Early Identification of Reading Difficulties
A Screening Strategy that Adjusts the Sensitivity to the Level of Prediction Accuracy
Poulsen, Mads; Nielsen, Anne-Mette Veber; Juul, Holger; Elbro, Carsten

Published in:
Dyslexia

DOI:
10.1002/dys.1560

Publication date:
2017

Citation for published version (APA):
Early identification of reading difficulties: A screening strategy that adjusts the sensitivity to the level of prediction accuracy

Mads Poulsen, Anne-Mette Veber Nielsen, Holger Juul, and Carsten Elbro
University of Copenhagen

This is the peer reviewed version of the following article: Early identification of reading difficulties: A screening strategy that adjusts the sensitivity to the level of prediction accuracy, which has been published in final form at


This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions

Abstract
Early screening for reading difficulties before the onset of instruction is desirable because it allows intervention that is targeted at prevention rather than remediation of reading difficulties. However, early screening may be too inaccurate to effectively allocate resources to those who need them. The present study compared the accuracy of early screening before the onset of formal reading instruction with late screening six months into the first year of instruction. The study followed 164 Danish students from the end of Grade 0 to the end of Grade 2. Early screening included measures of phonemic awareness, rapid naming, letter knowledge, paired associate learning, and reading. Late screening included only reading. Results indicated that reading measures improved substantially as predictors over the first six months of Grade 1, to the point where late reading measures alone provided as much information as the early measures combined. In the light of these results and a less than perfect early screening accuracy, a new strategy for screening is introduced and discussed. The strategy proposes multi-point screening with gradually increasing sensitivity to strike a balance between manageable screening procedures and outcomes and early identification of students who are most likely in need of extra resources.

Keywords: reading disability, early identification, early prevention, screening
Early identification of children at risk for reading difficulties is important because early intervention can to some degree prevent some of the negative effects associated with poor reading skills (Elbro & Scarborough, 2004a; Poskiparta, Niemi, Lepola, Ahtola, & Laine, 2003; Torgesen 2005; Torgesen, Rashotte, & Alexander, 2001). Ideally, children who are at risk for reading failure should be identified before the onset of reading instruction and provided with special attention and support to prevent (rather than remediate) reading failure. Screening tests can provide objective means for deciding who should be given special attention.

Most previous studies of early identification of reading difficulties have focused on which measures to include in screening batteries to increase the precision of early screening with the purpose of identifying as many who need intervention as possible (e.g. Catts, Nielsen, Bridges, Liu, & Bontempo, 2015; Compton et al., 2010; Bridges & Catts, 2011; Johnson, Jenkins, Petscher, & Catts, 2009). The continued development of sensitive screening batteries is important. But until screening tools with high accuracy and manageable costs are available, it is important to explore the ways in which existing screening tools can be used to make the most of the limited resources in schools. The present study was concerned with investigating practically manageable strategies for screening with existing tools. Although the strategies are explored with a particular data set and a particular set of tools, the purpose is to explore strategies for screening in general in the face of less than perfect precision in screening.

A basic problem is that screening is far from 100 % accurate (Catts et al., 2015; Catts, Petscher, Schatschneider, Sittner Bridges, & Mendoza, 2009; Compton et al., 2010; Johnson et al., 2009). Some children pass the screening but later experience difficulties, while other children fail the screening but do well in reading later on. Early screening before instruction faces the fundamental difficulty that reading ability cannot be measured directly without strong floor
effects (Catts et al., 2009). Certain precursor measures are known to predict later reading ability, for example letter knowledge, phonological awareness and rapid automatized naming (e.g. Catts et al., 2009; Elbro & Scarborough, 2004b; Landerl & Wimmer, 2008; Kirby, Parrila, & Pfeiffer, 2003; Lervåg, Bråten, & Hulme, 2009; Poulsen, Juul, & Elbro, 2014; Puolakanaho et al., 2008). But once under way, reading tends to be a very strong predictor of future reading (Manis, Seidenberg, & Doi, 1999; Parrila, Kirby, & McQuarrie, 2004; Wagner, Torgesen, & Rashotte, 1994). For instance, in the study of Wagner et al. (1994) kindergarten phonological analysis was the strongest unique predictor of reading ability one year later. In Grade 1, however, reading was the best predictor of reading ability one year later in Grade 2.

This leaves a dilemma. Early screening before instruction is preferable because it allows for timely intervention, but early screening may be imprecise because it cannot use the best predictor of future reading ability. The main empirical purpose of the present study was to investigate the differences in identification accuracy of future reading difficulty before and after the onset of formal reading instruction and to suggest ways of dealing with the dilemma.

**Evaluating identification accuracy**
Several statistics are useful for evaluating identification accuracy (Metz, 1978). The *true positive rate* (also known as the *sensitivity*) of a test is the proportion of children who are correctly identified as at risk (true positives) out of all those who actually become poor readers. This rate should be high, because the whole point of screening is to identify as many as possible of those children who are on a path to become poor readers so that they can be provided with intervention.

