of reading adequacy. Another important decision about the criterion measure is when to obtain it. If a follow-up test is given only at the end of the initial intervention year, schools will not have information about how well students are reading in later grades. Because sustained reading success is of chief importance, we recommend obtaining follow-up test scores after the 1st year of intervention and beyond. Ultimately, each school will have to decide if the most critical reading outcome for determining RD is at the end of first, second, third, fourth, or even later grades.

The advantages of using Step 3 are gaining information about the classification accuracy of Steps 1 and 2 and identifying false negatives (i.e., students who were not deemed at risk in Steps 1 or 2 but were found to have substantial reading struggles at a later time). At this point, teachers can provide remediation to those students or at least monitor them closely in the next grade. The disadvantage of Step 3 is the time and expense required to administer and interpret criterion measure(s). However, having this information can be very useful during Step 4 when decision criteria are revised to enhance classification accuracy in subsequent years. Another advantage of Step 3 is the ability to use logistic regression models to predict each student's probability of being at risk for later RD based on scores from early in the school year (i.e., screeners and progress-monitoring assessments). In more concrete terms, after school personnel have determined their base rate of RD (based on available resources) and have applied the corresponding cut score to the criterion measure, they can employ logistic regression models to assign each student a predicted probability of actually meeting the criterion for RD. Readers are referred to Peng, Lee, and Ingersoll (2002) for an introduction to logistic regression models. Different models can be fit to determine which combination of screeners and progress-monitoring data best classifies students. At the same time, feasibility of screening can be taken into account; if a model with four screening measures classifies students only slightly better than a model with two, it may be more feasible to use the two-screener model so that only two measures rather than four have to be given in subsequent years.

When a final model has been selected, schools can then adjust indices of sensitivity and specificity for the model to determine the most beneficial probability cutoff to use in subsequent years. For example, in a school of particularly high-performing readers, students may be considered to be at risk for RD if they have even a 30% predicted probability of having a RD. This is because in a population of good readers, even a 30% chance of having a RD is high enough to classify a student as at risk. However, in a school of fairly poor readers, students may have to reach 60% probability to be considered at risk for RD. As mentioned previously, risk is relative to local standards and local resources, and logistic regression provides a method of systematically assigning risk in each specific context. After the prediction model has been set for Year 1, students' scores from screening measures and progress monitoring in subsequent years are entered into the model to determine each student's probability of later RD. See Compton et al. (2010) for a more detailed description of how to replicate logistic regression models across samples of students.

Step 4: Revision of risk-classification criteria. Step 4 of the screening system is considered the gold standard because it includes Steps 1 to 3 and also requires updating decision criteria used in Steps 1 and 2 with the data obtained from

conducting follow-up testing in Step 3. Because grade-level populations change within a school from year to year, cut scores have to be adjusted in perpetuity. Moreover, implementation of new instructional practices will create the need to adjust cut scores. To set cut scores for Step 1, schools determine the highest score obtained by a student with RD on each of the screening measures included in the final prediction model. The highest score for each test becomes the cut score; anyone below the cut score is considered at risk and anyone above is considered not at risk. The screening measure that eliminates the highest number of true negatives (while retaining all the true positives) becomes the first step for determining the initial risk pool in the following year. Scores from the remaining screening measures are retained for use in logistic regression models.

Then to set cut scores for Step 2, slopes and intercepts from the various progress-monitoring measures are calculated for each student. This information gives schools an indication of how their populations are performing and establishes norms for the school. Among the students identified as having a RD by the criterion measure, the highest slope and intercept is recorded and used as a tool for eliminating false positives (students who looked at risk at initial screening but who performed well by the end of the year) in following years. As an alternative, schools establish the highest DA score of any student with RD and use that score as a cut score for eliminating false positives from the risk pool.

The major advantage of Step 4 is increasing classification accuracy in subsequent years. By revising cut scores in Steps 1 to 3 based on the criterion measure, schools have the potential to monitor the progress of fewer students, provide needed remediation to all students who need it most, and withhold remediation from students who will fare well simply by receiving classroom instruction. The disadvantage of this step is that it takes at least 1 year to gather data and revise cut scores.

