Published in final edited form as:

J Learn Disabil. 2009 ; 42(2): 163–176. doi:10.1177/0022219408326219.

Floor effects associated with universal screening and their impact on early identification of reading disabilities

Hugh W. Catts,
University of Kansas

Yaacov Petscher,
Florida State University, Florida Center for Reading Research

Christopher Schatschneider,
Florida State University, Florida Center for Reading Research

Mindy Sittner Bridges, and
University of Kansas

Katherin Mendoza
Florida Center for Reading Research

Abstract

Response to intervention (RTI) holds great promise for the early identification and prevention of reading disabilities. The success of RTI rests in part on the accuracy of universal screening tools used within this framework. Despite advancements, screening instruments designed to identify children at risk for reading disabilities continue to have limited predictive validity. In this study, we examine a common screening instrument for the presence of floor effects and investigate the impact that these effects have on the predictive validity of the instrument. Longitudinal data (kindergarten-3rd grade) from a large cohort of children were used. These data included children's performances on five measures from DIBELS (Dynamic Indicators of Basic Early Literacy Skills) and two reading achievement outcome measures. Results showed that DIBELS measures were characterized by floor effects in their initial administrations, and that these effects reduced the predictive validity of the measures. The implications of these findings for early identification are discussed.

Programmatic research over the last decade demonstrates that early intervention can significantly improve the reading outcomes of children at risk for reading disabilities (Denton & Mathes, 2003; Foorman, Francis, Fletcher, Schatschneider & Mehta, 1998; O'Connor, Fulmer, Harty, & Bell, 2005; Simmons, Coyne, Kwok, McDonagh, Harm, & Kame'enui, in press). To benefit maximally from this intervention, children need to be identified early (Cavanaugh, Kim, Wanzek, & Vaughn, 2004; Torgesen, 2002; Torgesen et al., 1999; Vellutino, Scanlon, Small, & Fanuele, 2006). Recently, response to intervention (RTI) has been proposed as a framework that holds great promise for the early identification of children at risk for reading disabilities (Fletcher, Coulter, Reschly, & Vaughn, 2004;

Fuchs, Fuchs, & Speece, 2002; Vaughn & Fuchs, 2003). Within the typical RTI model, all students receive periodic screening (i.e., universal screening) for risk for reading disabilities (RD). Students identified as at risk on the basis of this screening are provided with short-term intervention. Those who fail to respond to intervention are considered to be “truly” at risk for RD and are provided with more specialized intervention.

Despite the promise of RTI to improve the early identification of RD, numerous challenges remain. A major challenge is the development of universal screening tools that can accurately identify an initial pool of children at risk. Whereas some advancements have been made in the development of screening instruments (Catts, Fey, Zhang, & Tomblin, 2001; Foorman, Fletcher et al., 1998; Kaminski & Good, 1996; O'Connor & Jenkins, 1999; Wood, Hill, Meyer, & Flowers, 2005), these instruments continue to be characterized by high rates of under- or over-identification (Glover & Albers, 2007; Jenkins, 2003; Jenkins, Hudson, & Johnson, 2007). Such identification errors result in over-looking children who need intervention or wrongly identifying those who do not. Both types of errors can be costly in terms of instructional time and resources.

Speece (2005) has argued that a major challenge for early identification of RD is finding a screening instrument that can hit a “moving target.” She noted that children are developing the very skills that are being assessed on screening instruments, but screening procedures seldom take this into consideration. Part of this development involves the maturation of cognitive-linguistic abilities related to literacy (Catts, Fey, Tomblin, & Zhang, 1999). Literacy experience and instruction also play a major role in the underpinnings of literacy development (Adams, 1990; Mann & Wimmer, 2002). Experience/instruction with the alphabet and sound-letter correspondence impacts the skills that are tested on screening instruments. This is clearly the case for measures involving letter and word recognition (Mann & Wimmer, 2002). In addition, phonological awareness, which is often argued to be a precursor of reading (Gough & Hillinger, 1980; Wagner & Torgesen, 1987), is also influenced by experience with reading (Barron, 1991; Castles & Coltheart, 2004; Hogan, Catts, & Little, 2005; Lundberg & Høien, 1991; Morais, 1991).

Speece (2005) argued that it may be helpful to use measures of growth in screening in order to address the “moving target” problem. This would entail making screening decisions on the basis of multiple assessments given over an extended period of time. Although measures of growth may improve the accuracy of early identification (Compton, Fuchs, Fuchs, & Bryant, 2006; but see Schatschneider, Wagner, & Crawford, in press), the time and effort required to implement these measures may limit their widespread use in universal screening. Alternatively, decisions could be based on a universal screening instrument given at a single point in time, if this point was optimally selected. The optimum time point for such an instrument would be the point at which children have enough experience and the ability level to provide adequate variance on a screening scale. If a screening tool is given too early, many children will not have the experience or ability level to perform the task, and will score near the lower end of the distribution. Such a *floor effect* can lead to a high rate of over-identification and reduced accuracy of early identification.

The purpose of this paper was to examine the distributions of scores from a common screening instrument that had been administered to a large group of children at various time points in the early elementary grades. Particular attention was paid to the presence of floor effects (and their time course) in these distributions, and the impact that floor effects had on the predictive validity of the instrument. The screening instrument that we investigated in this study was The Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2003). Whereas this instrument was developed as a progress monitoring tool, it is now commonly used as a universal screening tool to identify children as at risk for RD (U.S. Department of Education, Office of Inspector General, 2007). Given DIBELS widespread use, it is important to investigate the role of floor effects in the prediction accuracy of this instrument. However, our results should also be relevant to other universal screening instruments that are susceptible to floor effects.

Method

Participants

Participants were 18,667 children who were enrolled in Florida *Reading First* schools and who began kindergarten in the 2003–04 school year. This was a diverse cohort of students in which 49.7% were female, 39.4% were white, 31.9% were African American, 22.4% were Latino, 15.2% were identified as English language learners, and 72.5% of the sample were eligible for free or reduced-price lunch.

Measures

The data used in this study were obtained from the Progress Monitoring and Reporting Network (PMRN) maintained by the Florida Center for Reading Research as part of its role in providing support for statewide *Reading First* programs. We examined data collected from our cohort while they were in kindergarten, first, second, and third grades. The PMRN database included data from five DIBELS (Good & Kaminski, 2003) assessments (Initial Sound Fluency, Letter Name Fluency, Phoneme Segmentation Fluency, Nonsense Word Fluency, and Oral Reading Fluency), and a standardized measure of reading comprehension. Each is described below.

