Springer, Early Childhood Education Journal

Manuscript Draft

Manuscript Number: ECEJ-D-17-00096

Title: Teacher's Ability to Identify Children at Early Risk for Reading Difficulties in Grade 1

Article Type: Research Paper

Keywords: assessment; teacher; pre-reading skills; at-risk students; sensitivity; specificity

Corresponding Author: Riitta Virinkoski

Corresponding Author's Institution: University of Jyväskylä

First Author: Riitta Virinkoski

Order of Authors: Riitta Virinkoski; Marja-Kristiina Lerkkanen; Leena Holopainen; Kenneth Eklund; Mikko Aro

Abstract: The aim of the study was to investigate what kinds of assessment practices class teachers and special educational needs (SEN) teachers use in assessing first grade students' pre-reading skills (letter knowledge and phonological skills). Further, we investigated to what extent teachers were able to identify difficulties in pre-reading skills of the lowest achievers. The accuracy of teacher ratings of students' pre-reading skills was studied by comparing teacher ratings to actual test scores. The data from two Finnish longitudinal studies were used: JLD sample (class teachers, n = 91; SEN teachers, n = 51; 200 students) and First Steps sample (class teachers, n = 136; SEN teachers, n = 34; 598 students). Results showed first, that most class teachers used qualitative assessment and SEN teachers also relied on tests. Secondly, although teacher ratings correlated with the test scores, closer investigation of sensitivity and specificity of the teacher ratings revealed that a number of children in need of extra support for their early reading development according to test scores remained unidentified. Moreover, there were some students identified by the teacher to have difficulties despite test scores not confirming that. The findings underline the importance for developing more specific and reliable assessment tools for teachers to use for pedagogical purposes, and respectively, the need to pay more attention to early identification of reading difficulties in teacher training program curricula.

Acknowledgements

This study has been carried out in the Centre of Excellence in Learning and Motivation Research, and financed by the Academy of Finland (No. 213486 for 2006–2011) and other grants from the same funding agency for the authors (Nos. 292466 for 2015–2019, 268586 for 2013–2017).

Teachers' Ability to Identify Children at Early Risk for Reading Difficulties in Grade 1

Teachers play a key role in identifying the need for early support in reading skill development because they generally observe the first signs of reading difficulties (RD; Bailey & Drummond, 2006; Compton et al., 2010). Previous studies have indicated that teachers' judgments of reading skills in kindergarten and at the beginning of school (first and second grade) generally correspond well with the scores of standardized reading achievement test results especially regarding the high-performing students (Bailey & Drummond, 2006; Begeny, Krouse, Brown, & Mann, 2011; Südkamp, Kaiser, & Möller, 2012) . The main purpose of teachers' evaluations of students should be to produce accurate knowledge of the students' skills in order to plan tailored instruction and support when necessary (Bailey & Drummond, 2006; Mesmer & Mesmer, 2008). Begeny et al. (2011) studied first- to fifth-grade students' oral reading fluency and found that accurate performance assessments could allow for providing early support, thereby preventing the need for intensive intervention. However, their findings revealed that it was difficult for teachers to judge students' reading levels as low-, average-, or high-performing. One explanation for low-judgement accuracy could be the lack of teacher training and practice in conducting assessments (Begeny et al., 2011).

Particularly, children with poor pre-reading skills who are potentially at risk for RD should be identified as early as possible. Early recognition of risk for RD would be needed to avoid prolonged or more serious problems. Flynn and Rahbar (1998) also indicate that researchers have disagreed on whether teacher ratings or screening tests best identify children at risk for reading failure. In support of screening instruments for early identification, prior studies have shown that screening batteries and standardized achievement tests predict those at risk for reading failure better than teachers' evaluations based on, for example, rating scales, whereby the latter have tended to produce high false-

negative rates (Fletcher & Satz, 1984; Flynn & Rahbar, 1998). Moreover, to lead to effective and early support for at-risk students, the screen must be relatively accurate, i.e., capable of distinguishing students who will subsequently have difficulties from those who will not (Johnson, Jenkins, Petscher, & Catts, 2009). Recent study by Catts, Nielsen, Bridges, Liu, and Bontempo (2015) indicated that using screening batteries containing measures of e.g., letter naming fluency and phonological awareness enabled accurate identification of good and poor readers at the end of first grade. It has also been shown that using teacher judgment more with the universal screening procedures could increase the classification accuracy rates of at-risk and not-at-risk students (Compton et al., 2010; Martin & Shapiro, 2011; Snowling, Duff, Petrou, & Schiffeldrin, 2011).

In addition, studies comparing class teachers' and special educational needs (SEN) teachers' assessments for identifying at-risk students are lacking. For example, both class teachers and SEN teachers in Finland have Master's Degree, and they receive different kinds of training to gain competence in identifying children at early risk for RD. SEN teacher training in Finland comprises theory and practice, related to individual and small-group instruction, e.g. application of various assessment tools, support in reading and writing, mathematics, and communication, but also in behavioral and socio-emotional challenges (Takala & Ahl, 2014). Respectively, Finnish class teacher training provides readiness to instruct a whole class within general education and adapt that instruction according to children's needs.

Consequently, the present study investigates class teachers' and SEN teachers' assessment practices and a matching of their ratings of pre-reading skills regarding especially the lowest achievers who have difficulties in letter knowledge and phonological skills. Further, we are interested in how well the teachers' ratings correspond to the test scores at the beginning of the first grade in the highly transparent Finnish orthography.

Assessment of Pre-Reading Skills by Teachers

Previous research indicates that the most common rationale for being identified as an at-risk student is problems in, for example, letter knowledge and identifying letter sounds. To ensure accurate identification, the screening batteries should cover several skill areas related to developing reading skills, such as phonological skills, orthographic and letter knowledge, word reading ability, vocabulary, and syntactic ability (Bailey & Drummond, 2006; Davis, Lindo, & Compton, 2007). To accurately classify students into two groups, at risk and not at risk for poor reading outcomes, it is important that the screens are targeted at reading skills, and that the content is age-appropriate (Jenkins, Hudson, & Johnson, 2007). However, the accuracy of screening measures differs with respect to sensitivity and specificity (Catts et al., 2015; Compton et al., 2010; Jenkins et al., 2007; Johnson et al, 2009). Sensitivity refers to the degree of true positives, meaning how accurately the measure identifies students at high risk for RD. Specificity, on the other hand, refers to the degree of true negatives, or how accurately the measure identifies students at low risk for RD. The fact that a test discriminates poor readers at the group level does not necessarily guarantee accurately predicting or identifying difficulties at the individual level (Puolakanaho et al., 2007). The quality of the predictor is determined by how well it is able to capture the true-positive cases that turn out to have RD at school age, and to avoid false-positive cases that predict risk for RD although the children do not have difficulties in reading at school age. According to the literature, teachers' assessment practices can be divided into three categories: tests comprising screening or individual test batteries, (performance-based assessment), curriculum-based measures (CBM), and qualitative assessments such as observations in the classroom (Bailey & Drummond , 2006; Südkamp et al., 2012). One way to assess student progress toward long-term curriculum goals in literacy learning is CBM, which is the main tool of screening difficulties learning difficulties and the risk for RD in the response to intervention (RTI) framework (Stecker, Fuchs, & Fuchs, 2005; Deno, 1985). CBM may be used to monitor students' progress in an entire school or classroom, to track an

individual's progress toward end-of-year benchmarks or individualized education program goals, or to screen students at a specific time point to determine their level of risk for academic failure (Deno, 1985, 2003; Madelaine & Wheldall, 2005; Zumeta, Compton, & Fuchs, 2012).

