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.

Hitting the Moving Target Known as Reading Development: Some Thoughts on Screening First Grade Children for Secondary Interventions

Deborah L. Speece

University of Maryland

When citing materials presented during the symposium, please use the following: “Speece, D. L. (2003, December). *Hitting the Moving Target Known as Reading Development: Some Thoughts on Screening First Grade Children for Secondary Interventions.* Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention Symposium, Kansas City, MO.”

Hitting the Moving Target Known as Reading Development: Some Thoughts on Screening First Grade Children for Secondary Interventions

Attempts to define early identification procedures have a long history in educational and psychological measurement. The task, of course, is to develop measures and procedures that successfully identify young children who will later experience reading problems. Despite the amount of effort devoted to this goal, a satisfactory solution has yet to emerge. This situation is particularly ironic in that we have available validated instructional practices that can assist many young children in their efforts to enter the world of literacy (e.g., Foorman & Torgesen, 2001; Fuchs et al., 2001; Torgesen et al., 1999). However, the instructional practices that are likely to help those who struggle the most are intensive and expensive (e.g., Torgesen et al. 2001; Vellutino et al., 1996). Thus, it behooves us to develop screening procedures that will allow scarce instructional resources to be targeted at those children who will otherwise experience reading failure.

The challenges of early identification are well documented (e.g., Fletcher & Satz, 1984; Jenkins & O'Connor, 2002; O'Connor & Jenkins, 1999; Scarborough, 1998). Measures that enjoy strong correlations with reading often fail as classification/screening tasks, yielding too many classification errors. There are two types of screening errors: over- and under-identification. Over identification means the screening procedure identified too many children as at risk for poor outcomes--children were identified who did not experience reading problems. Under identification is just the opposite--the screen missed children who later experienced poor outcomes. Which error is more egregious depends upon one's perspective. Over identification means children who don't need intervention will receive it; under identification means children who need intervention will not receive it.

Ideally we could devise a system that minimizes each error type, recognizing their reciprocal relationship--as one increases the other decreases. Some investigators focus efforts on minimizing under identification and set screening criteria so that no child with a poor outcome will be missed. Scanlon and Vellutino (1996) adjusted their kindergarten screening criterion for letter names from 10 to 20 and the number of children misidentified as poor readers in first grade more than tripled. O'Connor and Jenkins (1999) similarly set their first grade criteria so that no child was missed but the over identification rate (i.e., false positives; FP/FP + TP) ranged from 47 to 70%. Essentially, one must pick her poison.

What we are faced with in screening is developing methods that will hit a moving target. That is, children continue to develop on the very skills we use as screens but our methods rarely take this development into account. Scarborough (1998) illustrated this phenomenon by showing that kindergarten children who have poor phonological

Speece, D. L. (2003, December). *Hitting the Moving Target Known as Reading Development: Some Thoughts on Screening First Grade Children for Secondary Interventions.* Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention Symposium, Kansas City, MO.

awareness skills may or may not have difficulty getting into reading because phonological skills are quite learnable. Some children with poor initial skills will "get it", others won't, and we don't know how to discriminate these two groups of children using the extant one-step screening procedures. Phillips et al. (2002) provided another example of how slippery development can be. These investigators recently rebutted Juel's (1988) findings of reading status immutability by showing that the probability of being a poor reader in both first and sixth grades is no more than .50 as compared to .88 between first and third grades reported by Juel. Almost half the children who were below average in first grade were in the average achieving group by sixth grade, presumably due to usual exposure to the curriculum and instruction.

The point to be made is that children are not passive members of inert classrooms waiting for the next measurement occasion. This is so despite the fact that most attempts at screening/early identification assume this to be the case either implicitly or explicitly. One reason that screening efforts have not achieved an acceptable degree of accuracy may be the failure to attend to growth. There is accumulating evidence that measures of learning (i.e., growth) may be key to early identification efforts (Byrne, Fielding-Barnsley, & Ashley, 2000; Deno, Fuchs, Marston, & Shin, 2001; O'Connor & Jenkins, 1999; Speece & Case, 2001; Speece & Cooper, 1990). Byrne et al. (2000) reported that the number of phonological awareness training sessions needed by preschool children to demonstrate perfect performance differentiated disabled and nondisabled readers in elementary school and contributed significant unique variance (8% to 21%) to fifth grade literacy performance beyond the contribution of phonological awareness. Deno et al. (2001) found that first grade students in general education demonstrated over twice the growth in oral reading fluency compared to their counterparts in special education and that this discrepancy held when beginning reading levels (intercepts) were controlled. This evidence suggests that, in addition to level of performance, measurement should recognize growth.