Initial Sound Fluency (ISF)—This task measures children’s awareness of the initial sounds in words. In this task, the examiner provides the names of four pictures, and children are required to identify the picture that begins with the sound produced by the examiner. Children are also asked to produce the beginning sounds of words presented orally by the examiner. The amount of time taken to identify/produce the correct sounds is converted into the number of initial sounds correct in one minute. The alternative-form reliability of this measure is .72 (Good et al., 2004). Test-retest reliability completed on a subset of 2,408 children from our sample was .66.

Letter Naming Fluency (LNF)—This task measures children’s ability to rapidly identify upper- and lower-case letters of the alphabet arranged in a random order. The score is the number of letter names identified correctly in one minute. The alternative-form reliability of

this measure is .88 (Good et al., 2004). Test-retest reliability completed on a subset of 2,408 children from our sample is .90.

Phoneme Segmentation Fluency (PSF)—This task measures children's ability to segment words into phonemes. The examiner asks children to say individually each of the phonemes in three- and four-phoneme words. The number of phonemes said correctly in one minute is the index of performance. The alternative-form reliability of this measure is .79 (Good et al., 2004). Test-retest reliability completed on a subset of 2,408 children from our sample is .69.

Nonsense Word Fluency (NWF)—This task measures children's ability to use letter-sound correspondence to decode words. In this task, children are presented with printed vowel-consonant or consonant-vowel-consonant nonsense words (e.g., ov, sig, rav) and asked to produce verbally the individual letter sound of each letter or verbally produce, or read, the whole nonsense word. The number of letter-sound correspondences produced correctly in one minute is the index of performance. The alternative-form reliability of this measure is .83 (Good et al., 2004). Test-retest reliability completed on a subset of 2,408 children from our sample is .86.

Oral Reading Fluency (ORF)—This task measures children's reading accuracy and speed with connected text. Children read each of three previously unseen grade level passages aloud for one minute. Words omitted, substituted, and hesitations of more than three seconds are scored as errors. Words self-corrected within three seconds are scored as accurate. The number of words correctly read in one minute from each passage is recorded and the median value from the three passages is taken as the index of performance. The alternative-form reliability of this measure is .95 (Good, Kaminski, Smith, & Bratten, 2001). Test-retest reliability completed on a subset of 2,408 children from our sample is .96.

In addition to DIBELS assessments, the Reading Comprehension subtest from the Stanford Achievement Test (10th edition; SAT-10; Harcourt Brace, 2003) was also administered. For this group-administered measure, which is given in February of third grade, children read short passages silently and answered multiple choice questions. The reliability of this measure is .88.

DIBELS Administration Schedule—The schedule for the DIBELS assessment in the PNRN database was set within Florida *Reading First* programs. According to this schedule, DIBELS measures used in kindergarten through second grade were administered four times a year in September, December, February, and April. The DIBELS website also provides a schedule in which measures can be given 3 times a year (Fall, Winter, and Spring) rather than 4, but the initial administration time point for each measure is comparable to that in the Florida system.¹ In the Florida system, ISF was given during the first 3 testing sessions of kindergarten. LNF was administered during the four testing sessions of kindergarten and the first session of first grade. PSF was given in the last two sessions of kindergarten and all

¹The only appreciable difference is that the DIBELS website recommends that ORF be initially administered in Winter of first grade rather than Fall.

sessions in first grade. NWF was administrated in the last two sessions of kindergarten and all four sessions of first and second grades. ORF was given in all four sessions of first and second grades and 3 times a year in third grade. All assessments were administered by well-trained and reliable school- or district-based assessment teams, which included no classroom teachers

The entire sample did not participate at every time point. Complete data were available on 12,762 (68.4%). The remaining 5,905 participants had missing data for one or more assessments and/or time points. In some instances, missing data was due to incidental absences, whereas in other cases it resulted from children moving in and out of the *Reading First* schools. The total number of children who were tested at each time point for each measure is shown in Table 1. These data show that the sample size varied only slightly (plus or minus 2.8%) from a mean sample size of 17,848. Furthermore, multiple comparisons of the samples revealed that their composition in terms of gender, ethnicity, ELL, and free or reduced-priced lunch status did not differ appreciably from each other or that of the whole sample.

Criterion for Reading Outcomes

For some of our predictive analyses, it was necessary to classify children as good or poor readers based on reading outcome measures.² For predictive analyses involving ISF, LNF, PSF, and NWF, we used ORF (3rd grade, April) as the outcome measure. This measure was chosen because ISF, LNF, PSF, and NWF are intended to provide information relevant to word reading abilities. For these analyses, children were identified as poor readers if their ORF outcome score in April of third grade was below 110 words correct per minute. This is the recommended cut-off score for “some risk” at this time point in the DIBELS Data System (Good & Kaminski, 2003). Children with scores of 110 words correct per minute or higher were defined as good readers. This approach led to a base rate of poor readers of .29

For analyses in which ORF was the predictor variable, the Reading Comprehension subtest of the *SAT-10* (February of 3rd grade) served as the outcome measure. This measure seemed most appropriate because ORF is often used to determine if children are at risk for failure on statewide reading tests, which primarily measure comprehension (Buck & Torgesen, 2003; Vander Meer, Lentz, & Stollar, 2005; Wilson, 2005). In the case of the *SAT-10*, Florida, like many other states, uses the 40th percentile as a cut-off for below grade-level reading. Whereas this criterion is more liberal than that used in many classification studies of poor readers (e.g., Catts et al., 2001; O’Conner & Jenkins, 1999), it seemed appropriate to use in this study in order to obtain a comparable base rate to that found for the ORF poor reader criterion. Using the above *SAT-10* cut-score resulted in a base rate of .33.³

²We have chosen long-term outcomes to test the influence of floor effects on the accuracy of early identification. Whereas it is common to use universal screening instruments for long-term prediction (Catts et al., 2001; Silbergitt & Hintz, 2005), screening tools can also be used to predict reading outcomes over short-term periods in a successive manner (e.g., Hintz & Silbergitt, 2005).

