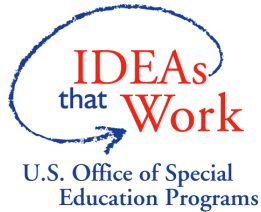




Responsiveness-to-Intervention Symposium

December 4-5, 2003 • Kansas City, Missouri

The National Research Center on Learning Disabilities, a collaborative project of staff at Vanderbilt University and the University of Kansas, sponsored this two-day symposium focusing on responsiveness-to-intervention (RTI) issues.



The symposium was made possible by the support of the U.S. Department of Education Office of Special Education Programs. Renee Bradley, Project Officer. Opinions expressed herein are those of the authors and do not necessarily represent the position of the U.S. Department of Education.

Candidate Measures for Screening At-Risk Students

Joseph R. Jenkins

University of Washington

When citing materials presented during the symposium, please use the following: “Jenkins, J. (2003, December). *Candidate Measures for Screening At-Risk Students*. Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention Symposium, Kansas City, MO.”

Neuropsychological Aspects for Evaluating Learning Disabilities

Most reading researchers have concluded that early identification and intervention beats late identification and intervention. This belief along with the “No Child Left Behind” legislation has focussed public policy on screening and early detection of children likely to encounter reading difficulties. My focus in this paper is on measurements that might be good candidates for screening students requiring secondary intervention (i.e., intervention that is more systematic and intense than generally effective classroom instruction). I approach this task by considering: the school context in which screening occurs; the attributes of an ideal screening system, important ideas to come out of research on screening, and measures that hold promise as screening tools.

Screening in the Context of a Responsive Intervention Model

A Responsive Intervention Model provides timely and correct intervention to every child who requires additional or different instruction from that given in normally effective general education classrooms. Intervention is timely if struggling readers receive help sooner rather than later, and the amount and type of intervention is adjusted *when* needed. Intervention is correct if its content, delivery, and amount are appropriate for and specific to children’s learning needs, and results in improved outcomes (i.e., interventions are effective). Screening is the mechanism for identifying “struggling readers” who lack prerequisite skills or who acquire reading skills at a pace that puts them at risk for unsatisfactory outcomes.

Attributes of an Ideal Screening Mechanism

Ideally, a screening procedure satisfies at least three criteria. First, it must accurately distinguish individuals that require intervention from those who do not. The former are referred to as “at risk” for poor outcomes, the latter as “not at risk” for poor outcomes. In the framework employed by Lichtenstein and Iretton (1984), a screening measure is characterized by its degree of *sensitivity* and *specificity*.

Sensitivity refers to the degree a screening mechanism identifies as “at-risk” individuals who in fact perform unsatisfactorily on a future criterion measure (i.e., “true positives”). *Specificity* refers to the degree a screening mechanism identifies as “not-at-risk” individuals who later performs satisfactorily on a future criterion measure (i.e., “true negatives”).

Sensitivity increases as one type of prediction error (false negatives) decrease. False negatives are individuals that a screening measure classifies not at risk, but who perform poorly on the later criterion measure. In a responsive intervention model, avoiding false negative errors is critical; otherwise individuals most in need of assistance do not receive it.

Specificity increases as another type of prediction error (false positives) decrease. False positives are individuals that a screening mechanism classifies as at risk, but who later perform satisfactorily on the criterion outcome. A screening instrument that produces too many false positive errors wastes precious in-

intervention resources on individuals who do not require them.

The second criterion for an ideal screening mechanism is that it is practical. To attain this criterion a screening procedure must be brief as well as simple enough to be implemented reliably on a wide scale under normal circumstances by normal people. In addition, school personnel must perceive the screening procedures as reasonable. If they perceive screening procedures as onerous, they will not reliably implement them.

A third criterion for an ideal screening system is that the net effect of its implementation is positive. Messick (1980, 1989) refers to this characteristic as consequential validity. Attaining consequential validity means that the screening procedure does no harm—that is, it avoids inequitable treatment; does not consume resources that could be put to better use; and is linked to effective interventions. In other words, screening procedures should be as simple and cost effective as possible for achieving results.