The dual focus on level and growth is the cornerstone of RTI conceptualizations proposed by Fuchs and Fuchs and their colleagues (e.g., Fuchs, 1995; Fuchs & Fuchs,1998; Fuchs, 2003). In their model, level and growth measures derived from weekly assessments on Curriculum-Based Measures (CBM) provide both screening and diagnostic information. Children who are below classmates or grade mates on both level and slope (Dually Discrepant) are initially considered at risk for poor academic outcomes. Following intervention phases in general education, children who continue to be nonresponsive are candidates for more intensive interventions (e.g., secondary interventions). Each of the screening and intervention phases typically last 6 to 8 weeks.

The work I have conducted with this model has focused primarily on the validation of the dual discrepancy criteria as a diagnostic (outcome) classification. In the domain of reading we found the classification to be valid with respect to construct and social consequential validity (Speece & Case, 2001). An important finding was that the DD classification was not biased on gender or ethnicity. We also found that children who were frequently identified as DD across three years, compared to other at-risk children, exhibited poorer performance on reading and reading-related measures, were rated lower by their teachers on academic competence, problem behaviors, and social skills, and were identified more frequently by school personnel as requiring assistance or attention beyond that provided in the general education classroom (Case, Speece, & Molloy, 2003)

These findings and others (Fuchs & Fuchs, 1998; Fuchs, 2003) provide support for the use of the dual discrepancy procedure as a diagnostic category. We have not investigated the accuracy of the initial DD classification as a screen. Placed in the context of one-shot screening approaches, the DD approach to screening may be viewed as a daunting undertaking: All children would be administered a CBM probe once a week for 8 weeks so that slope and level performances could be assigned to each child. These data are used to make a screening designation as at risk or not at risk. Although this screening procedure satisfies the criterion of acknowledging development by assessing children's responsiveness to the general education curriculum, a reasonable question is whether there is a more efficient method. For example, perhaps monthly screening across three or four months would produce either gain scores or slope estimates that are sensitive to responsiveness. The screen may take a month longer but require fewer resources to implement.

Perhaps efficiency is not the right question at the present time. Perhaps the right question is whether this screen yields accurate classification. Then the "right" question becomes: How do we implement? Nonetheless, the amount of philosophical, conceptual, and structural change required to implement weekly measurement in general education classrooms requires some recognition. In noting the difficulty of achieving school change, Erickson (1996) quipped "It reminds us of the joke about how many psychiatrists it takes to change a light bulb: Only one, but it takes a long time and the light bulb really has to want to change" (p. 91).

From this perspective, I used an existing database of two cohorts of first grade children to examine the utility of CBM measures collected in September and January to identify children who in May met various criteria for reading disability. To be sure, this was not the optimal data set to examine questions concerning growth. Monthly or weekly measures in the fall were not collected and Letter Sound Fluency (LSF) was the only CBM measure available on all children. Ideally, the questions would be examined with a variety of CBM and published measures varying the timing of measurement. Oral Reading Fluency (ORF) data were collected beginning in January because of expected floor effects earlier in the year. Both LSF and ORF data were collected either weekly or monthly beginning in January. The technical characteristics of these measures are strong and well established (e.g., Deno, 1985; Fuchs & Fuchs, 1998; 1999; Speece & Case, 2001). Although limited for investigating the issue of efficiency, the data are useful for illustrating several points and raising relevant research questions.

Sample

We screened all first grade children in two consecutive cohorts (*n*= 679) with two LSF probes in late September/early October to identify samples of at-risk (AR) and not-at-risk children (NAR). AR children were those who earned mean scores in the bottom 25% of their classrooms. NAR children scored at the 30^{th}, 50^{th} (two children), 75^{th}, and 90^{th} percentiles for their classrooms and represented the range of reading ability in the classroom. There were 140 AR children and 136 NAR children in the sample. In order to interpret the data from a normative perspective, the data for the NAR children were weighted by a factor of four as each

NAR child represented approximately four children in the population. This procedure yielded a weighted *N* of 684, which is close to the population *n* of 679. From this point forward the AR and NAR designations refer not to the above subject selection procedures but to classifications formed by the various screening methods used (described below).