A number of previous studies (e.g., Bailey & Drummond, 2006; Beswick, Willms, & Sloat, 2005) have shown that teachers' evaluations and their perceptions of a student's risk for literacy failure can be used as early as the beginning of kindergarten and the first grade to identify signs of RD. In Bailey and Drummond's (2006) study, kindergarten and first-grade teachers were asked to identify one to four students in their class who they perceived to be at risk for RD, but who were not receiving any formal remediation at the moment. They used literacy development checklists (LDC; Bailey et al., 2001) and also concept maps based on targeted early literacy skills, such as decoding, letter-sound correspondence and phonemic awareness. However, according to Bailey and Drummond (2006), the data teachers rely on when rating students' reading performance may not allow for making accurate judgments of particular pre-reading skills. Teachers' decisions seemed to be sometimes based on situational or other irrelevant factors (e.g., gender, behavior, students' ability to work in groups), instead of solely performance assessments (Beswick et al., 2005). They might also have insufficient knowledge or competence to identify students' RD (Bailey & Drummond, 2006). In addition, Bailey and Drummond (2006) noted that some teacher characteristics, such as years of teaching experience and personality, affect the accuracy of teacher judgments. Furthermore, teachers have been shown to have a tendency to underestimate the reading skills of those students who have had prior weaknesses in reading, and whose general cognitive skills are at a low level in combination with previously identified SEN (Soodla & Kikas, 2010).

Correspondence between Teacher Ratings and Test Scores

In most studies, the correlations between teacher ratings and test scores have varied between .40 and .70. For example, Südkamp et al.'s (2012) meta-analysis on teachers' judgment accuracy in a regular school system, from kindergarten through grade 12 over a 20-year period, indicated that the correlation between teacher judgments of students' academic achievement in language arts (reading, spelling, literature, and composition) and mathematics, and their actual test performance

was moderate, at .63. Their findings are in line with Hoge and Coladarci's (1989) study that investigated language arts, mathematics, and the social sciences, where the median correlation of .66 was reported. In their study, correlations between teacher judgments and the standardized tests ranged from .41 to .92.

However, in Bailey and Drummond's (2006) study, the correlations between teacher ratings and standardized tests regarding the emergent literacy/basic reading skills domain (e.g., print and graphic presentations) and the phonological awareness subskill of kindergarten and first-grade students were weak and not significant. They came to the conclusion that the low correlation resulted from the array of informal assessment procedures that teachers used, such as combining old curricular material with current material and observational information with in-class tests.

Despite relatively high correlations between teachers' evaluations and children's actual test scores, teachers may, however, systematically over- or underestimate student performance (Bates & Nettlebeck, 2001). Another salient limitation of teacher judgments may be revealed when the range of student competence is restricted, particularly regarding students who show low-academic performance (Graney, 2008; Soodla & Kikas, 2010). Teachers' judgments may also be related to some personal characteristics, such as their skills, training, future expectations, or perceptions (Kikas, Silinskas, & Soodla, 2015; Soodla & Kikas, 2010).

Flynn and Rahbar (1998) developed a theory-based screening instrument for teachers to assess reading competency, and their results suggest that teachers' predictions of children at risk for RD can be improved by using rating instruments that include research-validated antecedents of reading with behavioral descriptions of low and high achievement (Flynn & Rahbar, 1998). Further, in their study, Bailey and Drummond (2006) found that by using a literacy checklist, teacher evaluations can become more systematized and also lead to a higher identification rate of at-risk students.

The best predictors of a preschooler's or kindergartener's later reading achievement when the child has a familial history of dyslexia have turned out to be measures that require processing printed material, together with oral language proficiency measures and performance-IQ measures (Puolakanaho et al, 2007). Most studies evaluating the accuracy of teachers' judgments have used standardized tests as the comparative criterion for this investigation. Fletcher and Satz (1984) and Flynn and Rahbar (1998) found in their studies that compared to teacher ratings, standardized tests more accurately identified students who were potentially at risk for RD in the future. Teacher ratings usually had high false-negative rates and low true-positive rates. Flynn and Rahbar (1998) also found that combining class teacher ratings and screening tests in the first, second, or third grades increased the accuracy of identifying students who would experience reading failure in the future, with a correct identification percentage of 88%. In the same study, kindergarten teachers only used a traditional rating scale to predict future reading achievement, and the positive identification rate was rather low (30%). However, in this same study, using a project-developed, theory-based screening battery, the class teachers correctly identified 81% of poor readers. The prediction rate of the teacher ratings improved after some research-validated changes had been made, but remained below the identification accuracy of the screening test.

Learning to Read in Finnish

Finnish children attend kindergarten at age 6, and reading instruction begins at age 7 when they enter first grade. Upon entering school, letter knowledge seems to be one of the best predictors of reading and spelling accuracy in the Finnish language (Holopainen, Ahonen, & Lyytinen, 2001; Lerkkanen, Rasku-Puttonen, Aunola, & Nurmi, 2004). Also, phoneme identification and pseudoword repetition at school entry predict the development of accuracy in reading and spelling (Aro, 2006). The Finnish orthography is almost purely phonemic: the grapheme-phoneme correspondences are regular and symmetrical at the level of the single letter, and early reading instruction in Finnish is almost uniformly rests upon synthetic phonics (Aro, 2006; Seymour et al., 2003). In transparent orthographies, such as Finnish, the process of learning to decode accurately is rather fast (Seymour, Aro, & Erskine, 2003), and that might make the early identification of risk for RD, manifested mostly as problems in reading rate, even more challenging for teachers. Studies have shown that approximately 30% of Finnish children are able to decode before entering the first grade (Soodla et al., 2015), and highly accurate decoding skills are usually acquired within the first months of reading instruction (Lerkkanen et al., 2004). Even the nonreaders at school entry reach the level of early readers in reading accuracy during the first school year (Lerkkanen et al., 2004; Parrila, Aunola, Kirby, Leskinen, & Nurmi, 2005; Soodla et al., 2015). However, students whose growth is slow for letter knowledge and phonological awareness could encounter RD at the beginning of school (Lyytinen et al., 2006; Torppa, Poikkeus, Laakso, Eklung, & Lyytinen 2006). Additionally, a study identified a group of children with problems in phonological decoding in the end of the second grade, who remained behind their peers in reading accuracy still by grade 8 (Eklund, Torppa, Aro, Leppänen, & Lyytinen, 2015). In general, studies have shown that Finnish students who struggle with reading do not typically have problems with reading accuracy, but do experience persistent problems with reading fluency (Hintikka, Landerl, Aro, & Lyytinen, 2008). In the case of RD, the forms of support are remedial teaching during or after school by the class teacher, part-time special