A fourth criterion worth considering (but not essential) is the degree to which screening assessments are integrated with assessments used in other parts of a responsive intervention model. Ideally, assessments used for screening should be linked not only to future criterion outcomes, but also to assessments used for progress monitoring and formative evaluation. Thus, at risk students' responsiveness (or non-responsiveness) to secondary intervention should be apparent on assessments resembling those that identified the students as at risk in the first place. This criterion is particularly relevant to the focus of this conference (Response to Treatment as a model for Learning Disabilities Identification).

Highlights and Important Distinctions from Research on Screening

In this section, I highlight ideas, findings, and important distinction that come out of research on screening.

Grade-Specific vs. Multi-Grade Screening Systems

Researchers have examined the validity of various reading and reading-related measures that hold potential for screening, usually focusing on a specific grade level (e.g. kindergarten). In contrast, two research groups have worked to identify measures that can be used to screen across grades K-2. Good, Simmons, and Kame'enui (2001) developed the screening/progress monitoring system known as the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). Foorman, Fletcher, Francis, Carlson, Chen, Mouzaki, et al. (1998) developed the Texas Primary Reading Inventory (TPRI) which consists of screens and skill profiles. Both these systems have been scaled-up, with multi-district or state-wide implementations.

Satisfactory Reading Ability—How is the Criterion Defined?

Inasmuch as the immediate goal of screening is identifying those at risk for unsatisfactory reading outcomes, screening hinges on the selection of criterion reading measures and performance levels on those measures. Two decisions go into establishing a criterion. The first is deciding on a suitable measure of reading (i.e., *content* standard); the second is deciding the performance level (i.e., *performance* standard) that distinguishes between adequate and inadequate reading skill.

What Criterion Measure?

In the field of early intervention, the family of achievement tests developed by Richard Woodcock and associates come closest to a "gold standard" criterion test of reading ability. These include the

Woodcock-Johnson-Revised (WJ-R) and Woodcock Reading Mastery Test-Revised (WRMT-R) subtests (e.g., Letter-Word Identification, Word Identification, Word Attack, Passage Comprehension, or one or more cluster scales that combine various subtests). The various subtests and clusters assess different aspects of reading. Much of the screening research that follows uses one or more of the Woodcock tests as criterion measures, the studies differ in the tests, subtests, and scales used for their criterion measure of reading.

Choice of criterion measures is critical in evaluating a screen because students performing satisfactorily on one criterion may perform unsatisfactorily on a different criterion measure. For example in Speece et al. (2003), no first-graders read below the 25th percentile on the WJ-R Letter-Word Identification subtest, even though several students performed very poorly (reading fewer than 10 words correct per min) on a test of oral reading fluency. Moreover, the accuracy of a screening measure in predicting different criterion tests may differ. For example, when Speece et al.'s. (2003) used low ORF as the criterion for unsatisfactory reading at the end-of-grade 1, two screens (Non-sense Word Fluency and Letter Naming Fluency) demonstrated strong sensitivity (86%) and specificity (81%-88%) in identifying at risk kindergartners. However, the same screening measures were only 50% sensitive in identifying poor readers (i.e. missing half) when WJ-R Word Attack test was the criterion measure. Screens that are well linked to one criterion measure may not be well linked to another criterion. The moral is--choose criterion measures carefully.

Distal and Proximal Criterion Measures. Most states define reading proficiency according to a standards-based test adopted by that state. Typically, such tests are not given until third or fourth grade, too distant to serve

as criterion measures for early (e.g., kindergarten and first grade) screens. Thus, researchers must use more proximal criterion measures to evaluate the accuracy of screens. For example, to validate the TPRI kindergarten and beginning grade 1 screens, researchers used ending grade 1 achievement on the WJ-R Basic and Broad Reading Scales. To validate the beginning grade 2 screen, TPRI researchers used ending grade 2 WJ-R Broad Reading Scale.

For the DIBELS, Good, et al. (2001) used a different strategy in selecting proximal criteria to judge screening accuracy. Rather than linking DIBELS screens to external measures such as the WJ-R Scales, they linked them to subsequent DIBELS, CBM, and state tests. Specifically, Initial Sound Fluency (ISF) at mid-kindergarten is linked to Phoneme Segmentation Fluency (PSF) at end-of-kindergarten, which is linked to Non-sense Word Fluency (NWF) at mid grade 1, which is linked to Curriculum Based Measurement-Oral Reading Fluency (CBM-ORF) at end-of-grade 1, and so on until CBM-ORF is linked to the Oregon Standards Assessment at end-of-grade 3.