The characteristics of the weighted sample generally reflected a normal distribution: WJ-R Basic Reading Skills Cluster Score *M* = 107.2 (SD = 15.5); Full Scale IQ *M* = 103.1 (*SD* = 16.9). The May Oral Reading Fluency descriptive statistics were Level *M* = 46.0 (*SD* = 37.0); Slope *M* = 1.42 (*SD* = .97). The slope estimate approximates other early elementary school samples reported in the literature (e.g., 1.5 in Fuchs, 2003; Deno et al., 2001), while the level data are lower than some reports (e.g., *M* = 66 [*SD* = 41.9], Marston et al., 2003).

Screening Groups

Screening groups were formed using cutoff scores - 1 SD below the mean for the following variables: LSF Fall Screen, LSF January Level, LSF January Slope, LSF January Level and Slope (DD), and LSF Growth. The cutoff score for ORF January Level was defined as a score below the 25^{th} percentile. The LSF Growth variable was calculated by subtracting LSF Fall Screen from LSF January Level. The - 1 SD criterion was selected because a more reasonable percentage of children was identified as AR (Mdn. = 16.96, range 8.04 to 23.8) compared to a -.5 SD criterion (Mdn. = 33.3; range 18.71 to 42.7).

Criterion Variables

Two approaches were used to assess the validity of the screening groups. First, the AR and NAR children from each screen were compared on the available end-of-year reading variables: WJ-R Basic Reading Skills Cluster standard score (BRS SS; composite of Letter Word Identification and Word Attack), WJ-R Passage Comprehension standard score (PC SS), ORF level and ORF slope. ORF level was the average of the last two data points administered in May and reflected the number of words read correctly in one minute from connected text; ORF slope was determined through OLS regression and reflected growth from January to mid-May in terms of weekly increase in words read correctly. Children classified as AR should have lower end-of-year reading scores than children not classified.

Second, several "reading disability" classifications were created so that classification analyses could be conducted with the screen groups. These classifications were based on procedures either recommended or used in the literature to define reading problems. Four classifications were created: BRS SS 90 (BRS standard score less than 90 or the 25^{th} percentile); BRS SS 78 (BRS standard score less than 78 or approximately 1.4 SD below the normative mean); ORF 40 (ORF score below the benchmark criterion of 40 words per minute); and ORF DD (ORF slope and level -1SD below the weighted sample's mean).

All of the screening groups were crossed with all the reading disability groups to produce 2 x 2 tables for calculation of diagnostic utility indexes. Four indexes are reported: Sensitivity (TP/TP + FN), specificity (TN/TN + FP); Positive Predictive Power (TP/TP + FP); and Negative Predictive Power (TN/TN + FN). The extent to which each index approaches 100% is an indication of classification accuracy.

Table 1. *Effect Sizes for Reading Variables Comparing At-Risk and Not-At-Risk Children Defined by Screening Criteria*

Screen Group	% At Risk	End of Year Reading Variable	ES/d
LSF FALL	17.4		
		BRS SS	1.25
		PC SS	1.40
		ORF LEVEL	1.05
		ORF SLOPE	0.89
LSF GROWTH	16.96	BRS SS	0.54
		PC SS	0.45
		ORF LEVEL	0.37
		ORF SLOPE	0.50
LSF JAN LEVEL	19.3	BRS SS	1.30
		PC SS	1.21
		ORF LEVEL	1.04
		ORF SLOPE	0.86
LSF JAN SLOPE	14.33	BRS SS	0.30
		PC SS	0.20
		ORF LEVEL	0.29
		ORF SLOPE	0.05
LSF JAN DD	8.04	BRS SS	0.83
		PC SS	0.73
		ORF LEVEL	0.73
		ORF SLOPE	0.52
ORF JAN LEVEL*	23.83	BRS SS	1.52
		PC SS	1.81
		ORF LEVEL	1.52
		ORF SLOPE	1.35

Results

Group differences. Table 1 presents the percentage of children classified as AR for each screening group and the effect size (*d* adjusted for sample sizes) for each end-of-year reading variable. The largest AR group was produced by the ORF January cut point of the 25th percentile and the fewest AR children were identified by the LSF January DD criteria. Large effect sizes (ES) (all differences were statistically significant; $p < .0001$) were associated with the following screening groups: LSF Fall, LSF January Level, and ORF January Level. Moderate ES (all p values < .0005) were found for LSF Growth and LSF January DD with small ES for LSF January Slope ($p < .01$ BRS and ORF Level; $p > .05$ PC and ORF Slope). The groups that used level to define the cut points produced larger AR/NAR differences than did the groups that had some aspect of growth in the definition. Although it is possible that growth does not make for a good screen, it is just as likely that LSF is not a good measure for first grade progress monitoring. That is, it worked well as an initial measure to define research groups but by late Fall/early January word measures may be more appropriate.