education given by the SEN teacher individually or in small groups during school days, or co-teaching by the class teacher and the SEN teacher during normal literacy lessons (Lerkkanen, 2007). However, these forms of support do not require any formal diagnosis of a reading difficulty (Björn, Aro, Koponen, Fuchs, & Fuchs, 2016).

The Aim of the Present Study

The aim of the study was to investigate teachers' evaluation practices, and the sensitivity and specificity of their assessments of pre-reading skills, especially of the lowest achievers, and further, how the teacher ratings correspond to the reading test scores at the beginning of the first grade in the highly transparent Finnish language. By using two different samples, we intended to obtain a diverse overview of the assessment practices, as well as identification of children at early risk for RD, performed by regular class teachers as well as SEN teachers at the time of data collection. The research questions were as follows: (1) Which assessment practices do class teachers and SEN teachers use to assess pre-reading skills (e.g., letter knowledge, phonological skills) at the beginning of grade 1? According to previous studies (e.g., Bailey & Drummond, 2006; Beswick et al., 2005; Compton et al., 2010), we expected that teachers use screening batteries, CBM, and observation when assessing pre-reading skills (Hypothesis 1). Also expected were variations between the practices used by class teachers and SEN teachers.

(2a) Are teacher ratings associated with test scores in pre-reading skills? We expected the teacher ratings to correspond quite well with test scores (Hypothesis 2a, see e.g., Graney, 2008; Hoge & Coladarci, 1989).

(2b) How accurately do the teachers identify students' pre-reading difficulties to test scores, and what are the sensitivity and specificity rates of their assessments? According to previous studies (Fletcher & Satz, 1984; Flynn & Rahbar, 1998), we expected that teacher ratings

would have had high-false negative rates, and on the other hand, low true-positive rates in identifying at-risk students for reading. (Hypothesis 2b).

Method

The data for this study were drawn from two Finnish longitudinal studies. In both studies, parents and teachers were asked for written consent for the child's and their own participation in the study. In the Jyväskylä Longitudinal Study of Dyslexia (JLD), only the responses of SEN teachers were available concerning RQ1, but in the First Steps sample, we had the opportunity to study both class teachers' and SEN teachers' responses. Regarding RQ2, in the JLD sample, both class teachers' and SEN teachers' assessments were gathered, whereas in the First Steps sample, only SEN teachers' responses were available.

JLD Sample

Participants and procedure. In this study, we used the data from the fall of the first grade, and the data comprised class teachers (n = 91), SEN teachers (n = 51), and 200 first-grade students (*M* age = 7.19 years, SD = 0.26; 47% girls, 53% boys). The student data comprises four successive age cohorts born between 1993 and 1996, and half of the students had a familial risk of dyslexia, and the other half belonged to the control group. The at-risk children were defined by parents' self-reports of literacy difficulties and their reports of similar problems among their immediate relatives. The parents were sent a questionnaire that dealt with demographic information, and the occurrence of reading and writing difficulties during childhood, adulthood, and among relatives. In diagnostic tasks of reading and writing, the parents selected in the dyslexic sample had to obtain a -1 or less z-score in either accuracy or speed of oral text reading or spelling accuracy. Also, they had to obtain z-scores of -1.0 or less in two or more out of eight computer-aided measures. The children belonging to the control group or low-risk group did not have any reported parental literacy

difficulties nor in their first- or second-degree relatives (for more specific details, see Leinonen et al., 2001).

The research data consisted of teachers' questionnaires, teachers' student ratings, and test scores regarding first-grade students' letter knowledge and phonological skills upon entering school. The majority of the teachers' observation forms were returned in December and some after a reminder in February. Students participated in the individual tests in August.

Measures

Special education teachers' assessment practices. SEN teachers reported their reading assessment practices by responding to an open-ended question in the SEN teacher's observation form: "*What kinds of practices have you used in your assessment of learning to read and write and what sub-skills have you assessed*?" Nineteen (37%) out of 51 SEN teachers responded to this question. The teachers' written responses of assessment practices were classified in the following categories: 1) qualitative assessment (e.g., observation, check-lists, or discussions); 2) CBM (e.g., tests of ABC books or spelling from dictation); 3) reading tests, such as screening or individual tests; and 4) "Other," comprising teachers' self-developed assessment tools. Class teachers were not asked to report their assessment practices.

Teachers' ratings concerning children's reading and pre-reading skills. Class teachers and SEN teachers rated the students' reading performance on a five-point scale: 1 = "clear problem"; 2 = "mild problem"; 3 = "masters the skill"; 4 = "masters the skill quite well"; and 5 = "masters the skill very well." For this study, we selected three pre-reading skills rated by both class teachers and SEN teachers: letter naming, initial phoneme identification, and blending three sounds**.** The categories "clear problem" and "mild problem" were pooled together for the analyses; in addition, the categories "masters the skill," "masters the skill quite well," and "masters the skill very well" represented in the

analyses that the student did not have any problem with the skill from the teachers' point of view.

Letter knowledge. The letter-naming task was administered as an individual test in August. Twenty-nine uppercase letters were shown to the student by a trained tester on a sheet of paper in a fixed order, and the student was asked to name them as accurately and quickly as possible. The score was based on the number of correct responses.

Phonological skills. The phonological skills were assessed with two individual tasks drawn from Diagnostic Test Battery 1 (Poskiparta, Niemi, & Lepola, 1994). In the initial phoneme identification task, the trained tester first said the word to the student, and then the student said which sound was at the beginning of the word. In the phoneme-blending task, the experimenter said altogether 10 word items, phoneme by phoneme, and the student was instructed to say the resulting word. The sum scores were based on the number of correct items. Cronbach's alpha of phoneme identification was .94. and phoneme blending was .57. To enhance the reliability of the phonological awareness task, these variables were merged, and the mean of the two tasks was .80.