At-risk and for what—Unsatisfactory or Very Unsatisfactory Reading Outcomes?

Two types of performance standards have been used in research on screening resulting in screening students who will demonstrate either: (1) unsatisfactory reading, or (2) very unsatisfactory reading.

Unsatisfactory Reading as a Criterion. Most screens focus on predicting unsatisfactory reading outcomes, where unsatisfactory is defined as performing below a standard (e.g., performing more than one-half year below grade level, performing below a “high standard” like those used by some states in a standards-based reading test), Both the DIBELS and TPRI define unsatisfactory reading based on criterion levels that result in

a fairly large proportion of the population so classified.

Using state-mandated, standards-based reading tests to define unsatisfactory reading is on the increase. States use different performance standards to define satisfactory reading ability. This means that satisfactory reading in one state is not necessarily satisfactory reading in another. For example, Washington State sets a fairly high standard for passing its reading criterion, such that one-third of fourth-graders perform unsatisfactorily (i.e., do not meet the standard). Other states set lower standards (e.g., Texas, where fewer than 2% fail the state competency test in grade 3).

Another common convention for defining unsatisfactory reading is norm-referenced performance falling at or below the 25th percentile. This criterion also can result in designating a substantial proportion of the population as “unsatisfactory readers” depending upon how a school performs in relation to the national norm group.

Very Unsatisfactory Reading as the Criterion. Only the very lowest readers--those considered to have a reading or learning disability—qualify for this designation. O’Connor and Jenkins (1999) and Speece and Case’s (2001) screening efforts were focussed on identifying this group (i.e., performance below the 10th percentile).

In comparing the results of different screening procedures it is important to ascertain whether the screen seeks to predict unsatisfactory or very unsatisfactory reading skill. The proportion of students identified at risk by a screen depends on this distinction. To illustrate, the TPRI which focuses on identifying *unsatisfactory* readers (those ending first-grade performing below the 23rd percentile on WJ-R Basic Reading Scale) identified 56% of mid-kindergartners as at risk (Foorman et al., 1998). In contrast, O’Connor & Jenkins (1999) focusing on *very unsatisfactory* readers (those ending first-

grade below the 9th percentile on the WRMT-R) identified only 18% of mid-kindergartners as at risk. Depending on the performance standard used as the outcome (i.e., unsatisfactory or very unsatisfactory reading levels) the screening procedure will identify very different proportions of students as at risk.

How At Risk?

Screening approaches also differ in the degree of risk they specify. For example, screens can classify individual as *at risk* for not meeting a standard (their screening score does not strongly presage meeting a later criterion), or as *very at risk* for not meeting standard. The DIBELS system distinguishes between students at risk and students very at risk for unsatisfactory reading outcomes. As Table 1 shows, DIBELS uses different measures to identify risk status at different grades (e.g., ISF at mid-kindergarten; PSF at spring of kindergarten). Students who achieve a “benchmark” score on the various screens are very likely to attain the future criterion standard. Students performing below the benchmarks are not assured of achieving the future criterion of satisfactory reading; by implication, they are at risk. In addition, DIBELS specifies performance levels below which students are *very* unlikely to achieve the future standard. The latter group could be considered “very-at risk” because their likelihood of attaining the future criterion is slim.

Table 1 illustrates the different proportion of students that fell into the two risk levels for the various DIBELS measures, as reported by Good et al. (2001). At mid-kindergarten, DIBELS ISF identified 47% of students as at risk. In contrast, only 7% qualified as *very* at risk (i.e., fell below the “needs intensive service score of 10 ISF). Across DIBELS measures, on average 51% of the sample were classified as at risk vs. an average of 7% as very at risk. Also noteworthy was the considerable variability in the percent of students classified as at risk (33-71%)

or very at risk (3-20%) by the different measures.

In summary, depending on the goal of screening (which dictates the emphasis given to particular levels of risk and criterion performance), students can be variously classified: (1) at risk, or (2) very at risk for (3) unsatisfactory, or (4) very unsatisfactory reading outcomes. It is critical to distinguish between these different emphases when comparing screening approaches. The different conventions used for selecting criterion tests and performance levels make it difficult to compare the results from different screening measures.