Subsequent Years

Step 1: Universal screening. All screening measures included in the final prediction model from Year 1 are given to all students in subsequent years. Then, whichever screening measure performed the best in Year 1 with regard to eliminating true negatives is used to eliminate as many true negatives as possible without risking loss of true positives. To eliminate the true negative in subsequent years for universal screening, the cut score from the previous year is applied to the new population of students. Because cut scores will not replicate perfectly from year to year, we recommend that schools retain 10% more students in the risk sample than are identified by the initial screening measure's cut point. The additional 10% helps guard against true positives becoming false negatives in the initial step. In other words, there may be students who score above the cut point on the screening measure that go on to experience substantial difficulty in reading such that their criterion scores indicate RD status. Most likely, these students will have scores that are just above the cut point in Step 1, so keeping 10% of the "borderline" students would allow schools to distinguish true negatives (students who score above the cut point and fare well in reading) from false negatives (students who score above the cut point and fare poorly in reading).

Step 2: Progress monitoring. All students in the initial risk sample are then monitored using weekly assessments. Although 6 weeks of progress monitoring has been used in research settings (e.g., Compton et al., 2006; Compton et al., 2010; Gilbert et al., 2011), schools may monitor for a longer period of time if necessary. Readers are referred to Christ (2006) for a discussion on the relation between weeks of progress-monitoring and measurement error. After data collection, slopes and intercepts for each student are calculated and compared with the previous year's risk cutoff. At this point, schools can eliminate false positives by applying Year 1 cut scores to the new slopes and intercepts. For the remaining risk sample, the Year 1 logistic regression model is applied to scores from screening measures and progress-monitoring data, and each student is assigned a RD probability. Students above the school's probability cut point are then assigned to remediation or Tier 2 tutoring. Weekly assessments of progress are administered during Tier 2 so that a criterion for adequate response to instruction/RTI can be established for at-risk students in subsequent years.

Step 3: Follow-up testing. At the end of the year, all students who were identified as at risk (plus the 10% buffer) are administered on a follow-up test. Students from the first cohort will also be administered on a criterion measure. In fact, follow-up tests are given to all students up to the grade that the school has selected as the grade to determine RD.

Step 4: Revision of risk-classification criteria. Based on the Step 3 follow-up test for the Cohort 1 students, cut scores for Steps 1 and 2 are updated to reflect changes in classification criteria. As we mentioned previously, cut scores may need ongoing adjustment due to major changes in student population, in school resources, and/or in classroom instruction.

Implications for Practice

In this article, we presented a flexible, yet comprehensive four-step screening system to identify children who are at risk for developing later RD. Accurate identification of risk plays a major role in the success of RTI as a method of identifying and remediating academic difficulties associated with RD. Thus, the system we recommend provides schools with a method of objectively identifying academic risk so that early intervention can be implemented for the students who need it most. The core components of this

system are universal screening, progress monitoring, follow-up testing, and ongoing revision of risk-classification criteria. Although it is feasible to implement this system using fewer than all four steps, we do not recommend doing so. Schools have the best chance of accurately identifying students who are and are not at risk for RD if all steps of the system are in place. We hope that the flexibility of the system in terms of measures, cut scores, base rates, and timelines will encourage schools to incorporate the core components regardless of their student populations or amounts of personnel/monetary resources.

Authors' Note

Opinions expressed herein are the authors' and do not necessarily reflect the position of the U.S. Department of Education, and such endorsements should not be inferred.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The development of this article was supported in part by Grant R324GO60036 and R305B040110 from the Institute of Education Sciences, U.S. Department of Education.

References

Bradley, R., Danielson, L., & Hallahan, R. (Eds.). (2002). *Identification of learning disabilities: Research to practice*. Mahwah, NJ: Lawrence Erlbaum.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implications. *Language, Speech, and Hearing Services in Schools, 32*, 38–50. doi:0161-1461/01/3201-0038

Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review, 35*, 128–133. Retrieved from https://login.proxy.library.vanderbilt.edu/login?url=http://proquest.umi.com/pqdweb?did=1293575731&;Fmt=7&clientId=622&RQT=309&VName=PQD

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., & 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*, 327–340. doi:10.1037/a0018448

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*, 394–409. doi:10.1037/0022-0663.98.2.394

Elleman, A., Compton, D. L., Fuchs, D., & Fuchs, L. S. (2008, July). Exploring dynamic assessment as a means of identifying children at risk of developing comprehension difficulties. In D. L. Compton (Chair), *The utility of dynamic assessment in predicting concurrent and future academic performance.* Symposium conducted at the meeting of the Society of the Scientific Study of Reading, Asheville, NC.