On the other hand, there will be false positives, i.e. the screening will flag some children as at risk who will turn out to become normally developing readers. The *false positive rate* (also
known as 1-specificity) is the proportion of children who are incorrectly flagged as at risk (false positives) out of all those children who become normally developing readers. This rate should be low, because false positives engender unnecessary concern for the children, worries among and their parents, and unnecessary use of intervention resources, that could be used to improve intervention for those who truly need the intervention. Figure 1 presents an overview of these statistics.

[insert figure 1 about here]

The true positive and the false positive rates are determined by the accuracy of the screening tool and by the chosen cut-off value on the screening test. A relatively low cut-off will yield relatively many true positives, but also relatively many false positives, depending on the accuracy of the screening test.

Receiver operating characteristics (ROC) analysis allows evaluation of the accuracy of a screening tool independently of particular cut-off values (Metz, 1978). A ROC-curve plots the true positive rate on the y-axis against the false positive rate on the x-axis, where each point on the curve represents a cut-off value on the screening tool. More accurate tests will show sharp rises in the curve because high true positive rates will be associated with low false positive rates. The trade-off is quantified in a single statistic the area under the curve (AUC). Thus, the AUC compares the accuracy of tests irrespectively of particular cut-off values, and there are significance tests for comparing AUC’s (e.g. DeLong, DeLong, & Clarke-Pearson, 1988).

From a practical point of view, the true positive rate and false positive rate are more practically informative than AUC because they specify the consequences of particular cut-offs for the screening. Guidelines for minimum true positive rates range around 75-90% (Glover & Albers, 2007; Johnson et al., 2009).
It could be argued that from a school resources standpoint, an important statistic is the total positive rate, that is the proportion of students that will be flagged for special attention, the true plus the false positives. The total positive rate is the statistic that translates most directly into the cost of acting on the screening results. For example, it may not be economically feasible to provide intensive preventive intervention for 30% of a year group. Another way of looking at this is that, on a fixed budget, the total positive rate determines how much attention the individual at-risk student will get because the attention has to be divided between the true positives and the false positives. The total positive rate is usually heavily influenced by the false positive rate since most students do not end up having reading difficulties.

**Dealing with poor identification accuracy**

It is important to find manageable ways for schools and teachers to deal with poor identification accuracy. Much previous research has focused on improving early identification accuracy by intensifying and sophisticating the screening procedures. For example, some studies have devised ways of testing the potential for learning through short-term progress monitoring or dynamic testing (Catts et al., 2015; Compton et al., 2010; Compton, Fuchs, Fuchs, & Bryant, 2006; Fuchs, Compton, Fuchs, Bouton, & Caffrey, 2011; Sittner Bridges & Catts, 2011). Response to intervention (RTI) approaches are an extension of this where response to group-based “Tier 2” intervention is used as a filter for determining who gets more expensive, individual “Tier 3” intervention (e.g. Compton et al., 2012; Fuchs et al., 2011; Gilbert et al., 2013). Common to these approaches is that they are relatively resource demanding in terms of the time needed to administer the screening tests and possibly also in terms of the skills needed to administer and interpret the results (e.g. Compton et al., 2012). Researchers have recognized this and suggested the solution that only children who are picked out by less expensive screening
measures are tested with more resource demanding screening measures. For example, Compton et al. (2010) suggested a two-stage gated screening procedure that employed a relatively simple prescreening measure. This measure could with 100% sensitivity rule out around 40% of the true negatives, leaving 60% for more intensive testing to weed out additional the false positives. Such approaches work by the philosophy that sensitivity should be kept high initially. False positive rates of up to 50% have been mentioned as acceptable within Tier 1 screening of response to intervention frameworks (Catts et al., 2009). False positive rates in this range mean that more than half of the population needs follow-up attention in the form of further screening and possibly intervention. Thus, although the reduction in the number of students that need more intensive screening effort is substantial, the remaining proportion of students that need intensive testing is still high. Therefore, there is still practical reasons for exploring ways to deal with the problem of false positives with the screening tools at hand.

If it turns out that screening improves markedly after the onset of instruction when reading skill can be used as a predictor, a simple solution is to postpone screening to a point in time when the students have had a chance to respond to general instruction. Of course, this entails postponing intervention.

Another simple solution to the problem of high false positive rates is to lower cut-off values, which lowers both true and false positive rates - to meet affordable total positive rates. Of course, lowering sensitivity entails overlooking more children who might benefit from intervention.

Both of these simple approaches -- either postponing intervention or overlooking students in need of intervention -- have obvious drawbacks. But these drawbacks could potentially be mitigated by combining them in a multi-point screening strategy with increasing sensitivity to match screening accuracy. Early screening could be conducted with low sensitivity to identify
some of the students in need of attention while limiting the number of false positives. If late screening is more accurate, additional screening could then be conducted at later points with increasing sensitivity levels to identify the remaining students in need of attention with more manageable total positive rates. We will refer to this as an increasing sensitivity strategy. This represents a different multi-point strategy than the previously mentioned response to intervention approach, which aims to start with high sensitivity and uses consecutive screening and intervention to home in on the students that need intensive attention.