First Steps Sample

Participants and procedure. In this study, the data comprised class teachers (N = 136; *M* age= 42.69 years, SD = 9.1; 91% female, 9% male), SEN teachers (N = 34; *M* age = 45.62, SD = 9.60; 97% female, 3% male), and a subsample of 598 children selected for more intensive follow-up (47% girls, 53% boys) from four municipalities participating in the study in the fall of the first grade. The children were 7 years old at the beginning of the first grade (beginning of school; *M* = 7.18 years, SD = 0.30). The large majority of the class teachers (78%) had a master's degree in education from a class teacher program (69%). The rest had a master's degree in either special education (5%) or both programs (4%). One percent of the class teachers had a doctoral degree in education. Twenty-one percent (21%) had some other degree, typically a bachelor's (BA) degree, which was formerly sufficient for the

qualification as a class teacher or kindergarten teacher. The basic education for SEN teachers was a master's degree from a class teacher program combined with an SEN teacher qualification (53%), a master's degree from an SEN teacher program (44%), or something else (3%), such as a BA degree as a kindergarten teacher. Two class teachers and two SEN teachers did not report their education.

The sample of the present study contained both students identified as at risk (n = 277) and not at risk (n = 321) for RD. From the total sample of 1,880 children, the children's risk for reading problems was determined by the researchers at the end of the kindergarten year on the basis of four criteria: children's initial phoneme identification (indicator of phonological awareness), letter knowledge, rapid automatized naming, and parental report of their own RD (see Lerkkanen, Ahonen, & Poikkeus, 2011). The risk for RD was defined as a joint occurrence of at least two criteria out of three: low phonological awareness (i.e., scored clearly below age level in initial phoneme identification, ≤ 15%); poor letter knowledge (≤ 15%); and poor rapid automatized naming (≤ 15%; Kiuru et al., 2013; Lerkkanen et al., 2011). Furthermore, if parents reported having reading disabilities, a score below the 15th percentile in one of the three tests (phonological awareness, letter knowledge, or rapid automatized naming) was sufficient for identifying a risk for RD. The control children were randomly selected from the same classrooms as the children identified as being at risk for RD. The criteria resulted in one to six (typically two or three) children from each participating classroom being included in the more intensive follow-up. One to five from a maximum of six children were from the at-risk group (depending on the number of at-risk children in the classroom in each case), and the remainder were from the no-risk group.

SEN teachers were sent a list of the students who were followed more intensively in their school, but they did not know which group (at risk for RD or not) the individual children belonged to. They were asked to rate all the students who had received part-time

special education during the first grade in that particular school, irrespective of the reason for support (e.g., speech therapy, reading, math, behavioral problems) by December. In some cases, if the number of students exceeded six, the SEN teacher was allowed to select the students for the rating (usually students who needed more intensive support were selected).

The individual and the group tests at the beginning of the first grade were carried out in September. If the student was absent for the tests in September, the tests were implemented in October.

Measures

Class teachers' and SEN teachers' assessment practices. In December, both class teachers and SEN teachers reported the assessment practices they used with their students by answering the question on the SEN teachers' questionnaire "*How has the need for special educational support been defined?* (*What was assessed?/How was it assessed?/When did the assessment take place?)?"* The teachers' responses were classified in three categories similarly to the JLD sample: qualitative assessments, CBM, and tests.

Ratings by SEN teachers. SEN teachers were asked to evaluate their students' school entry pre-reading (e.g., letter knowledge) skills in December by filling in questionnaires concerning individual children. They rated the students' pre-reading skills using a three-point rating scale: 1 = "clear problem"; 2 = "mild problem"; and 3 = "no problem." Two variables were selected from the questionnaire for this study: letter naming and phonological skills (reading/spelling 3–4-letter syllables. The categories "clear problem" and "mild problem" were pooled together for the analysis because we were only interested in whether or not the student had difficulty from the teacher's point of view.

Letter knowledge. Letter knowledge was assessed in an individual situation using the ARMI test battery (Lerkkanen, Poikkeus, & Ketonen, 2006). The children were instructed to name 29 letters of the Finnish alphabet arranged randomly in three rows. The score was

the number of correctly named letters (max = 29). Cronbach's alpha for the naming letters task was .92.

Phonological skills. The phoneme-blending task (Poskiparta et al., 1994) was a group-administered test. The experimenter said words phoneme by phoneme, and after each word, the students were shown four pictures of objects, from which they had to choose the picture similar to the word formed by the phonemes. The score was the sum of correct items (maximum score 9). Cronbach's alpha was .70

Data analyses. The first research question was examined using descriptive statistics, and the analyses of the second research question were carried out using Spearman's rank-order correlation and cross tabulations. Next, we calculated the sensitivity and specificity scores of the teacher ratings in order to assess the overall accuracy of the teacher ratings with regard to identifying an early risk for problems. The cut-off score for low achievement in the test data was set to the lowest 15th percentile. Finally, logistic regression analyses were conducted to test statistically whether the teachers' ratings of students' pre-reading skills were significantly interconnected with the dichotomized students' test scores.

Results

Teachers' Assessment Practices

First, we examined the assessment practices teachers used to evaluate pre-reading skills (letter knowledge and phonological skills) at the beginning of grade 1.

JLD Sample

The number of assessment practices used by SEN teachers are summarized in Table 1. Most SEN teachers reported that they used only one type of assessment, qualitative or CBM being the most commonly used. If the SEN teachers used two types of assessment practices, they were usually tests combined with qualitative assessment. Further, when the

SEN teachers used three assessment practices, the most common combination was tests, qualitative assessment, and CBM. Altogether, tests were used by 47% of the SEN teachers.

Table 1

Number of Assessment Practices of the SEN Teachers in the JLD Sample (n = 19)

Number of assessment practices	n	%
One assessment practice	10	53
Two assessment practices	4	21
Three assessment practices	5	26
Total	19	100

First Steps Sample

Class teachers' and SEN teachers' assessment practices in the First Steps sample are summarized in Table 2. Assessment practices were unevenly distributed in the two groups (χ^2 (2, N = 51) = 6.57, $p < .05$). According to the standardized residuals class teachers used more often and SEN teachers less often than expected only one assessment practice (adjusted standardized residual for the cells = 2.4). Moreover, the use of two assessment practices was close to significant in favor of SEN teachers the adjusted standardized residual being 1.9. Most class teachers (58%) used qualitative assessment as their only practice. Further, when the class teachers assessed students using two types of practices, they were usually either CBM or tests combined with qualitative assessment. More than half of the SEN teachers relied on two types of assessment practices, most commonly tests combined with CBM. None of the class teachers and only 7% of the SEN teachers reported using all three types of assessment. Nearly 90% of the SEN teachers used tests in their assessment, either tests only, or tests with some other assessment practice. Whereas among the class

teachers, only 13% reported that they used tests as their only assessment practice or combined them with qualitative assessment.