³Whereas this base rate is comparable to that found for the ORF outcome, it is lower than that expected (.40) from a normative sample. It is unclear why our sample had a lower base rate for this criterion. One possible influence may have been the quality of instruction that our sample received in *Reading First* schools.

Results

Two sets of analyses were carried out in this study. In the first set of analyses, we examined the distributional characteristics of DIBELS measures looking specifically for the presence of floor effects. For large samples such as ours, floor effects are easily detected by visual inspection (Tabachnick & Ridell, 2001). Figure 1 displays frequency histograms for each of the DIBELS measures at each measurement time point. Examination of these series of histograms reveals a somewhat similar change in the shape of the distribution over time. Each DIBELS measure is initially characterized by an abnormal distribution with a strong floor effect. Over time the floor effects lessen in varying degrees. Examining each measure individually, the ISF distributions show a slight reduction in floor effects across time points but do not appear to normalize before the measure was eliminated from the testing protocol. The LNF distributions, on the other hand, show a more normal shape by the third or fourth time the measure was given. The PSF initially had a bimodal distribution with a strong floor effect. By the fourth administration, the PSF distribution appears to be unimodal with no evidence of a floor effect, and in fact, a negative skew appears to emerge. The NWF distributions show a reduction in floor effects by the fourth administration whereas the ORF distributions seem to become more normal only after the sixth administration.

Table 1 provides further data concerning the distributional characteristics of DIBELS assessments. Included in this table are the mean, standard deviation, kurtosis, and skewness. High standard deviations relative to the means in the initial administrations of each of the DIBELS measures are consistent with the floor effects seen in Figure 1. Kurtosis and skewness can also provide an indication of floor effects. Kurtosis is an index of the peakedness of a distribution. A high positive value is associated with a distribution that has a large peak and heavy tails. Whereas a strong floor effect may lead to high kurtosis, this need not be the case. Skewness is an index of the symmetry of the distribution. The higher the absolute value of skewness, the further the mean is from the middle of the distribution. Positive skewness is indicative of a mean that is closer to the lower end of the distribution and an asymmetric tail extending toward the higher end of the scale. Because the mean of a distribution is generally lowered by a floor effect, higher positive skewness values often accompany floor effects.

Tests of significance for skewness and kurtosis are available but were not run because almost any departure from normal would be significant with our large sample (Tabachnick & Ridell, 2001). Changes in kurtosis values shown in Table 1 correspond at least partially to changes in the floor effects seen in visual inspection. Kurtosis values were initially high (greater than 1) for ISF, NWF, and ORF, and showed some reduction across time points. Changes in skewness values more closely corresponded to changes seen in visual inspection of the distributions. Skewness values near or greater than 1 were found for the initial assessment of each of the measures, except PSF, and these values became smaller across time. The PSF was not characterized by positive skewness, but rather had a negative skew on most administrations, which can also be seen in visual inspection.

Whereas floor effects can sometimes be quantified by skewness and kurtosis statistics, these effects can be more directly identified by examining the percentage of the sample that

performed at the floor of each distribution. To calculate this percentage, we identified the floor as the lowest quarter of the range of scores in each distribution. Because the range of scores can be overestimated by outliers, we converted scores to a z-score scale and limited this range to scores that fall between -3 and $+3$. In a normal distribution, 99.5% of the scores fall within this z-score range. Additionally, within a normal distribution, approximately 7% of the scores would be expected to fall within the lowest quarter of the -3 to $+3$ range. The percentages of children with scores in the lowest quarter of each distribution are included in Table 1. These data clearly show that each of the DIBELS measures was characterized by strong floor effects in the initial administrations. Over time, the percentages of children in the lowest quarter of the range decreased and this mirrored changes seen in the visual inspection of the data. It is only after this percentage approached 7% that floor effects seemed to disappear from the distributions.

Predictive Analyses

In a second set of analyses, we examined the predictive validity of DIBELS measures. Special attention was given to the impact that floor effects in these assessments had on the prediction of subsequent reading outcomes. Again, ORF (April, 3rd grade) served as the outcome measure for ISF, LNF, PSF, and NWF, whereas *SAT 10* (February, 3rd grade) was the outcome measure for ORF. Figure 2 shows the quantile regression plots between the outcome measure and each of the DIBELS assessments at each administration time point. Quantile regression plots show the change in correlation between predictor and outcome for various levels of the predictor variable. Values from the quantile regression are calculated in a similar fashion as an ordinary least squares regression. In ordinary least squares analysis, when estimating the sample mean of a criterion given a predictor, the goal is to minimize a sum of squared residuals, to which a fit line is produced in a scatterplot. Quantile regression, similar to ordinary least squares regression looks to minimize a sum of absolute residuals in order to produce a *point estimate for the median*. In a symmetric distribution, the point estimate from the quantile regression (i.e., the median) will be the same as the value from the ordinary least squares regression. This process of estimating values for the median (i.e., .50 quantile) in quantile regression can be extended to obtain point estimates across a continuum of quantiles by asymmetrically weighting values above and below the quantile of interest (Koenker, 2005). Resulting coefficients from a quantile regression can be plotted to indicate the magnitude of the relationship between the criterion and the predictor (plotted on the y-axis) as a function of the quantile of interest (plotted on the x-axis).

Quantile regression plots show that there were low to moderate correlations between DIBELS assessments and outcome measures. For each DIBELS assessment, except for the PSF, quantile plots indicate that the predictability of the DIBELS measures typically improved across administration time points. For example, at the .50 quantile for LNF, the correlation changed from .35 in September of kindergarten to .53 one year later. However, for most measures, much of the improvement in predictability occurred at the lower quantile levels of the predictor variable. This improvement at the lower levels closely mirrored the reduction of floor effects seen in Figure 1. The quantile plot for PSF shows a different pattern. First, the correlations between PSF and reading outcome are generally quite low. In

addition, rather than the correlations going up across time points, the correlations involving PSF actually go down.