It may be argued that intensive screening approaches, such as RTI, are superior because they are more likely to ensure that children in need of attention will get it early. However, it is worth exploring other solutions, because some school systems may not be prepared for extensive screening or the structured RTI implementation. The difficulty of implementing comprehensive testing procedures has been acknowledged within the RTI literature (Compton et al., 2012). The danger is that high costs of screening and intervention could limit the amount of intervention given or even be prohibitive of making structured screening at all. In the Danish school system, there is a strong tradition for giving individual teachers considerable autonomy over what happens in their classroom, and most instruction and intervention is expected to take place in the classroom with limited outside help. A basic, non-intensive screening system that identified a manageable number of students in need of the teacher’s close attention might be a more realistic solution than the RTI framework in a foreseeable future. Other educational systems may not be prepared for RTI for different reasons. Finally, some recent studies have failed to demonstrate improved identification accuracy from using information about response to intervention (Catts et al., 2015; Compton et al., 2012; Tran, Sanchez, Arellano, & Swanson, 2011). Thus, although the intervention in RTI may be helpful (cf. Gilbert et al., 2013, for recent review), it is not clear that
RTI improves identification of who needs the help, at least not beyond information about how students respond to ordinary instruction. In summary, limited accuracy of early identification of reading difficulty is still a practical problem.

**The present study**

Given the problem of early, accurate, and manageable identification, the present study used data from an ongoing longitudinal study (AUTHORS) to compare the identification accuracy before and after formal instruction. The purpose was to inform practical strategies for screening using a manageable set of common predictor measures, not so much to evaluate the particular set of predictors as to explore ways of dealing with the realities of the identification accuracy of common screening tools. More specifically, the study compared the identification accuracy between pre- and post-instructional screening for end of Grade 2 reading difficulty in a sample of Danish children that was followed from end of Grade 0 to the end of Grade 2. Identification accuracy was compared both in terms of AUC, which is cut-off independent, and practically relevant numbers of false positives and total positives at different sensitivity levels. With the problems of high positive rates in mind, the present study departed from earlier studies (e.g. Bridges & Catts, 2011; Catts et al., 2009; Compton et al. 2010) in exploring sensitivity levels below the recommended 80% to evaluate different ways of dealing with potentially low identification accuracy.

Danish is a relatively deep orthography (Elbro 2005; Seymor, Aro, Erskine, & COST Action A8 network, 2003) like English. Grade 0 functions as a preparation for school. There is some instruction in letters and sounds, but most students cannot read at the end of the school year. Formal reading instruction sets in in Grade 1. The individual teacher has considerable autonomy over how the formal reading instruction is carried out. A phonics-approach is recommended in
the guidelines from the Ministry of Education (Danish Ministry of Education, 2014) and supported by the most widespread teaching materials (Borstrøm & Petersen, 1999; Jacobsen & Nielsen, 2012).

**Method**

**Participants**
One hundred and eighty-seven students in an ongoing longitudinal study (AUTHORS) were followed from May of Grade 0. The present study reports data from Grade 0 in 2009 to the end of Grade 2. One hundred and sixty-four students had complete data sets, 87 boys and 77 girls. The following analyses were based on this subset. In Denmark children usually enter Grade 0 in August of the year they turn six. The mean age of the sample was 6;10.

**Design and procedure**
Pre-instruction screening was carried out in April and May of Grade 0. It consisted of a comprehensive screening battery with multiple reading precursor measures. The reading abilities of the students were tested several times during Grade 1 and 2. For the present study, the reading measure from January of Grade 1 served as the post-instructional screening point. This point in time was selected because waiting until half way through Grade 1 would allow differences in response to the standard instruction to show on simple reading measures, while leaving time for intervention within the first year of instruction (although the study did not include special intervention).

Reading scores from May of Grade 2 were used as reading outcome measures.

**Measures**
In addition to well-known precursor measures of rapid naming, letter knowledge, and phonemic awareness, we selected a measure of paired associate learning for the pre-instructional screening
battery to maximize identification accuracy. Recent studies have suggested that verbal paired associate learning is correlate of reading (Elbro & Jensen, 2005; Litt & Nation, 2014; Poulsen, Juul, & Elbro, 2015; Warmington & Hulme, 2011, but cf. Lervåg & Hulme, 2009). All measures were administered individually by trained student assistants (cf. AUTHORS, for more comprehensive description of measures).

**Rapid automatized naming.** This task measured the students’ ability to rapidly name the digits 1-5 displayed in five rows of ten digits. Accuracy and completion time was measured from audio recordings using using the Praat speech analysis software (Boersma, 2001). The score was the number of correctly named digits per second. The correlation with RAN with objects in this sample has been reported at $r = .65$, which is a lower bound estimate of the reliability (AUTHORS).