Measurement Content of Screening Tests

Reading development follows a series of predictable stages (Chall, 1996; Ehri, 1992; 1998) with successive stages emphasizing different skills. Thus, the specific traits that forecast later reading success vary according to children's reading development. For screening measures to be useful they must be sensitive to the skills that pertain at successive stages and grade-levels. In kindergarten, phonemic awareness, letter knowledge, graphophonemic knowledge (letter-sound knowledge), and vocabulary are important building blocks of reading development.

In first grade, children continue to develop phonemic awareness, graphophonemic skill, and vocabulary, but the greatest growth occurs in phonemic spelling, decoding, word identification, and text reading. By second and third grade, reading growth is reflected in the number and type of words individuals can read, the difficulty of texts they can read and comprehend, and the fluency with which these tasks are accomplished. As grade level increases, comprehension of more difficult texts becomes the primary measure of reading development. Screening measures cannot adequately mark individual differences in reading development unless they are sensitive to the different reading skills emphasized at different grade levels. Studies confirm that screening measures valid at one grade are in-

Jenkins, J. (2003, December). *Candidate Measures for Screening At-Risk Students*. Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention Symposium, Kansas City, MO.

valid at other grade levels (O'Connor & Jenkins, 1999; Foorman et al., 1998).

Good and Poor Candidates for Screening Content. Research shows that some traits are better than others in forecasting reading success. This is important in weighing the merits of various candidates for the content of screening instruments. Perceptual and motor skills, once considered useful in identifying poor readers, do not appear to hold much promise for screening. In addition, general measures of receptive and expressive language ability do not seem to target very precisely children who will have difficulty acquiring beginning reading skills (Schatschneider, Fletcher, Francis, Carlson, & Foorman, in press). General language measures may be more useful in identifying students who will have difficulty in more advanced stages of reading acquisition where the emphasis is on reading comprehension.

Accuracy vs. Fluency Measures. Two types of performance have been used in screening: some emphasizing accuracy and some emphasizing fluency. Accuracy measures distinguish children according to number or percent of correct responses on tasks (number or percent of words segmented into phonemes, words identified), whereas fluency measures distinguish children according to the number of correct responses per minute. Accuracy measures reveal individual differences in knowledge; fluency reveals individual differences both in knowledge and speed of processing. O'Connor and Jenkins (1999) combined accuracy (phoneme segmentation) and fluency (rapid letter naming) in their screens, but most studies focus on accuracy (e.g., Foorman et al. 1998) or fluency (Good et al., 2001; Speece et al. 2003).

Combined Measures Beat Single Measure.

Another finding that emerges from research on screening is that approaches that combine several measures have better screening accuracy than approaches based on

single measures. Both Foorman et al. (1998) and O'Connor and Jenkins (1999) found improve classification accuracy from combining scores on several measures.

Validity—Types of Evidence

An efficient screening procedure is one that obtains significant information about many individuals in very little time. The efficiency requirement leads screening researcher to identify reading-related traits, design brief assessments of those traits, and evaluate their potential utility as screens. Screening researchers report two types of validity evidence: criterion validity and classification accuracy.

Criterion Validity. Studies of criterion validity correlate performance on the candidate screening measure with performance on an established reading measure, administered either concurrently and/or at a future time. The strength of the correlation between the candidate measure and established the measure provides evidence on the new measure's validity. But simply documenting a relationship between a screening measure and later reading ability, or even documenting that a particular trait accounts for unique variance in reading ability provides at best weak evidence regarding the utility of a measure for screening purposes. In fact, correlations between screening measures and criterion tests need not be particularly high as long as the screening measure distinguishes between individuals who perform poorly from those who perform satisfactorily on the criterion measure. Including more challenging measurement content (e.g. word-level or text-level tasks) could serve to distinguish between middle, strong, and very strong readers on a future criterion test (thereby increasing the correlation between the new and the criterion measure), but unless the challenging items are particularly sensi-

tive to individual differences in the low and middle skill ranges (i.e., distinguish between poor and satisfactory performers), including them is not an improvement.

Classification Validity. The strongest evidence for the validity of a screening measure derives from classification analyses. These analysis determine the accuracy (sensitivity and specificity) of a screen in distinguishing between students who perform satisfactorily on a future criterion measure from those who do not. Criterion validity studies are informative for identifying measures that hold potential as screens, but classification studies, however, are the sine qua non of screening research.