Logistic regression analyses were also undertaken to examine the predictability of DIBELS measures. These analyses provide classification indices concerning how accurate DIBELS assessments are in predicting which children will be good or poor readers on outcome measures. As reported above, third grade ORF served as the outcome measure for ISF, LNF, PSF, and NWF, whereas third grade *SAT-10* served as the outcome measure for ORF. Table 2 contains results from the logistic regression analyses. Area under the curve (AUC) provides an overall estimate of the predictability of a measure. It is an index of the area under the Receiver Operator Characteristic (ROC) curve. A ROC curve is a plot of true positive rate (i.e., the percentage of poor readers correctly identified by the predictor, also known as sensitivity) versus false positive rate (i.e., the percentage of good readers incorrectly identified by the predictor) for each of the possible cut-off scores of the predictor. AUC is an estimate of how accurate a screening tool will classify two randomly chosen individuals, one from the poor outcome group and one from the good outcome group. Thus, AUC values range from .5 (i.e., chance level) to 1.0 (i.e., perfect classification). Values above .80 are generally considered desirable for screening measures (Metz, 1978).

Beyond considering AUC, Jenkins (2003; Jenkins, Hudson, & Johnson, 2007) has argued that the most appropriate way to assess a measure's classification accuracy is to choose a desired sensitivity level and then evaluate the acceptability of the corresponding false positive rate. From a practitioner standpoint, it is also useful to examine the corresponding positive prediction power. This rate is an estimate of the percentage of children who are identified as at risk by the predictor and who turn out to be poor readers. This statistic is particularly valuable because it tells the practitioner how likely a positive finding is really positive. Although valuable, positive prediction power is influenced by the base rate of the disorder and can be artificially inflated when base rate is high (Meehl & Rosen, 1955; Schatschneider, Petscher, & Williams, 2008), as is the case in this study.

Table 2 provides the false positive and positive prediction rates for each measure and time point when the sensitivity level or true positive rate was set at .80, .85, and .90. A sensitivity level of .80 is considered a minimum or near minimum level of acceptability for a predictor of RD (Carran & Scott, 1992; Jansky, 1977; Kingslake, 1983), whereas a rate of .90 is a more optimal level (Jenkins, 2003). As far as the acceptability of false positives rates, there is little consensus in the literature. However, in most RTI frameworks a false positive rate of a 50% or less would seem acceptable given that Tier 2 is designed to reduce this further. An acceptable level for positive prediction power is more variable, primarily because this statistic is heavily influenced by the base rate of the outcome condition.

There are a number of noteworthy findings in Table 2. Results indicate that the ISF and PSF subtests had rather limited predictive validity. Both of these measures were associated with low AUC values (<.70) and high false positive rates (>.50). Also, note that the AUC values for PSF actually decreased across time points. Although unexpected, this finding is consistent with the data that were reported in the quantile plots of this measure. LNF and

NWF proved to be much better predictors of reading outcome than ISF and PSF. ORF was also a good predictor of reading outcome, however, the predictive results for ORF cannot be compared directly to those of other DIBELS assessments because of the difference in the outcome measures. Data indicate that for LNF, NWF, and ORF, the AUC values increased across time points with much of the increase occurring in the initial several administrations.⁴ A somewhat similar corresponding pattern was seen in false positive and positive prediction values. This early improvement in predictability is consistent with the lessening of the floor effects in each measure. LNF begins to reach its optimum predictive validity in the spring of kindergarten. At that time point, it appears to be a good predictor of ORF two years later. It has a low false positive rate and a reasonable positive prediction value at sensitivity levels below .90. NWF had less predictive validity than LNF when they were administered together in spring of kindergarten. However, by first grade, NWF showed similar predictability to LNF. At the beginning of second grade, NWF was a good predictor of ORF over a year and one half later. Finally, ORF proved to be a fair predictor of reading outcome based on the SAT 10 (February, 3rd grade). However, it was not until the beginning/middle of second grade that optimal rates of predictability were reached.

Discussion

In this study, we examined universal screening data from a large longitudinal database. This database included student performances on DIBELS assessments in kindergarten through second grade. In our initial analyses, we investigated the distributional characteristics of the scores on each of the DIBELS measures at each administration time point. Our results showed that these measures were initially characterized by strong floor effects. These floor effects were apparent in each of the DIBELS measures, and were not linked to a particular grade level but rather to the initial several administrations of each assessment. Further examination indicated that these floor effects lessened across administrations and eventually all measures except ISF showed a normal or near normal distribution.

The floor effects observed in student performances on the DIBELS measures impacted the ability of these measures to predict future reading outcomes. Analyses involving quantile correlations showed that the strength of the correlation between each of the DIBELS assessments, except PSF, and its corresponding outcome measure increased with subsequent administrations. Such a pattern is to be expected, given that with each subsequent administration, the time duration between the administration of the predictor and the outcome measures narrowed. However, most of the changes in the strength of the relationship between the predictor and the outcome measures occurred across the initial several administrations of each measure. Furthermore, these changes were especially apparent at the lower values of the DIBELS measures. Such a pattern corresponds well to what would be predicted given the presence of floor effects in the initial administrations of DIBELS assessments. That is, floor effects obscure individual differences among those who score in the lower end of the distribution and lessen the correlation between their scores and later outcome measures.

⁴It is possible to test the statistical significance of differences in AUC across time points. However, given the very small standard errors associated our data, almost any difference in AUC value is significant.

The impact of these floor effects were also seen in the results of the logistic regression analyses. These analyses showed that each of the DIBELS measures, except PSF, improved across administrations in its predictability of reading outcome. In addition, much of the improvement occurred across the first several administrations in correspondence with the lessening of floor effects. However, the PSF subtest actually became less predictive over time. This measure was associated with high false positives rates and low positive prediction values across administration time points. Others have also found PSF to have less predictive power than other DIBELS measures (Hintze, Ryan, & Stoner, 2003; Kloo, 2006). Whereas the ISF improved in its predictability of reading outcome over its three administration time points, its false positive rate remained high even at a .80 sensitivity level. Such a finding suggests that by itself this measure may have somewhat limited value as a universal screening instrument when long-term prediction is of primary concern. The other DIBELS measures proved to be better predictors of reading outcome. Both LNF and NWF showed good predictability by the end of kindergarten (>.80 AUC). However, these measures continued to be associated with high false positive rates at a sensitivity level of .90. Finally, by second grade, ORF proved to be a good indicator of later reading status based on the *SAT 10*. At that point, it had near acceptable false positive and positive prediction rates at the .90 sensitivity level. These findings are consistent with that reported by Roehrig, Petscher, Nettles, Hudson, and Torgesen (2008) for a different outcome measure in a sample that overlaps with the current sample. Our results are also similar to those reported by Silberglitt and Hintze (2005). They observed floor effects in a measure of oral reading fluency when it was administered in winter and spring of first grade and in the fall of second grade. These floor effects were reduced by winter of second grade. Predictability of third grade reading comprehension also improved as floor effects were reduced..