**Letter knowledge.** This task (from AUTHORS) measured the students’ ability to name each of the 29 upper-case letters in the Danish alphabet. Cronbach’s alpha for the full sample was .92.

**Phonemic awareness.** This task (adapted from AUTHORS) measured the students’ ability to identify and delete initial, medial or final phonemes in spoken words. The students heard a word and were asked to say the word that was left when a given phoneme was removed from the word. There were 19 items. Cronbach’s alpha was .91.

**Paired associate learning.** This visual-verbal learning task measured the students’ ability to associate nonsense word names with non-familiar cartoon animals. The examiner introduced three cartoon animals and their names one at a time. Afterwards the students had to produce the names in response to rows of the three animals in different orders over 15 trials. The score was the number of correctly named animals. The reliability of this measure is uncertain (cf. AUTHORS, for discussion).
**Grade 0 oral word reading accuracy.** For this experimenter-devised measure, the students were asked to attempt to read aloud 32 words (mainly content words), all of which were deemed to be well within the spoken vocabularies of six-year-olds. As the students had not yet received formal reading instruction at this point, the main purpose was to identify early readers. Testing was stopped if a student gave up on (or misread) three consecutive words. Fast reading was not encouraged, and reading speed was not recorded. The 32 words were divided between four increasingly demanding lists ranging from short regular words (e.g., *ko 'cow', hat 'hat*) to longer and somewhat irregular words (e.g., *frugt 'fruit', hjerte 'heart'). The score was the percentage of correctly read words. Cronbach’s alpha was .94.

**Grade 1 & 2 oral word reading accuracy.** This was an expanded version of the Grade 0 measure. The students were asked to read a total of 104 words as accurately and fast as they could. The words were divided equally between two booklets each containing lists of four words. The lists were increasingly demanding, ranging from short regular words such as *ko 'cow* and *hat 'hat* (six lists with a total of 24 words) to longer and somewhat irregular words such as *slikkepind 'lollipop* and *gyngestol 'rocking chair* (four lists with a total of 16 words). Testing was stopped if all items on a list were misread, or if three consecutive lists were read with only one or two correct items. The score was the percentage of correctly read words. The correlation between scores on the two booklets was *r* = .96 in Grade 1.

**Grade 1 & 2 oral word reading fluency.** This measure was based on the first 24 short regular words (presented in lists of four and divided between two booklets) in the word reading accuracy task. To ensure comparability, the fluency measure was confined to these relatively easy words lists because they were attempted by all students before any stop criteria applied. The completion time for each word list was measured from the recordings using the Praat speech
adjusting sensitivity to accuracy analysis software (Boersma, 2001). The score was the number of correctly read words pr. minute. The correlation between scores on the two booklets was \( r = .93 \) in Grade 1.

**Classification of reading difficulty in Grade 2**

Juul et al. (2014) found that accuracy and fluency development appear to depend on different prerequisites. Therefore, the students were categorized as having difficulties separately on accuracy and fluency measures. Students who scored among the lowest 15 percent on the number of correctly read words were considered as having accuracy difficulty. This corresponds to present practice in early grades in Denmark since it has been estimated that 15% of Danish Grade 3 students receive some sort of reading intervention (Møller et al., 2014). The percentage drops to 9% in Grade 5 and 6% in Grade 9. The cut-off score was 72% correct. Students without accuracy difficulties achieved this accuracy level between May of G1 (\( M = 69\%, SD = 23\% \)) and September of G2 (\( M = 78\%, SD = 24\% \)), thus the cut-off approximately represented a one-year lag behind the non-disabled mean. Likewise, students who scored among the lowest 15 percent on the number of correctly read words pr. minute were considered as having fluency difficulty. The cut-off was 45 correct words pr. minute. Students without fluency difficulties on average achieved this level of efficiency between March of G1 (\( M = 36, SD = 26 \)) and May of G1 (\( M = 50, SD = 31 \)). Again, the cut-off approximately represented a one-year lag behind the non-disabled mean.

Twenty-four students were considered as having accuracy difficulties and 24 students were considered as having fluency difficulties. Sixteen students were labeled as having both accuracy and fluency difficulties.
Results

A Little’s MCAR test on the full data set did not suggest that missing data were non-random ($p > .05$). Therefore, only data from the 164 students who completed all tasks were used in the following analyses.

The descriptive statistics and zero-order Spearman correlations between the variables are displayed in Table 1. We report Spearman correlations because the relationship between the accuracy and fluency measures were highly curvilinear due to many students having reached ceiling accuracy in Grade 2. This curvilinear nature of these correlations is not a problem for the logistic regression models to follow, where the outcome measures have been dichotomized.

[Insert Table 1 here]

All correlations were significant with the exception that paired associate learning did not correlate with RAN.