Speece's et al. (2003) illustrates the relative value of classification analysis over correlational or regression approaches in evaluating the measurement content of screening tests. They examined the utility of several potential end-of-kindergarten screening measures in relation to several end-of-Grade 1 criterion measures, including Woodcock-Johnson (WJ) Letter-Word Identification (Word ID), WJ Word Attack (WA), or oral reading fluency. Kindergarten phonemic awareness (phonemic blending and phonemic elision) accounted for significant variance in all three criterion measures, but its sensitivity was poor, ranging from 43% to 67%. In the best case, phonemic awareness failed to identify 33% of students who performed below criterion at the end of first grade (i.e., false negatives).

Cut-Points and Cross-Validation.

Designing effective screening tools is an empirical process. The task of the researcher is to determine whether cut-points can be found on the screen that clearly distinguish between satisfactory and unsatisfactory outcomes on the criterion measure. It is accomplished by working backward from the criterion measure to the screening measure. There

is always a trade-off between sensitivity (reducing false negatives) and specificity (reducing false positives). Because intervention researchers place a greater premium on sensitivity than on specificity, they select cut-points to limit false negatives. This was the strategy used by Foorman et al. (1998) and O'Connor and Jenkins (1999). The post hoc nature of the process guarantees acceptable sensitivity. Problems arise when attaining sensitivity produces unacceptable specificity (i.e., too many false positives, or over-identification). Because selecting cut points in a post-hoc fashion in effect guarantees desirable levels of sensitivity, this approach usually produces higher sensitivity levels than approaches where cut-point are selected more arbitrarily (e.g., using the 25th percentile of the screening measure to divide risk status). Speece and Case (2001) and Speece et al. (2003) employed the latter approach. These different approaches to marking cut-points must be considered when comparing the sensitivity of different screening measures.

The danger in generalizing sensitivity and specificity results from studies using a post-hoc procedure for setting cut-points is that the cut-points may not hold up in a cross validation. When O'Connor and Jenkins (1999) applied cut-points derived from their first cohort to subsequent cohorts, they did not achieve the same level of sensitivity as they had obtained with their first cohort. Caution is warranted in using cut-scores that have not been cross-validated.

Measures that Hold Promise for Screening

In this section, I review measures that have been used for screening and/or that hold promise as candidates for screening, based on evidence from classification or criterion validity studies. I confine my review to child performance assessments, leaving out assess-

ments that are based on teachers' ratings of children.

Tables 2-5 identify several candidate-screening measures and the evidence that supports them. Measures are shown according to the grade level where they pertain: Early and mid-kindergarten; late kindergarten; early grade 1; late grade 1 and early grade 2.

Early- and Mid-Kindergarten Screening Measures

Table 2 shows a selection of measures used in early and mid-kindergarten to predict poor reading outcomes at end-of-grade 1. These measures may be useful to schools that provide early intervention during the kindergarten year.

The table illustrates several of the study differences that I have already mentioned. First, most researchers used some version of the Woodcock end-of-grade tests for the grade 1 criterion, but they focused on different subtests and scales (e.g., WJ-R Readiness Cluster, Letter-Word ID, Broad Reading; WRMT-R Basic Reading Cluster). Second, the three classification studies (Foorman et al, 1998; O'Connor & Jenkins, 1999) employed different criterion performance levels (i.e., unsatisfactory reading--performance below the 23 percentile; very unsatisfactory reading--performance below the 8th percentile; teacher-judgment of severe reading difficulties). This in turn affected the percent of kindergartners identified as at risk by the researchers (56% -18%).

Third, all the kindergarten measures assessed some aspect of phonological awareness (e.g., blending onset-rimes) and/or letter knowledge (e.g., naming letters and/or their sounds). Fourth, some kindergarten measures assessed knowledge (e.g., ability to name letters), others assessed fluency (e.g., speed with which letters are named).

Fifth, two classification studies ascertained that a multiple assessments in combination worked better than single assessments

(e.g. letter naming) as screens. Specifically, Foorman, et al., (1998) and O'Connor & Jenkins (1999) both reported that screening with a combination of phonological awareness and letter knowledge measures produced high sensitivity (i.e., the screen correctly identified 95% to 100% of the students who later performed at an unsatisfactory level on the criterion measure). Results for specificity were also positive, but more variable (the screen correctly identified 56% - 87% of those who later performed at a satisfactory level on the criterion measure).