Factors Related to Floor Effects

The floor effects observed in the initial administrations of DIBELS assessments may result from multiple factors. It is possible that these measures were initially administered before many children had the “cognitive maturity” to perform one or more of the tasks. Limitations in linguistic or meta-cognitive abilities or a lack of attentional resources could impact performance on these measures. However, it seems more likely that many children did not have the necessary experience or instruction to perform well on the initial administrations of each of the DIBELS measures. Experience and instruction with the alphabet and sound-letter correspondence clearly impacts performance on literacy measures like the LNF, NWF, and ORF measures. Numerous studies highlight the differences in letter knowledge of children from print-rich versus print-impooverished homes (Burgess, Hecht, & Lonigan, 2002; Bus, van Ijzendoorn, & Pelligrini, 1995). Cross-linguistic comparisons also support the role of experience/instruction in literacy acquisition. Mann and Wimmer (2002), for example, reported that German kindergarteners scored much lower on tests of letter knowledge and word decoding than did American kindergarteners. Mann and Wimmer argued that this difference was the consequence of pedagogical differences such that children from German-speaking countries do not generally receive explicit literacy instruction until first grade, whereas American children receive such instruction in kindergarten or preschool.

A lack of literacy experience and/or instruction could also impact performance on measures of phonological awareness, such as the ISF and PSF measures. Although these measures require judgments about spoken language units, they can be impacted by literacy experience as well. Barron (1991), for example, has argued that knowledge of letters makes children more aware of the phonemic nature of spoken language. Alternatively, Castles, Holmes, Neath, & Kinoshita (2003) have proposed that learning to read may not change children's explicit awareness of phonemes as much as it allows them to use their spelling knowledge to perform phonological awareness tasks. Regardless of the explanation, there is strong evidence that literacy experience/instruction influences performance on phonological awareness measures (Castles & Coltheart, 2004; Hogan, Catts, & Little, 2005; Lundberg & Høien, 1991; Mann & Wimmer, 2002; Morais, 1991), and thus, may impact initial performance on assessments such as the ISF and PSF.

Implications

Our results have implications for universal screening within an RTI framework. The intent of universal screening is to identify a pool of children at risk for RD who will receive short-term intervention in Tier 2. A universal screening tool should identify as many children truly at risk as possible (i.e., true positives), while at the same time, limit the number of children falsely identified (i.e., false positives). Our results highlight the importance of not only choosing the right instrument for universal screening but the appropriate schedule for this screening. The screening schedule used in the Florida *Reading First* programs appears to be less than optimal. Each of the DIBELS measures was administered on a schedule that resulted in floor effects and poor predictability in its initial administrations. Although the Florida schedule differs from that found in some states (four vs. three testing points a year), the Florida schedule for the initial assessment of each measure is essentially the same as that in other states.

One option for reducing the floor effects and improving the accuracy of DIBELS screening would be to delay the initial administration of each of the measures until a more optimal time point. Such an approach could lead to much greater screening accuracy in the case of the LNF, NWF, and ORF. If LNF was given initially in Winter or Spring of kindergarten, NWF in Fall of first grade, and ORF in Fall or Winter of second grade, more optimal screening accuracy should be achieved in each case. Our data suggests that by themselves ISF and PSF may not reach acceptable levels of accuracy at any time point in their recommended schedule.

A major problem with delaying the administration of DIBELS screening until more optimal levels of accuracy can be achieved is that the identification of children at risk for RD will itself be delayed. Alternatively, DIBELS assessments could be administered according to the Florida *Reading First* schedule or other comparable schedules but supplemented with other screening measures. For example, ISF DIBELS assessments could be employed as a "first cut" to identify children who do not meet the benchmark, and a follow-up phonological awareness measure could be used to help reduce the false positives that are introduced by floor effects.⁵ It is our understanding that such an approach is already being followed in many school districts. Whereas this approach could potentially reduce the false positive

errors associated with the ISF measure, other static measures of phonological awareness can also be affected by floor effects, especially at the beginning of kindergarten. For example, most of the phonological awareness subtests from the Comprehensive Test of Phonological Processing (Wagner, Torgesen, & Rashotte, 1999) and the Phonological Awareness Test (Robertson & Salter, 1998) appear to show floor effects for beginning kindergarteners.

Rather than supplement DIBELS assessments with static measures, it might prove more beneficial to employ dynamic assessments. In dynamic assessment, children are provided with explicit feedback during testing. Dynamic assessment is linked theoretically to Vygotsky's (1987) notion of the "zone of proximal development." This zone represents the potential gain between what a child can do independently versus what he/she can do with assistance. Applied to screening, a dynamic assessment provides an index of how well a child might be expected to respond to instruction in the classroom. Rather than a measure of a learned product (i.e., static assessment), it is a measure of the potential to learn (Grigorenko & Sternberg, 1998). Furthermore, a dynamic assessment might be useful in reducing false positives by providing many children with the extra knowledge or experience they are lacking. If experiential deficits underlie the floor effects as argued above, the instruction in dynamic assessment, although limited, could be enough to reduce these effects. Indeed, one recent study (O'Connor & Jenkins, 1999) has shown that dynamic assessment of phonological awareness can reduce false positive errors in early identification. Dynamic assessments have also been developed to assess decoding ability (Fuchs, Fuchs, Compton, Bouton, Caffrey, & Hill, 2007). One disadvantage of these assessments is that they can be lengthy to administer. The dynamic assessments used by O'Connor and Jenkins (1999) and Fuchs et al. (2007) took approximately 30 minutes to administer. It may be possible to reduce the administration time of these instruments, but they are likely to still best be used as a supplement to DIBELS or other static assessments. As such these instruments could be administered to children who do not meet the benchmark or cut-off score on the initial screening instrument.