To predict Grade 2 reading difficulty, we constructed a number of logistic regression models with either accuracy difficulty or fluency difficulty as the dependent variable. The models are reported separately in Table 2. Early complex models used all the measures from May of Grade 0 as predictors. Late complex models used all the precursor predictors and the reading fluency measure from Grade 1 as predictors. Reading fluency was used to predict both accuracy and fluency difficulty on the grounds that the fluency measure included both accuracy and fluency information. To investigate whether late screening could be simplified, a final set of late, simple models used only reading fluency from Grade 1 as predictor.

[Insert Table 2 here]
Our first question was whether identification improved from early to late screening. The overall identification accuracy of the different models was compared with receiver operating characteristic (ROC) analyses. For accuracy difficulty, the AUC of the late models (.945 and .940) were significantly larger than for the early complex model (.851), DeLongs test for correlated ROC curves, complex: $z = -3.18, p < .01$; simple: $z = -2.81, p < .01$. For the models predicting fluency difficulty, the late complex model was significantly more accurate than the early model (AUC .871 vs .804, $z = 2.04, p = .04$), but the late simple model was not (AUC = .863, $z = 1.31, p = .16$). However, the small difference between the two late models were not significantly different ($z = 0.58, p = .56$). In summary, postponing screening until after instruction yielded reliably improved identification accuracy when reading difficulty was defined in terms of reading accuracy, but the results were mixed when difficulty was defined in terms of fluency.

Inspection of the individual model coefficients provides information on the predictive power of the precursor measures compared to actual reading measures. Before reading instruction, there was evidence that the precursor measures were stronger predictors than real reading measures. Table 2 shows that in the early complex models, RAN made unique contributions to the identification of both accuracy and fluency difficulty ($p < .05$ and $p < .001$). Paired associate learning made unique contribution to the identification of accuracy difficulty ($p < .01$). Letter knowledge made marginally significant contributions to identification of fluency difficulty ($p = .08$). But budding oral word reading in Grade 0 only made marginal unique contributions to the identification of accuracy difficulty ($p = .06$) and no unique contribution to the identification of fluency difficulty ($p > .10$). This is not to say the Grade 0 reading was not a predictor of future reading difficulty, only that it did not contribute beyond the other measures in the logistic
regression models. In contrast, Table 2 also shows that in the late complex models mid Grade 1 reading obviated the Grade 0 predictors in the identification of both accuracy and fluency difficulty; no other predictors made significant contributions (all other p’s > .10). This picture supports the above results showing that the AUC’s of the late complex models were not significantly different from the late simple models, confirming that none of the Grade 0 measures contributed to identification accuracy beyond Grade 1 reading.

To summarize, the results indicated that simple reading measures improved as predictors relative to Grade 0 precursor measures over the course of six months of institutional reading instruction, to the point where they obviated the precursor measures. For the accuracy definition of reading difficulty, the total identification accuracy improved reliably. For the fluency definition of reading difficulty, the total screening accuracy did not improve reliably.

**Explorative analyses of identification accuracy at different levels of sensitivity**

The above analyses provide statistics on the relative accuracy of pre- and post-instructional screening. Now we turn to evaluating the identification accuracy of the screening methods in more practical terms. To this end it is necessary to inspect the specifics of the identification accuracy. Table 3 presents AUC and the associated false positive rate and total positive rate of the screening tools at different levels of sensitivity. The proportion of overlooked children (false negative rate) is also of obvious concern, but can be computed as 1-sensitivity. Specific numbers should be interpreted with caution since they are point estimates without confidence intervals, but they convey the best estimates for use in an evaluation of the general practicability of different screening strategies.

[Insert Table 3 here]
Table 3 shows that correctly identifying 80% of the students with accuracy difficulty from the Grade 0 results entailed incorrectly assigning at-risk status to 29% of those who in fact did not turn out to have accuracy difficulty (i.e. the false positives). In total, 37% of the sample would be flagged as in need of attention (i.e. the total positive rate) at this level of sensitivity. Since reading difficulty was defined as scoring in the bottom 15 percent on the Grade 2 outcome measure, thus the optimal positive rate was 15%.

If the Grade 0 sensitivity was lowered to 60%, the false positive rate and the total positive rates improved to 14% and 21% respectively. But of course, this means overlooking 40% of those students who turned out to have accuracy difficulty.

By following a late screening strategy and postponing screening to mid Grade 1, 80% sensitivity could be achieved with a false positive rate of 7% and the total positive rate of 18%.

Table 4 presents the corresponding numbers for identifying students who turn out to have fluency difficulty. The table shows that correctly identifying 80% of the students with fluency difficulty in Grade 0 entailed a false positive rate of 34%, and a total positive rate of 41%. It should be kept in mind, that the above comparisons of AUC’s found no difference between the two late models. Thus, the differences in the detailed statistics between the two late models should be not be considered reliable and will not be commented further upon.

