INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI®

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

ACCURACY OF TEACHER ASSESSMENTS OF ESL CHILDREN AT-RISK FOR READING DISABILITY

by

Marjolaine Michèle Limbos

A thesis submitted in conformity with the requirements
for the degree of the Master of Arts
Department of Human Development and Applied Psychology
Ontario Institute for Studies in Education of the
University of Toronto

Accuracy of Teacher Assessments of ESL Children
At-Risk for Reading Disability

Marjolaine Michèle Limbos
Master of Arts, 1998
Human Development and Applied Psychology,
Ontario Institute for Studies in Education
of the University of Toronto

Abstract

This study examined the accuracy of teacher assessments in screening for reading disabilities among ESL and first language (L1) students. Three hundred and seventy-five children (262 ESL, 113 L1) were administered academic and oral language tests in grades 1 and 2. Concurrently, 48 teachers nominated children at risk for reading failure and completed rating scales assessing academic and oral language skills. Scholastic records were reviewed for teacher notation of concern or referral. The criterion measure for at-risk status was a standardized reading score based on 3 measures: (1) phonological awareness, (2) rate of access of phonological information and (3) word recognition. Results indicated that teacher rating scales and nominations were accurate in screening but the accuracy was considerably lower for ESL compared to L1 students. Relative to other forms of screening, teacher expressed concern in student files had a lower sensitivity. Finally, oral language proficiency was found to contribute to misclassifications.

Acknowledgments

I am indebted to Dr. Esther Geva for giving me the opportunity to study this field and for imparting her tremendous knowledge and experience. It has been a privilege to witness her broad-based knowledge of educational and practice issues as they apply to ESL children, as well as to learn from her presentation and manuscript preparation skills. I am also thankful for her support in this undertaking that not only involved the thesis but also the preparation of a paper for a conference presentation. Dr. Geva always went beyond what would be expected of a supervisor. I appreciate the thorough review and comments from Dr. Dale Willows, the second-reader of my thesis. I would also like to thank the entire ESL team at OISE, and in particular Zohreh Yaghoubzadeh, Alison Mack, Fataneh Farnia, Robin Sidhu and Barbara Schuster, for the many thought-provoking discussions in the ESL meetings and for their commitment to the collection and organization of the data. In addition, I would like to thank the many school teachers who took time out of their busy schedules to meet with me in the completion of the interviews. I would also like to greatly acknowledge my dear husband, Dr. David Joyce, for his unrelenting support in the completion of this work and the pursuit of my studies. His breadth of knowledge continues to amaze me, and his ability to communicate statistical and epidemiological concepts was of tremendous assistance to me. Finally, I would like to thank my father for his support and for teaching me the value of a good work ethic, and my mother for her interest in my work and insightful comments on the study.

Table of Contents

List of Tables

Accuracy of Teacher Assessments of ESL Children At-Risk for Reading Disability

Recently, major interest has developed regarding the early identification of learning disabilities among second language learners. Historically, minority children in certain immigrant groups were discriminated against and regarded as educationally inferior, and were much more likely to be identified as learning disabled and placed in special education classes (Cummins, 1984). Special education placement has been criticized as reflecting socioeconomic, linguistic and cultural factors rather than psychoeducational factors. Further, there is a growing awareness that educational difficulties may reflect linguistic and acculturation processes. In response to these criticisms, professional and school personnel have been delaying diagnosing English as a Second Language (ESL) children as learning disabled for up to 4-5 years in order to allow time for proficiency in the language to develop. It has been argued that it is only then that such difficulties can reliably be attributed to a learning difficulty. Although this growing awareness indicates an increased sensitivity by professionals to the language needs of ESL children, it has also led to a tendency to ignore the possibility that ESL learners are having difficulties not only because of insufficient ESL oral proficiency but because they are having word decoding or language processing problems typical of a learning disability. Delaying assessment for these children takes away any preventive opportunities and possibilities of instituting remediation strategies. Consequently, professionals are reconsidering the early identification of reading disabilities amongst second language students.

The process of early identification, however, is highly complex, with major concerns related to measurement, prediction, and the benefits of early remediation. For instance, knowledge is only beginning to emerge about the reading development of second language

learners. Linguistic, cognitive and language factors involved in reading are being examined to determine if they are specific to L1 students, or if they may also be applicable to ESL students (e.g., Geva, 1998; Schiff-Myers, 1992; Westernoff, 1995).

Insight from the first language literature indicates that there is a causal relationship between phonological awareness and reading, particularly from kindergarten through the second grade (Elbro, Borstrom & Paterson; Stanovich, 1988; Wagner & Torgeson, 1987). Phonological processing refers to an individual's mental operations that make use of the phonological or sound structure of oral language when he or she is learning how to decode written language. The kinds of phonological processing skills and knowledge that have been most frequently studied include phonological awareness, phonological memory, and rate of access for phonological information. Phonological awareness and rapid serial naming have consistently emerged as two of the most powerful predictors of subsequent reading achievement during the first grade (Bowers, 1995; Felton & Wood, 1989; Meyer, Wood, Hart & Felton, 1998; Torgeson, Wagner, & Rashotte, 1994; Wolf, 1997).