Limitations

One potential limitation of this study is that data were drawn exclusively from *Reading First* schools. By designation, these schools include a higher percentage of children at risk. Indeed our sample had more minority representation (60.6% vs. 53.3%), more English-language learners (15.2% vs. 8.8%) and more children qualifying for free or reduced-price lunch (72.5% vs. 45.8%) than the State of Florida as a whole. Therefore, it is possible that the high-risk status of the sample contributed to the presence of the floor effects in the initial administrations of the DIBELS measures. The extent of this contribution is unclear; but given that floor effects seem to also be present at the lower age/grade range of similar standardized assessments (Robertson & Salter, 1998; Torgesen, Wagner, & Rashotte, 1999; Wagner, Torgesen, & Rashotte, 1999) with more normative samples, we would expect these effects to generalize, at least in part, to samples with lower *Reading First* representation. Confirmation of this generalization, however, must await similar data using more normative

⁵Using ISF as a first cut would require that a cut-off value be selected that would assure high sensitivity.

samples. However, our findings should generalize well to other *Reading First* schools and/or to the many other school districts with similar student demographics.

One additional limitation is imposed by our sample. In the Florida *Reading First* schools, as in many other schools, DIBELS assessments are used in a programmatic manner to identify children for intervention and monitor progress. Thus, for many children in our sample, there was an intervening agent (intervention) between universal screening assessments and outcome measurements. Such procedures make the interpretation of data concerning classification accuracy more difficult. As Good, Cummings, and Powell-Smith (2008) point out, successful intervention improves the outcomes of children at risk, and by doing so, changes true positives to false positives. As a result, estimates of sensitivity and specificity are compromised. Whereas this is a legitimate concern, it is almost unavoidable in an educational setting. In such a setting there will always be differentiated instruction/intervention for children at risk and withholding such instruction/intervention for several years until “nature runs its course,” is clearly unethical. The issue of an intervening agent does not, however, negate the need for improving the accuracy of universal screening tools. Our results suggest that universal screening tools such as DIBELS can be influenced by floor effects. These effects, which are in fact likely to be independent of intervention, need to be reduced in order to improve screening accuracy.

References

- Adams, MJ. *Beginning to read: Thinking and learning about print*. Cambridge: MIT Press; 1990.
- Barron RW. Proto-literacy, literacy, and the acquisition of phonological awareness. *Learning and Individual Differences*. 1991; 3:243–255.
- Buck, J.; Torgesen, J. The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test (Technical Report 1). Tallahassee: Florida Center for Reading Research; 2003.
- Burgess SR, Hecht SA, Lonigan CJ. Relations of the home literacy environment (HLE) to the development of reading-related abilities: A one-year longitudinal study. *Reading Research Quarterly*. 2002; 37:408–426.
- Bus AG, van Ijzendoorn MH, Pelligrini A. Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research*. 1995; 65:1–21.
- Carran DT, Scott KG. Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics in Early Childhood Special Education*. 1992; 12:196–211.
- Castles A, Coltheart M. Is there a causal link from phonological awareness to success in learning to read? *Cognition*. 2004; 91:77–111. [PubMed: 14711492]
- Castles A, Holmes VM, Neath J, Kinoshita S. How does orthographic knowledge influence performance on phonological awareness tasks? *The Quarterly Journal of Experimental Psychology*. 2003; 56A:445–467. [PubMed: 12745843]
- Catts H, Fey M, Zhang X, Tomblin JB. Language basis of reading and reading disabilities: Evidence from a longitudinal study. *Scientific Studies of Reading*. 1999; 3:331–361.
- Catts HW, Fey ME, Zhang X, Tomblin JB. Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools*. 2001; 32:38–50.
- Cavanaugh CL, Kim A, Wanzek J, Vaughn S. Kindergarten reading interventions for at-risk students: Twenty years of research. *Learning Disabilities: A Contemporary Journal*. 2004; 2:9–21.

- Compton DL, Fuchs D, Fuchs LS, Bryant JD. Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*. 2006; 98:394–409.
- Denton, CA.; Mathes, PG. Intervention for struggling readers: Possibilities and challenges. In: Foorman, BR., editor. *Preventing and remediating reading difficulties: Bringing science to scale*. York: Timonium, MD; 2003. p. 229–251.
- Fletcher JM, Coulter WA, Reschly DJ, Vaughn S. Alternative approaches to the definition and identification of learning disabilities: Some questions and answers. *Annals of Dyslexia*. 2004; 54:304–331. [PubMed: 15741940]
- Foorman BF, Francis DJ, Fletcher JM, Schatschneider C, Mehta P. The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*. 1998; 90:37–55.
- Fuchs D, Fuchs LS, Compton DL, Bouton B, Caffrey E, Hill L. Dynamic assessment as responsiveness to intervention: A scripted protocol to identify young at-risk readers. *Teaching Exceptional Children*. 2007; 39:58–63.
- Fuchs LS, Fuchs D, Speece DL. Treatment validity as a unified construct for identifying learning disabilities. *Learning Disability Quarterly*. 2002; 25:33–46.
- Glover TA, Albers CA. Considerations for evaluating universal screening assessments. *Journal of School Psychology*. 2007; 45:117–135.
- Good, RH.; Cummings, KD.; Powell-Smith, KA. ROC done right? Part 3 DIBELS Benchmark goals; San Diego, CA. Paper presented at the annual Pacific Coast Research Conference; Feb. 2008
- Good, RH.; Kaminski, RA. DIBELS Indicators of Basic Early Literacy Skills. Longmont, CO: Sopris West Educational Services; 2003.
- Good, RH.; Kaminski, R.; Shinn, M.; Bratten, J.; Shinn, M.; Laimon, L., et al. Technical adequacy and decision making utility of DIBELS (Technical Report No. 7). Eugene, OR: University of Oregon; 2004.
- Good, RH.; Kaminski, R.; Smith, MR.; Bratten, J. Technical adequacy and second grade DIBELS Oral Reading Fluency (DORF) passages (Technical Report No. 8). Eugene, OR: University of Oregon; 2001.
- Gough PB, Hillinger ML. Learning to read: An unnatural act. *Bulletin of the Orton Society*. 1980; 30:179–196.
- Grigorenko EL, Sternberg RJ. Dynamic testing. *Psychological Bulletin*. 1998; 124:75–111.
- Harcourt Educational Measurement. Stanford Achievement Test. 10th ed. San Antonio, TX: Author; 2003.
- Hintze JM, Ryan AL, Stoner G. Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review*. 2003; 32:541–556.
- Hintze JM, Silberglitt B. A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review*. 2005; 34:372–386.
- Hogan TP, Catts HW, Little TD. The relationship between phonological awareness and reading: Implications for the assessment of phonological awareness. *Language, Speech, and Hearing Services in the Schools*. 2005; 36:285–293.
- Jansky, JJ. A critical review of 'some developments and predictive precursors' of reading disabilities. In: Benton, AL.; Pearl, D., editors. *Dyslexia*. New York: Oxford University Press; 1977.
- Jenkins JR, Hudson RF, Johnson ES. Screening for at-risk readers in a response-to-intervention (RTI) framework. *School Psychology Review*. 2007; 36:582–600.
- Jenkins, JR. Candidate measures for screening at-risk student Paper presented at the Conference on Response to Treatment as Learning Disabilities Identification. Kansas City, MO: National Research Center on Learning Disabilities; 2003 Nov.
- Kaminski RA, Good RH III. Toward a technology for assessing basic literacy skills. *School Psychology Review*. 1996; 25:215–227.
- Kingslake B. The predictive (in)accuracy of on-entry to school screening procedures when used to anticipate learning difficulties. *Special Education: Forward Trends*. 1983; 10:23–26. [PubMed: 6665590]