[Insert Table 4 here]

**Discussion**

The present paper reported a study that predicted reading difficulties at the end of Grade 2 with test results from the end of Grade 0 and the middle of Grade 1. Identification of future reading
accuracy difficulties improved from the end of Grade 0 to the middle of Grade 1, after formal reading instruction had set in. Furthermore, the improved prediction accuracy in Grade 1 could be achieved by simply measuring reading ability, compared to using a time-consuming battery of tests in Grade 0. When reading difficulties in Grade 2 were defined in terms of reading fluency, identification accuracy was poorer and only improved marginally from Grade 0 to Grade 1, but the same identification accuracy could be achieved with a single reading measure. The result that early prediction is less accurate than later prediction is not surprising given the literature on prediction of reading. But follow-up analyses explored different ways dealing with the lower early identification accuracy. In the following, we discuss how such information can guide screening strategies when screening accuracy is less than perfect.

The usefulness of a screening procedure is dependent on what follow-up actions that the procedure allows (Glover & Albers, 2007; Messick, 1995). Low screening accuracy means that there is a tough choice between finding most of those who need attention while limiting the number of children flagged for attention in order to have the resource to provide them with attention. Authors have recently focused on finding a minimum of 75-90% of those that need attention (Glover & Albers, 2007; Johnson et al., 2009). To meet an 80% sensitivity standard with early screening before reading instruction one third of a year group would be flagged for attention in the sample of the present study where we defined reading difficulty as the bottom 15%. In Denmark, 15% corresponds to the present proportion of students in Grade 3 who receive reading intervention (Møller et al., 2014). A proportion as high as one third could rapidly deplete the teaching resources of schools. As far as we can tell, the identification accuracy of the present screening battery is no worse than what has been found in several similar recent studies of early predictive validity (Catts et al., 2009; Jenkins, Hudson & Johnson, 2007; Johnson et al., 2009;
Sittner Bridges & Catts, 2011). Some studies have in some cases demonstrated better results when using more comprehensive testing (Catts et al., 2015; Compton et al., 2010; O'Connor & Jenkins, 1999), but dramatically better results than the present cannot be guaranteed. For example, Catts et al. (2015) reported improvements of AUC’s from .86 to .90 on two out of three outcome measure by adding progress monitoring to precursor measures taken in kindergarten (improvement from .92 to .95 on the third outcome measure).

If comprehensive testing is not a practically viable option, easy solutions to the problem with high positive rates are to either postpone screening or to lower the sensitivity of early screening. Postponing screening would improve positive rates of identification of accuracy disability from 37% to 18% with the same sensitivity level of 80%. Another solution is to lower sensitivity from 80% to 60% in early screening, which would improve positive rates from 37% to 21%. However, these solutions come at the cost of either postponing intervention or overlooking 40% of the students who need attention in Grade 0.

One way of dealing with the imperfections of the easy solutions could be to combine the two solutions in an increasing sensitivity strategy. Early screening with low sensitivity would enable early intervention for, say, 60% of those who need it, and most of the remaining students with reading accuracy difficulty would be identified half a year later when the identification accuracy of simple reading measures has improved to allow higher sensitivity with manageable positive rates. Thus, the idea of the increasing sensitivity strategy is to take advantage of the fact that screening accuracy increases over time by tailoring the sensitivity of the screening to the identification accuracy at the time of the test. This may be in line with what already happens in a lot of practical settings, where the amount of resources determines the number of students that are offered intervention. But an explicit increasing sensitivity strategy may help devise series of
test points that strike an informed balance between early intervention and having the resources to carry out the intervention. Although multiple test points imply added costs for testing, the results of the present study indicate that, after the onset of instruction, screening can be simplified to relatively simple measures of reading.

We have emphasized the total positive rate as a useful target statistic for organizing screening in practice. It should be kept in mind that the positive rate is dependent on sample characteristics. First, the positive rate is influenced by the base rate of reading difficulty. The total positive rate will approach the base rate as identification accuracy approaches 100%. Thus, the total positive rate will tend to be high if the base rate is high. In the present study, the base rate was set to match estimates of how many students receive intervention in Danish practice, but practice may differ (Ise et al. 2011). In other screening studies, base rates have varied quite substantially from 3% (Compton et al., 2006) to 20% and 40% (Johnson et al., 2009), which probably reflects that studies differ in who they want to identify, for example either students who can be expected to experience academic difficulties, or students who could benefit from some sort of additional attention.

Second, studies that oversample students with difficulties should take care estimate the population total positive rate, for example, with weighting procedures, because the sample total positive rate cannot be expected to reflect the population rate when the base rates of reading difficulty differ between the sample and the population.

The present study found that identification accuracy was substantially lower for reading difficulties when they were defined in terms of fluency compared to accuracy (AUC of .86-.87 compared to .94-.95). In a Finnish study, Puolakanaho et al. (2008) also found that Grade 2 reading fluency was harder to predict than reading accuracy from precursor measures. One
possible explanation is that reading fluency development – which may be largely distinct from accuracy development (cf. Juul et al., 2014) – has only just begun for many students in Grade 2. It might thus be too early to consider fluency an outcome measure by the end of Grade 2. Many students may not have had the chance to develop the speed aspect of fluency, and thus low speed may not yet be a stable sign of a learning difficulty. Therefore, the above discussion has concentrated on the identification of accuracy difficulty.