Table 2

The Assessment Practices of Class Teachers and the SEN Teachers in the First Steps Sample

Number of assessment practices	Class teachers		SEN teachers	
	n	%	n	%
One assessment practice	17	70	10	37
Two assessment practices	7	30	15	56
Three assessment practices	0	0	2	7
Total	24	100	27	100

Association between the Teacher Ratings and the Test Scores

Regarding our second research question, we wanted to determine the associations between teacher ratings and reading test scores, especially of the lowest achieving students. Spearman's correlations (see Table 3 for the JLD sample and Table 4 for the First Steps sample) showed that associations between teachers' ratings and the test scores were moderate. The letter-knowledge task in the JLD sample correlated quite well (.52) between the SEN teachers' letter-knowledge ratings and the class teachers' phoneme-blending ratings. The best correlation in the First Steps sample appeared in the letter-knowledge task (.50).

Table 3

Correlations between Class Teacher and SEN Teacher Ratings and Test Scores in the JLD Sample

	Test scores *(n = 40–44)*	Test scores *(n = 34–35)*

	Letter knowledge	Phonological awareness[a]		Letter knowledge	Phonological awareness[a]
Class teachers' ratings			SEN teachers' ratings		
Letter knowledge	.42**	.41**	Letter knowledge	.52**	.35*
Phoneme identification	.43**	.33*	Phoneme identification	.45**	.46**
Phoneme blending	.52**	.30	Phoneme blending	.49**	.32

Note. [a] The phonological awareness test variable comprises the variables of initial phoneme identification and blending phonemes.

Note. **) Correlation is significant at the .01 level (two-tailed); *) correlation is significant at the .05 level (two-tailed)

Table 4

Correlations between the SEN Teacher Ratings and the Test Scores in the First Steps Sample

	Test scores	
SEN teachers' ratings	Letter knowledge *(n = 69)*	Phoneme blending *(n = 69)*
Letter knowledge	.50**	.17
Phoneme blending[a]	.24	.29*

Note. **) Correlation is significant at the .01 level (two-tailed); *) correlation is significant at the .05 level (two-tailed)

Note. [a] The phoneme-blending test corresponds to reading/writing 3–4-letter syllables.

Finally, we analyzed how accurately the class teachers and the SEN teachers were able to identify students at risk for reading failure (sensitivity), and on the other hand, those who were not at risk (specificity).

JLD Sample

In letter naming, the cut-off score used to indicate problems was 19 correct letters out of 29 in the individual letter-naming task, and 31 students scored below this score. In the phoneme-identification task, the cut-off score was 3 or fewer correct answers out of 10 phonemes, and there were 30 students in this group. Further, in the phoneme-blending task, the lowest achieving students had 1 or 0 correct responses out of 10 items in the test (the cut-off score), and there were 34 students in this group. Tables 5 (class teachers) and 6 (SEN teachers) present the true positives, the false negatives, the true negatives, and the false positives, according to the test scores. According to logistic regression analysis the class teachers' ratings of students' letter knowledge and the students' categorical test scores were close to significant (χ^2 (1) = 2.80, p = .09). In addition, class teachers' ratings of students' phonological awareness were not associated with their categorical test scores (χ^2 (1) = 0.90, p = .34 and χ^2 (1) = 1.20, p = .27 for phoneme identification and phoneme blending, respectively).

Table 5

Identification of Students at Risk for RD Based on the Class Teacher Ratings and the Test Scores in the JLD sample

	True positives	False negatives	True negatives	False positives
Pre-reading skill (n = number of students rated by class teacher)	n (%)	n (%)	n (%)	n (%)

Letter knowledge (n = 44)	4 (9)	9 (21)	28 (63)	3 (7)
Phoneme identification (n = 42)	2 (5)	10 (24)	28 (66)	2 (5)
Phoneme blending[a] (n = 42)	7 (17)	8 (19)	19 (45)	8 (19)

Note. [a] The phoneme-blending task corresponds to blending three sounds in the class teachers' ratings.

The sensitivity of class teacher ratings in letter knowledge was 31% and specificity was 90%, which means that 69% of the at-risk students remained unidentified, and 10% of the students with no difficulties were falsely identified as at-risk. In phoneme identification, the sensitivity rate was 17% and the specificity rate was 93%, which reflects the fact that, in general, teachers very rarely identified problems in phoneme identification. Finally, in phoneme blending, the sensitivity rate was 46% and the specificity rate was 70%, which indicates that the class teachers did not identify about half of the at-risk students; additionally, they identified 30% of the not-at-risk students as having difficulties in phoneme blending. The results indicate that it was very challenging for the class teachers to identify the difficulties, in general; albeit in phoneme blending, the ratings were more in line with the test scores. According to logistic regression analysis the SEN teachers' ratings of students' letter knowledge were associated with students' categorical letter knowledge test scores (χ^2 (1) = 5.6. p = .018). Regarding phoneme identification the SEN teachers' ratings and students' categorical test scores were close to significant (χ^2 (1) = 3.0, p = .08), and in phoneme blending the SEN teachers' ratings and the categorical test scores were not associated (χ^2 (1) = 2.3, p = .13).

Table 6

Identification of Students at Risk for RD Based on the SEN Teacher Ratings and the Test Scores in the JLD sample

Pre-reading skill (n = number of students rated by teacher)	True positives n (%)	False negatives n (%)	True negatives n (%)	False positives n (%)
Letter knowledge (n = 36)	10 (28)	8 (22)	15 (42)	3 (8)
Phoneme identification (n = 33)	6 (18)	8 (24)	16 (49)	3 (9)
Phoneme blending[a] (n = 35)	8 (23)	3 (9)	13 (37)	11 (31)

Note. [a] The phoneme-blending test variable corresponds to blending three sounds in the SEN teachers' ratings.

The sensitivity of SEN teacher ratings in letter knowledge was 55% and specificity was 83%, which means that about half of the at-risk students were identified, but also 17% of the not-at-risk students, according to the tests, were unnecessarily identified. In phoneme identification the sensitivity rate was 43% and specificity rate 84%. This shows that SEN teachers had difficulties especially in recognizing the at-risk students struggling with phoneme identification. As in phoneme blending, the sensitivity rate was 72% and specificity was 54%. These results indicate that the majority of the at-risk students were identified, but also the rate of unnecessarily recognized students was quite high. These results show that it was also challenging for the SEN teachers to identify at-risk students who had difficulties with phonological skills. However, the SEN teachers seemed to identify RD more than the class teachers, and somewhat more accurately. Nonetheless, they also missed most students who were having difficulties.