- Kloo, AM. Unpublished doctoral dissertation. 2006. The decision-making utility and predictive power of DIBELS for students' reading achievement in Pennsylvania's Reading First schools.
- Koenker, R. Quantile regression. NY, NY: Cambridge University Press; 2005.
- Lundberg, L.; Høien, T. Initial enabling knowledge and skills in reading acquisition: Print awareness and phonological segmentation. In: Sawyer, DJ.; Fox, BJ., editors. *Phonological awareness in reading: The evolution of current perspectives*. New York: Springer-Verlag; 1991. p. 74-95.
- Mann V, Wimmer H. Phoneme awareness and pathways to literacy: A comparison of German and American children. *Reading and Writing: An Interdisciplinary Journal*. 2002; 15:653–682.
- Meehl PE, Rosen A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*. 1955; 3:195–216.
- Metz CE. Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine*. 1978; 8:283–298. [PubMed: 112681]
- Morais, J. Phonological awareness: A bridge between language and literacy. In: Sawyer, DJ.; Fox, BJ., editors. *Phonological awareness in reading: The evolution of current perspectives*. New York: Springer-Verlag; 1991. p. 31-71.
- O'Connor RE, Fulmer D, Harty K, Bell K. Layers of reading intervention in kindergarten through third grade: Changes in teaching and child outcomes. *Journal of Learning Disabilities*. 2005; 38:440–455. [PubMed: 16329445]
- O'Connor RE, Jenkins JR. Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading*. 1999; 3:159–197.
- Robertson, C.; Salter, W. The Phonological Awareness Test. East Moline, IL: LinguiSystems; 1997.
- Roehrig AD, Petscher Y, Nettles SM, Hudson R, Torgesen JK. Accuracy of the DIBELS Oral Reading Fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology*. 2008; 46:343–366. [PubMed: 19083363]
- Schatschneider, C.; Petscher, Y.; Williams, KM. How to evaluate a screening process: The vocabulary of screening and what educators need to know. In: Justice, L.; Vukelic, C., editors. *Achieving excellence in preschool language and literacy instruction*. New York: Guilford Press; 2008. p. 304-316.
- Schatschneider C, Wagner RK, Crawford E. The importance of measuring growth in response to intervention models: testing a core assumption. *Learning and Individual Differences*. (in press).
- Silbergliitt B, Hintze J. Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psycheducational Assessment*. 2005; 23:304–325.
- Simmons DC, Coyne MD, Kwok O, McDonagh S, Harn B, Kame'enui EJ. Indexing response to intervention: A longitudinal study of reading risk from kindergarten through third grade. *Journal of Learning Disabilities*. 2008; 41:158–173. [PubMed: 18354935]
- Speece DL. Hitting the moving target known as reading development: Some thoughts on screening children for secondary interventions. *Journal of Learning Disabilities*. 2005; 38:487–493. [PubMed: 16392689]
- Tabachnick, BG.; Fidell, LS. *Using multivariate statistics*. Boston: Allyn and Bacon; 2001.
- Torgesen J. The prevention of reading difficulties. *Journal of School Psychology*. 2002; 40:7–26.
- Torgesen, JK.; Wagner, RK.; Rashotte, CA. *Test of Word Reading Efficiency*. Austin, TX: Pro-Ed; 1999.
- Torgesen JK, Wagner RK, Rashotte CA, Rose E, Lindamood P, Conway T, et al. Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology*. 1999; 91:579–593.
- U.S. Department of Education, Office of Inspector General (OIG). The department's administration of selected aspects of the Reading First program: Final audit report (ED-OIG/A03G0006). Washington, DC: Author; 2007 Feb. Available at <http://www.ed.gov/about/offices/list/oig/auditreports/a03g0006.pdf>
- Vander Meer, CD.; Lentz, FE.; Stollar, S. The relationship between oral reading fluency and Ohio proficiency testing in reading (Technical Report). Eugene, OR: University of Oregon; 2005.

- Vaughn S, Fuchs LS. Redefining learning disabilities as inadequate response to intervention: The promise and potential problems. *Learning Disabilities Research and Practice*. 2003; 18:137–146.
- Vellutino FR, Fletcher JM, Snowling MJ, Scanlon DM. Specific reading disability (dyslexia): What have we learned in the past four decades. *Journal of Child Psychology and Psychiatry*. 2004; 45:2–40. [PubMed: 14959801]
- Vygotsky, LS. The collected works of L.S. Vygotsky. Vol. 1. New York: Plenum Press; 1987.
- Wagner RK, Torgesen JK. The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*. 1987; 101:192–212.
- Wagner, RK.; Torgesen, JK.; Rashotte, CA. *Comprehensive Test of Phonological Processing*. Austin, TX: Pro-Ed; 1999.
- Wilson, J. The relationship of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency to performance on Arizona Instrument to Measure Standards (AIMS) (Technical Report). Eugene, OR: University of Oregon; 2005.
- Wood FB, Hill DF, Meyer MS, Flowers DL. Predictive assessment of reading. *Annals of Dyslexia*. 2005; 55:193–216. [PubMed: 17849193]

Note: The range of values on the y-axis varies across histograms to allow for optimum display of frequency distributions. The range of values on the x-axis are consistent within each measure

Figure 1.