**Limitations**

It is a general limitation of this kind of prediction study that it probably underestimates the predictive validity of the screening measures because the participating schools and teachers already intervene to some degree based on impressions of student ability, whether it comes from formal testing or teacher observation. In the predictive analyses, successfully remediated students would incorrectly look as if they did not need the attention, that is true positives may look like false positives. Thus, poor ROC statistics could stem from either poor screening measures or successful instructional efforts. This affects true positive rates and false positive rates, but total positive rates should be unaffected because they simply reflect the cut-off on the screening battery.

The screening time point in the middle of Grade 1 was motivated by the idea that during the first half year of Grade 1, all of the students have an opportunity to learn to read, and that this increases the predictive accuracy of reading measures. There is, however, nothing in the design that rules out that it is simply the shortening of how far into the future the prediction reaches that explains the improved accuracy.

Likewise, the results suggest that reading improves dramatically as a predictor over the first six months of systematic instruction compared to precursor measures. But the precursor
measures were only taken in Grade 0 and thus did not benefit from the shortening of the prediction span or the opportunities to learn during Grade 1. Some longitudinal studies have found unique contributions beyond reading (autoregressors) of precursor measures such as phonological awareness or RAN (Manis et al., 1999; Parrila et al., 2004; Wagner et al., 1997), so it is possible that adding these measures to the Grade 1 test battery might have improved the prediction.

Finally, there is a strong argument that there is more to reading difficulty than word identification difficulties, which was the focus of this and most other screening studies. Some students experience difficulties with reading comprehension that are not attributable to word identification (Bishop & Snowling, 2004; Catts, Adlof, & Weismer, 2006). A complete screening for reading difficulties should be able to identify separate comprehension difficulties, and it would probably require predictor measures that are sensitive to such difficulties, such as vocabulary, grammar comprehension, and inference tests (Adlof, Catts, & Lee, 2010; Oakhill & Cain, 2012).

Conclusion
Early screening before instruction tends to be imprecise. Rather than simply postponing screening and intervention to a point in time where screening is more accurate, an increasing sensitivity strategy (ISS) with a combination of early screening with low sensitivity and later screening with high sensitivity provides a way of focusing intervention resources on the students who need it the most - at the earliest practical opportunity. We do not claim that the strategy is preferable to highly structured programs such as Response to Intervention, and further research into improving early identification accuracy is important within any framework. But the ISS is a possibility in school systems that are not prepared to provide the comprehensive testing and
intervention of RTI. Intervention is still needed, but can take the form of what is appropriate under the given conditions. For example, if false positive rates are high, it may be advisable that intervention take the form of special in-class instruction and activities for a group of students, in part to avoid stigma, rather than visible out-of-class activities. In other words, an informed ISS strategy may be preferable to no screening at all.

References


Danish Ministry of Education (2014). Læseplan for faget dansk [Curriculum for Danish]. Retrieved from The Danish Ministry of Education website:

http://www.emu.dk/sites/default/files/Dansk%20Læseplan_0.pdf


doi:10.1177/0022219410374235


doi:10.1037/a0037100


doi:10.1016/j.jml.2013.10.005


doi:10.1207/s1532799xssr0801_2


Tables

Table 1. Means, Standard Deviations and Spearman Correlations.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>G0 RAN</td>
<td>1.02</td>
<td>0.23</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>G0 Letter knowledge</td>
<td>.85</td>
<td>.18</td>
<td>.27</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>G0 Phoneme awareness</td>
<td>.35</td>
<td>.28</td>
<td>.22</td>
<td>.41</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>G0 Paired associate learning</td>
<td>32.46</td>
<td>6.97</td>
<td>.07</td>
<td>.29</td>
<td>.37</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>G0 Reading aloud acc</td>
<td>.10</td>
<td>.16</td>
<td>.20</td>
<td>.60</td>
<td>.66</td>
<td>.27</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>G1 Reading aloud acc</td>
<td>.40</td>
<td>.27</td>
<td>.40</td>
<td>.50</td>
<td>.56</td>
<td>.42</td>
<td>.61</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>G1 Reading aloud fluency</td>
<td>23.14</td>
<td>20.46</td>
<td>.50</td>
<td>.46</td>
<td>.49</td>
<td>.38</td>
<td>.50</td>
<td>.89</td>
<td>-</td>
</tr>
<tr>
<td>8.</td>
<td>G2 Reading aloud acc</td>
<td>.88</td>
<td>.20</td>
<td>.42</td>
<td>.40</td>
<td>.46</td>
<td>.38</td>
<td>.44</td>
<td>.75</td>
<td>.79</td>
</tr>
<tr>
<td>9.</td>
<td>G2 Reading aloud Fluency</td>
<td>85.53</td>
<td>36.28</td>
<td>.49</td>
<td>.30</td>
<td>.24</td>
<td>.30</td>
<td>.26</td>
<td>.61</td>
<td>.75</td>
</tr>
</tbody>
</table>