Sixth, two classification studies (Foorman et al., 1998; O'Connor & Jenkins, 1999) identified screening cut-points in a post-hoc fashion, selecting cut score on the screen to minimize false negatives, and one (Scanlon & Vellutino, 1996) identified several screening cut-points and their classification accuracy. O'Connor and Jenkins demonstrated the importance of cross-validation to ensure that cut-points appropriate for one sample apply as well to other samples. Foorman et al., selected cut-points based on statistically estimated generalizability.

What is *not* shown in Table 2 are the measures that the researchers eliminated as screening candidates. For example, before arriving at their set of screening measures O'Connor and Jenkins' (1999) eliminated PPVT-R, rhyming, blending syllables, segmenting syllables, blending phonemes, and isolating the first sound in spoken words because these measures did not discriminate risk status as well as those that were selected for the screen. Likewise, in arriving at their set of screening measures Foorman et al. (1998) eliminated phoneme blending, comparing first sounds in spoken words, phoneme elision, sound categorization, phoneme segmentation, phonological memory, PPVT-R, rapid naming, visual-motor integration, and visual perceptual matching. It is interesting that some measures eliminated by one research group (e.g. phonological memory eliminated by

Forman et al.) strongly resembled those chosen by another research group (sound repetition chosen by O'Connor and Jenkins, 1999).

End-of-Kindergarten Screening Measures

Table 3 provides classification and criterion validity results for several end-of-kindergarten screening measures used to predict end-of-grade 1 reading. As before, researchers employed different performance levels to divide unsatisfactory from satisfactory reading levels on the criterion reading measure, which in turn affected the proportion of the school population designated at-risk by the screen. Also as before, some researcher identified risk-groups in a post-hoc fashion by linking performance on the criterion test to cut-scores on the screening test. In contrast, Speece et al. (2003) arbitrarily designated the bottom 25% of their sample as at risk.

Sensitivity levels attained by candidate screening measures varied depending on the criterion measure used to measure reading outcomes. Letter Naming Fluency (LNF) and DIBELS NWF showed poor (50%) sensitivity when WJ-R Word Attack was the criterion measure, but good sensitivity (88%) when CBM-ORF was the criterion measure (Speece et al., 2003).

The highest sensitivity levels (93% and 100%) were reported by Foorman et al. (1998) and O'Connor and Jenkins (1999), respectively. The former used a combination of Letter Name/Sound Knowledge (LN/S) and blending onset-rimes in screening, and the latter used a combination of phoneme segmentation, LNF, and sound repetition. However in evaluating the sensitivity achieved by the various screens, it is important to note that both Foorman et al. and O'Connor and Jenkins established screening cut-scores by working backward from the criterion measure--a way to guarantee reasonable sensitivity. Their specificity ranged from 63% to 87%.

Beginning First-Grade Screens.

Table 4 shows results of screening conducted in first-grade to predict ending first-grade performance. Findings from the different studies are hard to compare because of design differences—e.g., the definition of unsatisfactory reading outcomes (i.e. reading disabled vs. reading skills in the lowest quartile), the method for determining screening cut-points (post hoc vs. arbitrary cuts), and the proportion of the population considered at risk (i.e., ranging from 17% to 48%). As before, higher sensitivity was reported for screens that combined phonological and letter knowledge measures and that used post hoc procedures to identify cut-scores on the screening instrument. Based on the classification studies, the best candidates for screening measures appear to be a combination of LNF, phoneme blending, and sound repetition (O'Connor & Jenkins, 1999) and a combination of phoneme blending, LN/S, and word reading (Foorman et al., 1998).

The addition of word identification to the TPRI screen is notable (Foorman, et al., 1998), because it is the first time that word reading emerges as a sufficiently sensitive measure for discriminating risk groups. Fuchs et al. (2003) criterion validity study of at risk first-graders (Table 4) also bolsters the case for using word identification to screen. Their Word Identification Fluency measure (CBM-WIF) produced strong concurrent and predictive validity coefficients with WRMT-R Word Identification and Word Attack. Remarkably, fall CBM-WIF was a stronger predictor of spring WRMT-R Word ID than was fall WRMT-R Word ID (.63 vs. .49). Validity coefficients of the CBM-WIF surpassed those of DIBELS NWF on several measures for the same group of first-graders, suggesting that word identification fluency may be a better screening measure than either WRMT-R or DIBELS NWF for marking risk status of first graders. A comparative classification study could settle this question.