Table 2. Diagnostic Utility Indexes for Screening and Diagnostic Groups and Percentage of Reading Disabled Children in Each Diagnostic Group

	Reading Disability Criteria															
Screening Group	BRS SS 90 (18.3 %)				BRS SS 78 (6.9%)				ORF 40 (55.9 %)				ORF DD (11.6%)			
	Sens	PPP	Spec	NPP	Sens	PPP	Spec	NPP	Sens	PPP	Spec	NPP	Sens	PPP	Spec	NPP
LSF FALL	48.8	51.3	89.6	88.7	40.3	16	84.3	95	27.8	89.1	95.7	51.2	55.7	37	87.6	93.8
LSF GROWTH	32	34.5	86.4	85	32	12.9	84.1	94.4	20.4	67.2	87.4	46.5	33.6	49.4	87.3	93
LSF JAN LEVEL	51.2	48.5	87.8	89	42.6	15.2	82.4	95.1	30.1	87.1	94.4	51.6	69.6	41.7	87.3	95.7
LSF JAN SLOPE	19.2	24.5	86.7	82.8	23.4	11.2	86.3	93.9	17	66.3	89.1	45.9	27.8	22.5	87.4	90.3
LSF JAN DD	15.2	34.6	93.6	83.2	14.9	12.7	92.5	93.6	12.0	83.6	97	46.6	27.9	40	94.6	90.9
ORF JAN LEVEL	63.2	48.5	85	91	57.5	16.6	78.7	96.2	41.6	97.6	98.7	57.2	**94.9**	**46**	**85.5**	**99.2**

Classification Analyses

Table 2 presents the percentage of children identified as reading disabled (RD) for each criterion group and the classification indexes. As expected, the BRS SS < 78 group produced the fewest RD children (6.9%) and the ORF < 40 group produced the most (55.9%). Overall, the specificity indexes were quite high indicating that children who were *not* RD in May were identified as such by the screen. For the most part, the negative predictive power (1 - false negative rate) was also high. Children who were NAR on the screen were likely not RD at the end of the year. The exception to this is the RD group defined by ORF < 40 which is predictable since over half the sample was identified as RD.

Sensitivity, the percentage of RD children picked up by the screen, and positive predictive power, percentage of AR children who were classified as RD, were more variable. For all four reading disability criteria, ORF January Level produced the highest sensitivity index and in two of four cases it produced the highest positive predictive power. In most cases, however, the level of accuracy would be considered unacceptable. The ORF January screen and the ORF DD reading disability criterion produced the most accurate classification: sensitivity = 94.9%, positive predictive power = 46% specificity = 85.8%, and negative predictive power = 99.2%.

Discussion

The high false positive rate (54%; 1 - positive predictive power) of the January ORF Level/May ORF DD classification is similar to other efforts in first grade (e.g., O'Connor & Jenkins, 1999) and remains a vexing problem not solved by the use of LSF measures and attempts to assess LSF growth. It would seem that, in addition to evaluating the power of word-level tasks (both CBM and published measures), a second screening step that incorporated a quick hitting intervention may be one way to reduce the number of false positive cases. It must be kept in mind that a dynamic assessment task used as part of a first grade screen in the O'Connor and Jenkins (1999) study reduced false positive cases but the rate still may be considered high at 47.4%. However, they used a very strict criterion to define reading disability (-1.4 SD) that will necessarily increase the number of false positive cases.

To say that more work needs to be done is both true and unsatisfying to those who wish to implement *now*. Jim Gallagher once described the frustration that policy makers experience with researchers. He noted that researchers have to be right 95 times out of 100 to make a recommendation whereas policy makers would be ecstatic with a recommendation that holds a 50-50 chance of success. If we accept as reasonable a 50% false positive rate given high sensitivity, it means we are comfortable identifying twice the number of children for secondary interventions than who actually need them. If this is the case then the fall first grade procedures identified by O'Connor and Jenkins (1999) might work (dynamic segmentation, sound repetition, rapid letter naming), as might a January screen on Oral Reading Fluency with a cutoff value of 6 wpm (our 25th percentile value). There are two obvious difficulties with these recommendations that must be understood: The O'Connor and Jenkins battery incorporated a dynamic assessment task, which added 30-35 min to the screen, and the January ORF screen, while efficient, may be viewed by some as too late and as ignoring the importance of growth.