First Steps Sample

To score below the cut-off point for low achievement in letter knowledge, the student had to correctly name a maximum of 14 out of 29 letters, and there were 85 students in this group. The SEN teachers had rated 26 of those students in the letter-knowledge task. If in the phoneme-blending task, the student got a maximum of 5 correct answers out of 10, the student belonged to the lowest-achieving group. The number of students who scored below this cut-off score was 114, and the SEN teachers had rated 24 of those students.

Table 7 presents the true positives, the false negatives, the true negatives, and the false positives in the First Steps sample.

Table 7

Identification of Students at Risk for RD Based on the SEN Teachers' Ratings and the Test Scores in the First Steps Sample

	True positives	False negatives	True negatives	False positives
Pre-reading skill (n = 69, number of students rated by SEN teachers)	n (%)	n (%)	n (%)	n (%)
Letter knowledge	26 (38)	0 (0)	10 (14)	33 (48)
Phoneme blending[a]	24 (35)	0 (0)	4 (5)	41 (60)

Note. [a] The phoneme-blending test corresponds to reading/spelling 3–4-letter syllables in the SEN teachers' ratings.

Regarding the First Steps sample, the results first showed that sensitivity of the SEN teacher ratings for letter knowledge was 100% and specificity was 23%, which means that all at-risk students were identified; however, 77% of the students were identified as at-risk even though, according to their test scores, they did not have difficulties with letter knowledge. Further, the sensitivity of teacher ratings for phoneme blending was 100%, whereas specificity was only 9%. Thus, the SEN teachers identified all at-risk students, but they also estimated that 91% of the students who managed quite well in the tests had difficulties with phoneme blending. According to logistic regression analyses and letter knowledge the SEN teachers' ratings were highly associated with the students' categorical letter knowledge test scores ($\chi^2(1) = 10.46$, $p = .001$), and also in phoneme blending to some extent ($\chi^2(1) = 3.5$, $p = .06$).

Discussion

The aim of this study was to get answer to three research questions. First, we wanted to describe the assessment practices the teachers used in identifying difficulties in students' pre-reading skills (letter knowledge and phonological skills) upon entering school in the first grade. The results first showed that the class teachers mostly used one single assessment practice, whereas the SEN teachers often used a combination of several assessment practices. Second, it turned out that the correlations between teacher ratings and test scores were mostly weak or moderate. In addition, we studied the accuracy of the class teachers' and SEN teachers' ability to identify the lowest achievers based on the test scores. To investigate this, we counted the sensitivity and specificity of the ratings. For the JLD sample, there were differences between the accuracy of the class teachers' and the SEN teachers' ratings, and in the First Steps sample, the specificity rate, in particular, was very low.

First, we were interested in finding out the kinds of assessment tools the teachers used to evaluate students' pre-reading skills. We expected (Hypothesis 1) that all teachers would have used versatile assessment practices (Graney, 2008). Instead, most class teachers

relied on qualitative assessment, unlike the SEN teachers. A minority of the SEN teachers reported that they used qualitative assessment solely, and a few combined qualitative assessment with some other means of assessment. It has been shown (Bailey & Drummond, 2006; Martin & Shapiro, 2011) that the qualitative data sometimes used by teachers is not sufficiently accurate or reliable for making decisions on particular skills.

Further, contradictory to what was expected (Hypothesis 2a), the correlations between the teacher ratings and the actual test scores were significant but mostly moderate. The main reason for this finding might be that the teachers had rated the students' skills with 3- and 5-point scales, and the test scores were continuous variables. In previous studies (Feinberg & Shapiro, 2003; Flynn & Rahbar, 1998), the rating scales and instruments have been more consistent with each other. Our study is also in line with Südkamp et al. (2012), who found that achievement tests usually measure very specific areas of academic ability, while teachers' ratings can be much broader evaluations of a skill (e.g., overall ability in reading). Additionally, according to previous research (Flynn & Rahbar, 1998; Martin & Shapiro, 2011; Speece et al., 2011), teacher ratings combined with screening tests has proven to be the most accurate instrument for detecting students who might later confront RD. For example, in Flynn and Rahbar's (1998) study, 88% of at-risk students were discovered by combining both methods.

Finally, partly as we expected (Hypothesis 2b), there were high false-negative rates in both class teachers' and SEN teachers' ratings (JLD sample). Also, the true-positive rate was low in the class teachers' ratings in the JLD sample (Fletcher & Satz, 1984; Flynn & Rahbar, 1998). Contradictory to what was expected (Hypothesis 2b), in the First Steps sample, the true-positive rate was high, but remarkably, the false-positive rate was also extremely high. One explanation might be that teachers are more used to evaluating more comprehensively students' reading and writing skills, instead of specific sub-skills. It could also be difficult for SEN teachers to recognize when the student no longer needs support or

how well the student's skills have developed. Perhaps this finding can be explained by the fact that the SEN teachers in this study only rated those students who had previously received support for their learning, and not necessarily RD (see Soodla & Kikas, 2010).

The current study differs from previous studies in that both class teachers' and SEN teachers' data were available. This enabled, to some extent, drawing comparisons between the two teacher groups. According to this study (JLD sample), the SEN teachers appeared to identify at-risk students a bit more accurately than the class teachers, because their valid positive rate was higher. An explanation for this might be that SEN teachers' have more opportunities to evaluate students and are also in a better position to support individual students than class teachers. Also, SEN teacher education provides SEN teachers with the competency and knowledge to use various assessment tools in their work, compared to class teachers.

A key finding in this study was that there was only a weak link between the teachers' ratings and the test scores. Both underestimations and overestimations of the difficulties were made, especially by the class teachers (JLD sample). Also, in the First Steps sample, the SEN teachers identified significantly more difficulties in pre-reading skills than the students actually had, according to their test scores. The SEN teachers' assessments could have been conducted by the fact that those students had previously received part-time special education for some learning difficulties (Soodla & Kikas, 2010).

Some questions still need to be discussed. First, are the teachers' assessment practices sensitive enough so that most, if not all, of the students in need for support can be detected by using them? In addition, could using several assessment practices improve the accuracy of teacher ratings? According to our findings, at least some SEN teachers have assessed the students for difficulties in pre-reading skills using several assessment practices. Unfortunately in this study we could not show whether use of multiple practices had resulted to more accurate identification of reading difficulties. Anyway, using multiple

assessment practices could enable teachers to provide targeted and individually designed support measures to improve a certain skill when a difficulty is carefully defined. However, this study also shows that, at times, the SEN teachers had evaluated the students' skills using multiple practices, even though the students' test scores were above the cut-off scores. Thus, there is discrepancy between the SEN teachers' perceptions and the actual test scores. Second, this study raises the question of reliability and stability of the test results, as well as the teachers' ratings, especially regarding those students who had been identified as false positives at the beginning of the first grade. One longitudinal study has indicated that late-emerging dyslexia seems rather difficult to predict (Torppa, Eklund, van Bergen, & Lyytinen, 2015). In this study (the First Steps sample), most SEN teachers used tests to assess their students' skills, either alone or with some other assessment practice. That might be the starting point for further and more specific investigation of the difficulty, using additional assessment practices. The need to better understand teachers' impressions stems from research showing that information from formal screening tests and teacher ratings together increase the accuracy of detecting RD in the early elementary grades (Bailey & Drummond, 2006; Flynn & Rahbar, 1998; Martin & Shapiro, 2011).