Jenkins, J. (2003, December). *Candidate Measures for Screening At-Risk Students*. Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention Symposium, Kansas City, MO.

End of First-Grade and Beginning Second-Grade Screening

Table 5 shows the results of two classification analyses, along with selected concurrent and predictive validity results for this age group. The TPRI screens employ brief word reading tests at the end of first-grade and beginning of second-grade. The ending first-grade screen also includes a measure of phoneme blending. As with the other TPRI screens, this was designed to produce high sensitivity (above 90%). Not surprisingly, specificity was higher for the beginning grade 2 screen (85%) than for the ending grade 1 screen (77%).

In their classification analysis, Speece and Case (2001) used CBM-ORF to screen beginning second-graders. Students who scored in the lowest quartile on the screen were designated at risk. For the end-of-grade 2 performance criterion, Speece and Case used “very unsatisfactory” reading, defined as a dual discrepancy (level and slope) on CBM-ORF. The CBM-ORF screen achieved a sensitivity of 77% against this criterion, and its specificity was 80%.

Table 5 also shows criterion validity studies of several candidate-screening measures, all of which assessed some aspect of fluency (CBM-ORF, CBM-WIF, DIBELS NWF). In the only comparative study of concurrent validity, Fuchs et al. (2003) reported stronger end of first-grade concurrent validity coefficients for CBM-WIF than for DIBELS NWF for an at risk sample of students.

Beyond Grade 2

Annual Achievement Tests as Screens. There are surprisingly few studies of screening measures beyond grade 2. Where it is possible, results from district- or state-wide annual achievement tests can be used to identify at risk students. This practice should result in reasonably good predictions, given that spring-spring and fall-spring achievement correlations are typically strong. For exam-

ple, Jenkins & Jewell (1992) reported a correlation of .89 between Gates-McGinitie Reading Test (fall) and Metropolitan Achievement Test-Reading (spring).

CBM-Maze. Espin, Deno, Maryuma, and Cohen (1989) developed the Basic Academic Skills Samples (BASS) which includes a maze reading test designed to screen students at risk for failure. For the BASS, students read 3 passages (Spache readability = 2.3) with every 7th word deleted. For the missing words, students attempt to select the correct word from among 3 choices. Distracters are clearly incongruous with the context of the story. Each passage is scored for the number of words correctly restored in one min. The student's score is the median of the score of the three passages. For a sample of grade 3-5 students, Espin et al. reported a concurrent validity coefficient of .81 for the BASS-Maze, using CBM-ORF as the criterion test. Jenkins and Jewell (1992) examined the concurrent validity of the BASS Maze and its degree of overlap with the Gates-McGinitie Reading Test and the Metropolitan Achievement Test--Reading. Concurrent validity ranged from .63 to .78, depending on grade level and the achievement test used as the criterion (Table 6). On average there was approximately 60% overlap between the bottom 15% of students on BASS and the bottom scoring 15% on the achievement tests. Unfortunately, Jenkins and Jewell did not report sensitivity and specificity levels. Research is needed on the classification accuracy of CBM-Maze screens.

CBM-ORF. Although studies too numerous to report have documented strong criterion validity for CBM-ORF, there is little information on the classification accuracy that can be achieved using ORF cut-points. In one of the few screening classification studies, Stage and Jacobson (2001) used ORF to screen beginning fourth-graders for unsatisfactory performance on the state-mandated

standards based reading test. Using an ORF fall cut-score of 100 correct words, they reported low sensitivity and specificity (66% and 76%, respectively). In an analysis using an ORF of 50 words correct (the marker for the lowest scoring 10% of the sample), Stage and Jacobson reported sensitivity of 31% and specificity of 96%. More research is needed to determine if CBM-ORF can produce reasonable specificity levels once sensitivity requirements are satisfied.

Conclusion

Based on this selective review, there are plenty of good *candidates* for screening measures. However, if we are to deepen our knowledge about accurate screening, we need to establish that the candidate measures produce satisfactory sensitivity and specificity in classifying students into risk or risk categories. This will require a certain kind of research on the candidate measures. Specifically, we need studies that adopt the following research protocol.