With respect to research there are several avenues that may yield more insight on how to identify children for secondary interventions. The only fairly firm conclusion to be drawn from the classification data I presented is that Letter Sound Fluency is likely not a good variable to pursue in first grade. Thus, as mentioned, comparing word-level tasks across a variety of measurement time points would address questions of efficiency and growth. Before accepting a 50% false positive rate, these approaches to screening need to be evaluated.

Given the number of studies devoted to early identification and their disappointing accuracy results, the conclusion that "growth matters" seems inescapable. After all, that is the definition of a false positive case: the child "grew" despite expectations to the contrary. Traditionally, screening has been conceptualized as a two-part process: screen and outcome defined by individual differences. I recommend we view screening as a three-legged stool that incorporates not only individual differences at the screen and outcome but also development conceptualized as rate of learning. Continuing to search for the right combination of child characteristic measures without attention to learning is likely a bankrupt strategy.

References

Byrne, B., Fielding-Barnsley, R., & Ashley, L. (2000). Effects of preschool phonemic identity training after six years: Outcome level distinguished from rate of response. *Journal of Educational Psychology, 92,* 659-667.

Case, L. P., Speece, D. L., & Molloy, D. E. (2003). The validity of a response-to-instruction paradigm to identify reading disabilities: A longitudinal analysis of individual differences and contextual factors. *School Psychology Review, 32, 557-582.*

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52,* 219-232.

Deno, S.L., Fuchs, L.S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30,* 507-524.

Erickson, F. (1996). Inclusion into what? Thoughts on the construction of learning, identity, and affiliation in the general education classroom. In D.L. Speece & B.K. Keogh (Eds.), *Research on classroom ecologies: Implications for inclusion of children with learning disabilities* (pp. 91-105). Mahwah, NJ: Erlbaum. Fletcher, J., & Satz, P. (1984). Test-based versus teacher-based predictions of academic achievement: A three-year longitudinal study. *Journal of Pediatric Psychology, 9*, 193 – 203.

Foorman, B.R., & Torgesen, J. (2001). Critical elements of classroom and small group instruction promote reading success in all children. *Learning Disabilities Research & Practice, 16,* 203-212.

Fuchs, L. S. & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review, 28,* 659-671.

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

Fuchs, D., Fuchs, L.S., Thompson, A., Al Otaiba, S., Yen, L., Yang, N.J., et al. (2001). Is reading important in reading – readiness programs? A randomized field trial with teachers as program implementers. *Journal of Educational Psychology, 93,* 251-267.

Fuchs, L. S. (1995, May). *Incorporating curriculum-based measurement into the eligibility decision-making process: A focus on treatment validity and student growth.* Paper presented at the Workshop on IQ Testing and Educational Decision Making, National Research Council, National Academy of Science, Washington, DC.

Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research & Practice, 81,* 172-186.

Jenkins, J.R., & O'Connor, R.E. (2002). Early identification and intervention for young children with reading/learning disabilities. In R. Bradley, L. Danileson, & D.P. Hallahan (Eds.), *Identification of learning disabilities: Research to Practice* (pp. 99-149). Mahwah, NJ: Erlbaum.

Juel, C. (1988). Learning to read: A longitudinal study of 54 children from first through fourth grade. *Journal of Educational Psychology, 80,* 437-447.

Marston, D., Muyskens, P., Lau, M., & Carter, A. (2003). Problem-solving model for decision making with high-incidence disabilities: The Minneapolis Experience. *Learning Disabilities Research & Practice, 81,* 187-200.

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

Phillips, L. M., Norris, S. P., Osmond, W. C., & Maynard, A. M. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. *Journal of Educational Psychology, 94,* 3-13.

Scanlon, D. M., & Vellutino, F. R. (1996). Prerequisite skills, early instruction, and success in first-grade reading: Selected results from a longitudinal study. *Mental Retardation and Developmental Disabilities, 2*, 54 - 63.

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.

Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology, 93,* 735-749.

Speece, D.L., & Cooper, D.H. (1990). Ontogeny of school failure: Classification of first grade children. *American Educational Research Journal, 27,* 119-140.

Torgesen, J.K., Alexander, A.W., Wagner, R.K., Rashotte, C.A., Voelher, K.K.S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34,* 33-58, 78.

Torgesen, J.K., Wagner, R.K., Rashotte, C.A., Rose, E., Lindamood, P., Conway, T., et al. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology, 91,* 579-593.

Vellutino, F.R., Scanlon, D.M., Sipay, E.R., Small, S.G., Pratt, A., Chen, R.S., et al. (1996). Cognitive profiles of difficult to remediate and readily remediated poor reader: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88,* 607-638.