Frequency histograms for each of the DIBELS measures at each of the time points.

Figure 2.
Quantile regression plots of DIBELS measures and corresponding outcome measures

Table 1

Number of children, mean, standard deviation, kurtosis, skewness, and the percentage of children in the lowest quarter of the range for each of the DIBELS assessments.

	N	Mean	SD	Kurtosis	Skewness	Lowest Quarter of Range
ISF Sept K	17,204	10.051	8.640	12.008	1.847	64.2
ISF Dec K	17,740	19.814	12.708	4.639	1.266	45.1
ISF Feb K	18,245	25.281	13.928	1.815	1.299	34.1
LNF Sept K	17,214	15.677	14.102	0.059	0.819	54.2
LNF Dec K	17,739	30.313	16.310	-0.140	0.255	28.9
LNF Feb K	18,244	37.027	15.734	0.188	0.256	14.9
LNF April K	18,645	45.574	16.527	0.094	0.177	8.1
LNF Sept 1 st	17,827	49.403	15.587	0.340	-0.046	4.5
PSF Feb K	18,245	28.308	18.509	-1.082	0.059	36.1
PSF April K	18,644	36.063	16.358	-0.350	-0.369	17.9
PSF Sept 1 st	17,823	40.236	16.667	0.146	-0.471	13.4
PSF Dec 1 st	16,960	44.692	13.976	0.734	-0.472	6.9
PSF Feb 1 st	17,831	50.947	13.009	1.743	-0.603	4.9
PSF April 1 st	17,835	48.258	12.807	2.029	-0.064	5.1
NWF Feb K	18,245	22.163	16.329	3.679	1.161	52.0
NWF April K	18,643	35.256	19.608	4.483	1.234	40.5
NWF Sept 1 st	17,826	37.022	21.624	4.118	1.357	44.5
NWF Dec 1 st	16,963	52.507	24.571	3.284	1.339	22.4
NWF Feb 1 st	17,831	58.654	28.059	2.697	1.261	23.4
NWF April 1 st	17,833	69.064	31.879	1.282	1.033	21.5
NWF Sept 2 nd	18,233	65.947	31.657	1.407	1.089	23.1
NWF Dec 2 nd	18,361	80.922	37.201	0.834	0.925	22.7
NWF Feb 2 nd	18,546	87.094	38.624	0.576	0.794	20.6
NWF April 2 nd	18,644	95.204	42.377	0.417	0.728	21.6
ORF Sept 1 st	17,825	21.080	19.390	6.575	2.230	77.8

	N	Mean	SD	Kurtosis	Skewness	Lowest Quarter of Range
ORF Dec 1 st	16,963	32.905	25.233	2.323	1.431	61.3
ORF Feb 1 st	17,832	42.375	30.689	0.736	1.087	54.6
ORF April 1 st	17,836	52.648	30.543	0.762	0.879	38.8
ORF Sept 2 nd	18,234	56.907	31.666	0.517	0.749	33.4
ORF Dec 2 nd	18,362	69.178	31.822	0.346	0.488	21.3
ORF Feb 2 nd	18,547	81.695	34.580	0.138	0.234	15.0
ORF April 2 nd	18,644	92.967	35.250	0.351	0.166	10.2

Table 2

Area under the curve (AUC) and false positive frequency and positive prediction power for various levels of sensitivity.

Time Point	AUC	Sensitivity					
		.80		.85		.90	
		FP	PPP	FP	PPP	FP	PPP
ISF Sept K	.629	.68	.33	.77	.31	.86	.30
ISF Dec K	.678	.58	.36	.70	.33	.78	.32
ISF Feb K	.693	.54	.38	.65	.35	.76	.33
LNF Sept K	.711	.59	.36	.69	.34	.79	.32
LNF Dec K	.786	.40	.45	.52	.40	.62	.38
LNF Feb K	.822	.32	.51	.42	.46	.57	.40
LNF April K	.842	.28	.54	.41	.46	.55	.40
LNF Sept 1 st	.853	.24	.58	.35	.50	.52	.42
PSF Feb K	.682	.59	.36	.68	.34	.72	.34
PSF April K	.682	.55	.38	.65	.35	.72	.34
PSF Sept 1 st	.669	.58	.36	.68	.34	.72	.34
PSF Dec 1 st	.633	.60	.36	.68	.34	.72	.34
PSF Feb 1 st	.606	.59	.36	.72	.33	.78	.32
PSF April 1 st	.604	.66	.33	.72	.33	.80	.32
NWF Feb K	.775	.40	.45	.51	.41	.68	.35
NWF April K	.815	.36	.48	.44	.44	.58	.39
NWF Sept 1 st	.849	.28	.54	.38	.48	.51	.42
NWF Dec 1 st	.843	.26	.56	.39	.47	.52	.42
NWF Feb 1 st	.869	.22	.60	.31	.53	.49	.43
NWF April 1 st	.878	.21	.61	.30	.54	.44	.46
NWF Sept 2 nd	.893	.20	.62	.26	.58	.38	.50
NWF Dec 2 nd	.879	.20	.62	.28	.56	.42	.47
NWF Feb 2 nd	.897	.19	.64	.25	.58	.42	.47
NWF April 2 nd	.898	.18	.65	.22	.62	.41	.48
ORF Sept 1 st	.743	.52	.43	.58	.41	.70	.38

Time Point	AUC	Sensitivity					
		.80		.85		.90	
		FP	PPP	FP	PPP	FP	PPP
ORF Dec 1 st	.770	.45	.46	.53	.44	.63	.41
ORF Feb 1 st	.784	.40	.49	.52	.44	.59	.42
ORF April 1 st	.791	.39	.50	.50	.45	.58	.43
ORF Sept 2 nd	.805	.36	.52	.47	.47	.55	.44
ORF Dec 2 nd	.814	.34	.53	.42	.49	.52	.46
ORF Feb 2 nd	.818	.33	.54	.40	.51	.50	.47
ORF April 2 nd	.821	.33	.54	.40	.51	.48	.48