1. Identify a criterion measure and the performance level (score) required for satisfactory reading.
2. Choose 1-3 candidate screening measures (possibly from those mentioned in this review) that are: correlated with the criterion measure, relatively brief, and can be reliably administered. In choosing these candidate measures, give priority to those that could also be used for subsequent progress monitoring.
3. Decide on an acceptable level of false negatives (e.g., 5-10%) or sensitivity (90-95%)
4. Administer the candidate measures during the desired screening period to a large sample of representative students.
5. Follow the students and administer the criterion measure.
6. Identify cut-scores on the candidate measures that net 90-95% (depending on the decision in step 3) of the students that failed to attain the criterion score.

7. Choose the screening measure that produces the highest specificity (the fewest false positives).

8. Determine if the sensitivity and specificity can be improved by combining the results of more than one candidate measure.

9. Cross-validate the cut-scores, sensitivity, and specificity either by (a) repeating the procedure with another group or (b) applying the cut scores to a subsample that you reserved from the larger sample.

Currently research using this protocol is in short supply. In the meantime, practitioners can rely on this review for selecting the measures that hold the most promise. Whether we can settle on *one best approach* for screening is another matter. Local preferences for criterion measures, criterion performance levels, and tolerance for under- and over-identification rates will lead to different choices for screening. What is critical at the local level, however, is ensuring that screens select all or nearly all students who require secondary intervention and progress monitoring. Only then can “Response to Treatment” be implemented to identify students with learning disabilities.

References

- Chall, J.S. (1996). *Stages of reading development*. (2nd ed.). Fort Worth, TX: Harcourt-Brace.
- Ehri, L.C. (1992). Reconceptualizing the development of sight word reading and its relationship to recoding. In P.B. Gough, L.C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 107-143). Hillsdale, NJ: Erlbaum.
- Ehri, L.C. (1998). Grapheme-phoneme knowledge is essential for learning to read words in English. In J.L. Metsala & L.C. Ehri (Eds.), *Word recognition in beginning literacy* (pp.3-40). Mahwah, NJ: Lawrence Erlbaum Associates.
- Foorman, B. R., Fletcher, J. M., Francis, D. J., Carlson, C. D., Chen, D., Mouzaki, A., Schatschneider, C., Wristers, K., & Taylor, R. (1998). *Technical Report Texas Primary Reading Inventory Technical (1998 Edition)*. Houston, TX: Center for Academic and Reading Skills and University of Houston.
- Foorman, B. R., Francis, D.J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk-children. *Journal of Educational Psychology*, *90*, 37-55.
- Fuchs, L. M., Fuchs, D., Compton, D. (2003).
- Jenkins, J. R. & O'Connor, R. E. (2002). Early identification and intervention for young children with reading/learning disabilities. In R. Bradley, L. Danielson, and D. P. Hallahan (Eds.), *Identification of Learning Disabilities: Research to Practice* (pp. 99-150). Mahwah, NJ: Erlbaum.
- Jenkins, J. R., & Jewell, M. (1992). An examination of the concurrent validity of the Basic Academic Skills Samples (BASS). *Diagnostique*, *17*(4), 273-288.
- Lichtenstein, R., & Ireton, H. (1984). *Pre-school screening and identification of young children with developmental and educational problems*. Orlando, FL: Grune & Straton.
- O'Connor, R. E., & Jenkins, J. R. (1999). The prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading*, *3*, 159-197.
- Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75-107). Timonium, MD: York Press
- Schatschneider, C., Fletcher, J.M., Francis, D.J., Carlson, C, & Foorman, B.R. (In press). Kindergarten predictors of reading skills: A longitudinal Comparative Analysis. *Journal of Educational Psychology*.
- Speece, D., & Case, L. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology*, *93*, 735-749.
- Speece, D., Mills, C., Ritchey, K., & Hillman, E. (2003) Initial evidence that letter fluency tasks are valid indicators of early reading skill. *Journal of Special Education*, *36*, 223-233.
- Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Chen, R., Pratt, A., & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology*, *88*, 601-638.

Jenkins, J. (2003, December). *Candidate Measures for Screening At-Risk Students*. Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention Symposium, Kansas City, MO.