With the establishment of a robust relationship between phonological skills and reading in first language learners (L1), researchers are beginning to examine whether similar predictors apply to second language learners. Recent studies that have compared L1 and ESL basic literacy skills among bilingual or second language learners suggest that there are certain universal cognitive and linguistic factors such as phonological processing, working memory, orthographic knowledge and speed of lexical access which appear to play a similar role in L1 and ESL reading skills acquisition (Bruck & Genesee, 1995; Chitiri, Sun, Willows & Taylor, 1992; Durgunoğlu, Nagy & Hancin-Bhatt, 1993; Geva & Clifton, 1993; Geva, Wade-Wooley & Shany, 1993; Geva & Siegel, in press; Gholamain & Geva, in press).

Given the strong data regarding the role of phonological awareness in predicting subsequent reading abilities found in the first language literature, and emerging evidence demonstrating the generalizability of these findings to the ESL population, school personnel and researchers alike are considering the feasibility of applying cognitive factors to the assessment and diagnosis of reading disability. For some time, researchers have been questioning the traditional methods of classification based on the discrepancy definition of learning disabilities. Although many school-identified learning disabled students do meet commonly applied criteria (e.g., a 15-point difference between ability and achievement) some do not (Algozzine & Ysseldyke, 1986). Furthermore, many low achieving students, never classified as learning disabled also meet the same criteria (Epps, Ysseldyke & Algozzine, 1983), and many normal students are classified using these criteria (Ysseldyke, Algozzine, & Epps, 1983). In criticism of these methods, Stanovich (1986, 1988, 1992) has argued that the ability to decode words in the early stages of reading is dependent solely on phonological awareness and that it is independent of intelligence or IQ scores. It is only when comprehension is considered that intelligence becomes a significant determinant of reading achievement. Stanovich's arguments are supported empirically by the finding that phonemic awareness is a stronger predictor of reading achievement than global measures such as intelligence or traditional measures of reading readiness (Juel, Griffith, & Gough, 1986; Siegel, 1988, 1993; Stanovich, Cunningham, & Cramer, 1984; Tunmer & Nesdale, 1985; Vellutino & Scanlon, 1987). Based on the findings of these previous studies demonstrating the strong predictive ability of phonological awareness, the current research considers phonological skills to be the best early indicator of reading difficulty.

In practice, not all children are formally assessed for adequate development of phonological awareness because of financial and time constraints. Instead, it has been generally assumed that the best complementary approach for early identification of children with learning disabilities/difficulties is through teacher assessments, as obtained through verbal nominations or ratings scales (Lindsay & Wedell, 1982). Teachers' input may be invaluable due to their ample experience, extended direct contact with each child, and cost-effectiveness as compared to formalized assessment or traditional educational screening tests. Authors have repeatedly stated that teachers are the best predictors of a child's subsequent success or failure in school. However, despite this general assumption, there is a paucity of literature examining teacher screening of reading disabilities. There is only one well-designed study of teacher screening for reading disability that was conducted by Salvesen and Undheim (1994). These authors found that the degree of correspondence between teacher assessments of low achievement and corresponding objective testing results was relatively good (Salvesen & Undheim, 1994). However, despite the excellent study design, the authors did not report standard measures of accuracy of screening instruments (e.g. sensitivity, specificity, and so on) that are considered essential in the decision regarding the utility of a screening test (Altman & Bland, 1994).

The usual method for evaluating the utility of educational screening methods has been through correlation coefficients. More recently, epidemiologists and researchers investigating early detection and intervention have emphasized the utility of other indices of accuracy such as the sensitivity, specificity, likelihood ratio, and the kappa statistic (Mantzicopoulos & Morrison, 1994; Meisels, 1988; Jaeschke, Guyatt, & Sackett, 1994; Sackett, Haynes, & Tugwell, 1985). These indices offer an enhanced interpretation of the

results of a screening test, extending beyond the relatively non-specific information provided by correlations. For instance, a highly significant relationship may be revealed through simple correlations, but this finding tells the researcher little about the ability of the test to correctly classify individuals as at-risk or not at-risk. Conversely, the sensitivity and specificity of a test have important clinical value. Screening tests with a high sensitivity give valuable information on the importance of a negative screening result; if a student is said not to be at-risk, there is a high likelihood that he/she is truly not at-risk. However, knowing that a test is highly sensitive gives little information on the utility of a positive screening result, in that many students could still be falsely positive. That is, they would be identified as at-risk when, in fact, they are not.