Finally, does the high rate of false positives lead to the fact that teachers are giving support to students who may be able to learn to read quite well without support, and instead, some of the at-risk students are not getting the support they need? Fletcher and Satz (1984) suggested in their study that students identified as at-risk could be included in classroom-based small-group interventions targeted to the skill deficits identified by the screening battery. Working with these small groups, teachers could reassign children who progress rapidly to other activities, while continuing to intervene with those who struggle with their reading. This kind of flexible teaching and support model is already being used in Finland (see Lerkkanen, 2007; Björn et al., 2016), when the class teacher and the SEN teacher work together in the classroom. According to this study, most class teachers used only

qualitative assessment, which is a parallel finding with previous studies (e.g., Bailey & Drummond, 2006). For this reason, we see that collaboration between class teachers and SEN teachers on assessment issues is desirable, if not necessary.

Teachers have a unique position for the early identification of students' RD, and this requires expertise, as well as the appropriate assessment tools. In order to be able to identify at-risk students and to deliver effective support and interventions in reading, SEN teachers, as well as class teachers, must be able to recognize students' deficits accurately and as early as possible. Early identification, and also intervention in specific deficit areas, can improve students' reading skill levels immediately, as well as prevent later difficulties. The results of this study indicate that teachers need reliable tools, not only to identify difficulties, but also to follow-up on skill development.

Limitations

Before drawing any generalizations from the findings, there are some limitations that should be highlighted. First, in both samples there were missing data, and accordingly, the comparison between class teachers' and SEN teachers' data, for example, was rather complicated. Further, the rather small sample of regular class teachers and SEN teachers did not allow for an analysis of teachers' assessment practices and their relationship to the accuracy of their judgements. In both samples, the teachers were also aware of the fact that there were more students with difficulties among the samples than there would have been if the sample had been based on unselected samples. Thus, it is possible that, in some cases, the teachers assumed the student had difficulties in reading, based on their prior knowledge about the student's low achievement. Furthermore, the variables in the tests and the teachers' ratings (i.e., what teachers were asked to assess) were not entirely comparable to each other.

Conclusions

The results of this study add to our understanding of class teachers' and special education teachers' essential role, and also their ability to evaluate students' pre-reading skills at the beginning of the first grade. The present study revealed that SEN teachers were able to quite accurately identify students at-risk for RD, however they seemed to face challenges in monitoring the progress in their students' literacy skills. Apart from identifying the need for support at early stage of learning to read it is also as important to evaluate students' development of literacy skills using dynamic assessment practices. That could help the SEN teachers to decide when some student no longer is in need for support, and they could have more resources in supporting the at-risk students.Our findings suggest that more attention should be paid to teacher training, as well as developing reliable assessment tools for teachers. Especially, every teacher's expertise in various assessment practices for the early identification of students at risk for RD should be ensured. Further, the current findings emphasize the need for developing high-quality tools that would also enable a systematic and reliable follow-up of a student's skills.

References

Aro, M. (2006). Learning to read: The effect of orthography. In R. M. Joshi & P. G. Aaron (Eds.), *Handbook of orthography and literacy* (pp. 531–550). New Jersey: Lawrence Erlbaum Associates, Inc.

Aro, M., & Björn, P. M. (2016). Preservice and inservice teachers' knowledge of language constructs in Finland. Annals of Dyslexia, 66, 111–126. DOI 10.1007/s11881-015-0118-7

Bailey, A. L., Cano, L., Fischer, D., Freeman, S., Jacobs, J., Heritage, M., et al. (2001). The LDC manual: A guide to using the Literacy Development Checklist (Rev. ed.). Los Angeles: University of California Regents.

Bailey, A. L., & Drummond, K. V. (2006). Who is at risk and why? Teachers' reasons for concern and their understanding and assessment of early literacy. *Educational Assessment,* 11, 149–178. doi:10.1207/s15326977ea1103&4_2

Bates, C., & Nettlebeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology, 21*(2), 177–187. doi:10.1080/01443410020043878

Begeny, J. C., Krouse, H. E., Brown, K. G., & Mann, C. M. (2011). Teacher judgments of students' reading abilities across a continuum of rating methods and achievement measures. *School Psychology Review, 40*(1), 23–38.

Beswick, J. F., Willms, J. D., & Sloat, E. A. (2005). A comparative study of teacher ratings of emergent literacy skills and student performance on a standardized measure. *Education, 126*(1), 116–138.

Björn, P., Aro, M., Koponen, T. K., Fuchs, L. S., & Fuchs, D. H. (2016). The many faces of special education within RTI frameworks in the United States and Finland. *Learning Disability Quarterly*, *39*(1), 58–66. doi:10.1177/0731948715594787

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015) Early identification of reading disabilities within an RTI framework. *Journal of Learning Disabilities, 48*(3) 281–297

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., . . . Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*(2), 327–340.

Davis, G. N., Lindo, E. J., & Compton, D. L. (2007). Children at risk for reading failure. *Teaching Exceptional Children, 39*(5), 32–37.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219–232.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, 37, 184–192. doi:10.1177/00224669030370030801

Eklund, K., Torppa, M., Aro, M., Leppänen, P. H. T., & Lyytinen, H. (2015). Literacy skill development of children with familial risk for dyslexia through grades 2, 3, and 8. *Journal of Educational Psychology, 107*(1), 126–140.

Feinberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly, 18,* 52–65. doi:10.1521/scpq.18.1.52.20876

Fletcher, J., & Satz, P. (1984). Test-based versus teacher-based predictions of academic achievement. A three-year longitudinal follow-up. *Journal of Pediatric Psychology, 9*(2), 193–201. doi:10.1093/jpepsy/9.2.103

Flynn, J. M., & Rahbar, M. H. (1998). Improving teacher prediction of children at risk for reading failure. *Psychology in the Schools, 35*(2), 163–172.

Graney, S. B. (2008). General education teacher judgments of their low-performing students' short-term reading progress. *Psychology in the Schools, 45*(6), 537–549.