Note. *r* > .15 are significant at the .05 level. *r* > .19 are significant at the .01 level.
Table 2. Logistic Regression Models for Predicting Grade 2 Difficulty by Accuracy or Fluency Criterion.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy difficulty</th>
<th>Fluency difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$B$</td>
<td>$SE$</td>
</tr>
<tr>
<td>Model 1: Early complex</td>
<td>.851</td>
<td></td>
</tr>
<tr>
<td>G0 RAN</td>
<td>-0.593</td>
<td>0.291</td>
</tr>
<tr>
<td>G0 Letter knowledge</td>
<td>-0.093</td>
<td>0.231</td>
</tr>
<tr>
<td>G0 Phoneme awareness</td>
<td>-0.028</td>
<td>0.411</td>
</tr>
<tr>
<td>G0 Word reading acc.</td>
<td>-4.184</td>
<td>2.203</td>
</tr>
<tr>
<td>G0 PAL</td>
<td>-0.676</td>
<td>0.247</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.972</td>
<td>1.093</td>
</tr>
<tr>
<td>Model 2: Late complex</td>
<td>.945</td>
<td></td>
</tr>
<tr>
<td>G0 RAN</td>
<td>-0.003</td>
<td>0.361</td>
</tr>
<tr>
<td>G0 Letter knowledge</td>
<td>0.192</td>
<td>0.266</td>
</tr>
<tr>
<td>G0 Phoneme awareness</td>
<td>0.204</td>
<td>0.480</td>
</tr>
<tr>
<td>G0 PAL</td>
<td>-0.486</td>
<td>0.302</td>
</tr>
<tr>
<td>G1 Word reading fluency</td>
<td>-9.919</td>
<td>2.650</td>
</tr>
<tr>
<td>Constant</td>
<td>-8.637</td>
<td>2.121</td>
</tr>
<tr>
<td>Model 3: Late simple</td>
<td>.940</td>
<td></td>
</tr>
<tr>
<td>G1 Word reading fluency</td>
<td>-9.369</td>
<td>2.235</td>
</tr>
<tr>
<td>Constant</td>
<td>-8.139</td>
<td>1.857</td>
</tr>
</tbody>
</table>

Note. AUC = area under the curve.
Table 3. Accuracy of Identifying Students with Accuracy Difficulty at Different Levels of Sensitivity.

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>AUC</th>
<th>SE</th>
<th>FPR</th>
<th>PR</th>
<th>FPR</th>
<th>PR</th>
<th>FPR</th>
<th>PR</th>
<th>FPR</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model 1 (early) .85 .04 .14 .21 .15 .23 .29 .37 .33 .41
Model 2 (late complex) .95 .02 .03 .11 .06 .16 .08 .19 .20 .31
Model 3 (late simple) .94 .02 .04 .12 .04 .14 .07 .18 .25 .35

Note. AUC = area under the curve; SE = standard error of AUC; FPR = false positive rate; PR = total positive rate. 15% of the students were considered as having reading difficulty on the Grade 2 outcome measure.
Table 4. Accuracy of Identifying Students with Fluency Difficulty at Different Levels of Sensitivity.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sensitivity</th>
<th>AUC</th>
<th>SE</th>
<th>FPR</th>
<th>PR</th>
<th>FPR</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 (early)</td>
<td>.60</td>
<td>.80</td>
<td>.04</td>
<td>.23</td>
<td>.28</td>
<td>.29</td>
<td>.35</td>
</tr>
<tr>
<td>Model 2 (late complex)</td>
<td>.70</td>
<td>.87</td>
<td>.03</td>
<td>.10</td>
<td>.18</td>
<td>.18</td>
<td>.26</td>
</tr>
<tr>
<td>Model 3 (late simple)</td>
<td>.80</td>
<td>.86</td>
<td>.04</td>
<td>.11</td>
<td>.19</td>
<td>.17</td>
<td>.25</td>
</tr>
<tr>
<td>Model 4 (late simple)</td>
<td>.90</td>
<td></td>
<td></td>
<td>.34</td>
<td>.41</td>
<td>.36</td>
<td>.44</td>
</tr>
</tbody>
</table>

*Note.* AUC = area under the curve; SE = standard error of AUC; FPR = false positive rate; PR = total positive rate. 15% of the students were considered as having reading difficulty on the Grade 2 outcome measure.

**Figures**

Figure 1. Summary of identification accuracy statistics

<table>
<thead>
<tr>
<th>Outcome status</th>
<th>At risk</th>
<th>Not at risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor reader</td>
<td>True positive (TP)</td>
<td>False negative (FN)</td>
</tr>
<tr>
<td>Normally developed</td>
<td>False positive (FP)</td>
<td>True negative (TN)</td>
</tr>
</tbody>
</table>

*Sensitivity* = TP / poor readers (TP+FN).

*False positive rate* (1-Specificity) = FP / normally developed (TN+FP).

Total positive rate = TP + FP / Sample size