Fuchs, D., Compton, D. L., Fuchs, L. S., Bouton, B., & Caffrey, E. (2011). The construct and predictive validity of a dynamic assessment of young children learning to read: Implications for RTI frameworks. *Journal of Learning Disabilities.* Advance online publication. doi:10.1177/0022219411407864

Fuchs, D., & Deshler, D. D. (2007). What we need to know about responsiveness to intervention (and shouldn't be afraid to ask). *Learning Disabilities Research & Practice, 22*, 129–136. doi:10.1111/j.1540-5826.2007.00237.x

Fuchs, L. S., Compton, D. L., Fuchs, D., Hollenbeck, K. N., Craddock, C. F., & Hamlett, C. L. (2008). Dynamic assessment of algebraic learning in predicting third graders' development of mathematical problem solving. *Journal of Educational Psychology, 100*, 829–850. doi:10.1037/a0012657

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice, 13*, 204–219.

Fuchs, L. S., Fuchs, D., & Speece, D. L. (2002). Treatment validity as a unified construct for identifying learning disabilities. *Learning Disability Quarterly, 25*, 33–46.

Gilbert, J. K., Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Barquero, L. A., & Cho, E. (2012). *Testing the efficacy of multilevel supplemental tutoring in a first-grade responsiveness-to-intervention model: A randomized control trial*. Manuscript submitted for publication.

Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Available from http://dibels.uoregon.edu

Harcourt Brace. (2003). *Stanford Achievement Test* (10th ed.) (Tech. Data Rep.). San Antonio, TX: Author.

Hintze, J. M., & Christ, T. J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review, 33*, 204–217.

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*, 582–660. Retrieved from https://login.proxy.library.vanderbilt.edu/login?url=http://proquest.umi.com/pqdweb?did=1410784981&;Fmt=7&clientId=622&RQT=309&VName=PQD

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention, 35*, 131–140. doi:10.1177/1534508409348375

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice, 24*, 174–185. doi:10.1111/j.1540-5826.2009.00291.x

Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability, Development, and Education, 52*, 33–42. doi:10.1080/10349120500071886

National Joint Committee on Learning Disabilities. (2005). Responsiveness to intervention and learning disabilities. *Learning Disabilities Quarterly, 28*, 249–260. Retrieved from http://find.galegroup.com/gtx/infomark.do?&;contentSet=IAC-Documents&type=retrieve&tabID=T002&prodId=AONE&docId=A138582088&source=gale&srcprod=AONE&userGroupName=tel_a_vanderbilt&version=1.0

O'Connor, R. E., & Jenkins, J. R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading, 3*, 159–197.

Peng, C. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research, 96*, 3–14.

Samuels, S. J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency. *Reading Research Quarterly, 42*, 563–566. Retrieved from http://vnweb.hwwilsonweb.com/hww/jumpstart.jhtml?recid=0bc05f7a67b1790e73515a7d2d85093cc6d0631d4ba7eb7b7704756e29a4c65e824f051bc136da50&;fmt=H

Simmons, D. C., Coyne, M. D., Kwok, O., McDonagh, S., Harn, B. A., & Kame'enui, E. J. (2008). Indexing response to intervention: A longitudinal study of reading risk from kindergarten through third grade. *Journal of Learning Disabilities, 41*, 158–173. doi:10.1177/0022219407313587

Sittner-Bridges, M., & Catts, H. (2008, July). Dynamic assessment of phonological awareness. In D. L. Compton (Chair), *The utility of dynamic assessment in predicting concurrent and future academic performance*. Symposium conducted at the meeting of the Society of the Scientific Study of Reading, Asheville, NC.

Spector, J. E. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. *Journal of Educational Psychology, 84*, 353–363. doi:10.1037/0022-0663.84.3.353

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285–1293. Retrieved from http://www.jstor.org/stable/1701052

VanDerHeyden, A. M. (2011). Technical adequacy of response to intervention decisions. *Exceptional Children, 77*, 335–350. Retrieved from http://find.galegroup.com/gtx/infomark.do?&;contentSet=IAC-Documents&type=retrieve&tabID=T002&prodId=AONE&docId=A252289583&source=gale&srcprod=AONE&userGroupName=tel_a_vanderbilt&version=1.0

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research & Practice, 18*, 137–146. doi:10.1111/1540-5826.00070