Hintikka, S., Landerl, K., Aro, M., & Lyytinen, H. (2008). Training reading fluency: is it important to practice reading aloud and is generalization possible? *Annals of Dyslexia, 58,* 59–79. doi:10.1007/s11881-008-0012-7

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59,* 297–313. doi:10.2307/1170184

Holopainen, L., Ahonen, T., & Lyytinen, H. (2001). Predicting reading delay in reading achievement in a highly transparent language. *Journal of Learning Disabilities, (34*)5, 401–414.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*(4), 582–600.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice, 24*(4), 174–185.

Kikas, E., Silinskas, G., & Soodla, P. (2015). The effects of children's reading skills and interest on teacher perceptions of children's skills and individualized support. *International Journal of Behavioral Development, 39*(5), 402–412.

Kiuru, N., Lerkkanen, M.-K., Niemi, P., Poskiparta, E., Ahonen, T., Poikkeus, A.-M. & Nurmi, J.-E. (2013). The role of reading disability risk and environmental protective factors in students' reading fluency in grade 4. *Reading Research Quarterly, 48*(4), 349–368.

Leinonen, S., Müller, K., Leppänen, P. H. T., Aro, M., Ahonen, T., & Lyytinen, H. (2001). Heterogeneity in adult dyslexic readers: Relating processing skills to the speed and accuracy of oral text reading. *Reading and Writing: An Interdisciplinary Journal, 14*, 265–296. doi:10.1023/A:1011117620895

Lerkkanen, M.-K. (2007). The beginning phases of reading literacy instruction in Finland. In P. Linnakylä & I. Arffman (Eds.) *Finnish reading literacy. When quality and equity meet* (155–174). Jyväskylä: University of Jyväskylä, Institute for Educational Research.

Lerkkanen, M.-K., Ahonen, T., & Poikkeus, A.-M. (2011). The development of reading skills and motivation and identification of risk at school entry. In M. Veisson, E. Hujala, P. K. Smith, M. Waniganayake, & E. Kikas (Eds.) *Global perspectives in early childhood education: Diversity, challenges and possibilities* (pp. 237–238). Frankfurt am Main, Germany: Peter Lang.

Lerkkanen, M. -K., Niemi, P., Poikkeus, A. -M., Poskiparta, M., Siekkinen, M., & Nurmi, J. -E. (2006). *The first steps study [Alkuportaat], ongoing.* University of Jyväskylä, Finland.

Lerkkanen, M.-K., Poikkeus, A.-M., & Ketonen, R. (2006). *ARMI. Luku- ja kirjoitustaidon arviointimateriaali 1. luokalle. [ARMI – a tool for assessing reading and writing skills in grade 1].* Helsinki: WSOY.

Lerkkanen, M.-K., Rasku-Puttonen, H., Aunola, K., & Nurmi, J.E. (2004). Reading performance and its developmental trajectories during the first and the second grade. *Learning and Instruction, 14*(2), 111–130. doi:10.1016/j.learninstruc.2004.01.006

Lyytinen, H., Erskine, J., Tolvanen, A., Torppa, M., Poikkeus, A.-M., & Lyytinen, P. (2006). Trajectories of reading development: a follow-up from birth to school age of children with and without risk for dyslexia. *Merrill-Palmer Quarterly, 52*(3), 514–546.

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

Martin, S. D. & Shapiro, E. S. (2011). Examining the accuracy of teachers' judgments of DIBELS performance. *Psychology in the Schools, 48*(4), 343–356. doi: 10.1002/pits.20558

Mesmer, E. M., & Mesmer, H. A. E. (2008). Response to intervention (RTI): What teachers of reading need to know. *The Reading Teacher, 62*(4), 280–290. doi:10.1598/RT.62.4.1

Parrila, R., Aunola, K., Kirby, J. R., Leskinen, E., & Nurmi, J. E. (2005). Development of individual differences in reading: Results from longitudinal studies in English and Finnish. *Journal of Educational Psychology, 97* (3), 299–319. doi:10.1037/0022-0663.973.299

Poskiparta, E., Niemi, P., & Lepola, J. (1994). *Diagnostiset testit 1. Lukeminen ja kirjoittaminen.* Turku: Turun yliopisto, Oppimistutkimuksen keskus. [Diagnostic Tests 1, Reading and Writing].

Puolakanaho, A., Ahonen, T., Aro, M., Eklund, K., Leppänen, P. H. T., Poikkeus, A.-M., . . . & Lyytinen, H. (2007). Very early phonological and language skills: estimating individual risk of reading disability. *Journal of Child Psychology and Psychiatry 48*(9), 923–931. doi:10.1111/j.1469-7610.2007.01763.x

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*(2), 143–174.

Snowling, M. J., Duff, F., Petrou, A., & Schiffeldrin, J. (2011). Identification of children at risk of dyslexia: the validity of teacher judgments using 'Phonic Phases'. *Journal of Research in Reading, 34*(2), 157–170. doi: 10.1111/j.1467-9817.2011.01492.x

Soodla, P., & Kikas, E. (2010). Teachers' judgment of students' reading difficulties and factors related to its accuracy. In A. Toomela (Ed.) *Systemic Person-Oriented Study of Child Development in Early Primary School* (pp. 73–94). Pieterlen, Switzerland: Peter Lang.

Soodla, P., Lerkkanen, M.-K., Niemi, P., Kikas, E., Silinskas, G., & Nurmi, J.-E. (2015). Does early reading instruction promote the rate of acquisition? A comparison of two transparent orthographies. *Learning and Instruction, 38,* 14–23. doi:10.1016/j.learninstruc.2015.02.002

Speece, D. L., Schatschneider, C., Silverman, R., Pericola, Case, L. P., Cooper, D. H., & Jacobs, D. M. (2011) Identification of Reading Problems in First Grade Within a Response-to-Intervention Framework. *The Elementary School Journal, 111*(4), 585–607.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*(8), 795–819. doi:10.1002/pits.20113

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. doi:10.1037/a0027627

Takala, M. & Ahl, A. (2014). Special education in Swedish and Finnish schools: Seeing the forest or the trees? *British Journal of Special Education 41*(1), 59–81. doi:10.1111/1467-8578.12049

Torppa, M., Poikkeus, A.-M., Laakso, M.-L., Eklund, K., & Lyytinen, H. (2006). Predicting delayed letter name knowledge and its relation to grade 1 reading achievement in children with and without familial risk for dyslexia. *Developmental Psychology, 42*(6), 1128–1142.

Torppa, M., Eklund, K., van Bergen, E., & Lyytinen, H. (2015). Late-emerging and resolving dyslexia: A follow-up study from age 3 to 14. *Journal of Abnormal Child Psychology, 43*(7), 1389–1401. doi:10.1007/s10802-015-0003-1

Zumeta, R. O., Compton, D. L., & Fuchs, L. S. (2012). Using word identification fluency to monitor first-grade reading development. *Exceptional Children, 78*(2), 210–220.