To better define the usefulness of a positive result of screening, one must also examine the specificity or likelihood ratio. A test with a high specificity when positive makes the probability of the child being truly at-risk very high. Likewise, a test with a high likelihood ratio indicates that the test is very good at increasing the certainty about a positive identification of at-risk. Using all of these test characteristics concurrently allows a clear definition of the merits and weaknesses of the screening test and avoids some of the pitfalls of using tests without adequate knowledge of their accuracy. For example, Meisels (1988) indicates that parents of children falsely identified as having a school difficulty (as might occur in screening tests with high sensitivity) may experience considerable stress and anxiety. Mantzicopoulos and Morrison (1994) warn, however, that a high specificity of a screening measure is also of concern, as a classification based only on high specificity may potentially deprive children of the benefits of early identification and intervention. Such

children may experience repeated failures and frustrations with academic tasks before they are actually identified and before they receive appropriate intervention.

Not only have the accuracy of teacher screening instruments not been investigated comprehensively, but the accuracy of such measures has not been examined in more specific subgroups. Specifically, the accuracy of teacher ratings and nomination in the determination of achievement levels and learning disabilities has not been examined for ESL pupils. This subgroup presents several unique challenges to the assessment process, as the correct identification of learning difficulties in these students may be influenced by the complex interaction of second language proficiency, different cultural norms and differences in rates of development. For example, for some time researchers have been hypothesizing that ESL children may be misclassified due to their oral language skills. Because of their rudimentary oral second language skills, ESL students' reading skills may be mistakenly assessed to be poor. Indeed, in schools ESL students are frequently placed in classes where they are taught vocabulary and oral language skills, with the assumption that this will result in improvements in other academic subjects. It is presumed that it is only when students have mastered the basic oral communication skills that they are prepared to use their explicit knowledge about language structure, such as its sounds, to "crack" the orthographic or alphabetic letter code. However, this developmental relationship between oral language and reading is not yet clear and is only beginning to be researched (Geva, 1998; Petrulius-Wright, 1998).

Preliminary research indicates that the development of oral language skills does not ensure mastery of learning to read. The language skills required for adequate oral communication may not be the same as those mobilized for reading (Snyder & Downey, 1991). For example, oral language skills such as narrative and communicative adequacy

have not been found to have a significant correlation with pre-reading variables such as phonemic awareness, print production and decoding (Dickinson & Snow, 1987). Other research indicates that such oral language skills as vocabulary and grammatical knowledge were either marginally related or not related to word-identification performance (Durgunoğlu, Nagy, & Hancin-Bhatt, 1993; Geva & Siegel, in press; Gholamain & Geva, in press). Listening comprehension has been found to be more predictive of reading comprehension than of basic reading skills in ESL children (Petrulius-Wright, 1998). Decontextualized oral language skills including syntax may have a greater association with prereading variables, although this continues to be questioned (Gottardo, Stanovich & Siegel, 1996). While syntactic awareness may contribute to decoding, it appears that it is secondary to phonological awareness (Gottardo et al., 1996). The difficulties in assessing oral language and understanding the association between oral language and reading in ESL students may account for these children being either overlooked or otherwise misjudged to be at-risk for reading failure on the basis of their oral language skills.

The present study was motivated by the unique challenge of the assessment of ESL children for reading difficulties. The study objectives were to: (1) examine the degree of correspondence between teacher rating scales and objective testing results in identifying reading disabilities; (2) determine the difference in the degree of accuracy between the teacher assessment of reading achievement among ESL and L1 students; (3) determine the degree of agreement between teacher rating scales, nomination of at-risk children and formal academic records and (4) explore the influence of oral language proficiency factors on the misclassification of students.

Methods

Participants

Students The battery of cognitive, linguistic and reading tasks was administered to 382 first-grade children. Teacher assessments were obtained for these children as well. Seven students were excluded from the analysis because of incomplete data or unclear language designation, leaving a total of 375 students. All children were assessed at the beginning of grade 1 (Time 1). The same battery was administered a year later, in the fall of grade 2 (Time 2), to 144 children comprising Cohort 1. The other 192 students who were assessed only at Time 1 comprised Cohort 2. So far the data for Cohort 2 at Time 2 is not available. Thirty-nine of the initial 183 in Cohort 1 did not complete testing, equivalent to an attrition rate of 21.3%.

The demographic characteristics of all the participants are summarized in Table 1. In both cohorts, children were 6-7 years old (M=79.33, SD= 6.38 months) attending grade 1 in 10 schools in three different areas of a large metropolitan multi-ethnic city. There were 262 ESL (69.86%) and 113 (30.13%) non-ESL students. An ESL student was defined as a student whose first spoken language was not English. It should be pointed out that many of the children were born in Canada but did not speak English until they began to attend school. Eleven students were initially eliminated from the sample because their first language was an English dialect. The most common first language of the ESL students was Punjabi or a dialect of Punjabi (68.32%). This was followed by Cantonese (17.56%), Portuguese (8.01%) and other (6.11%). Although the first language of L1 students was English, several of these students had an ethnic origin other than Canadian. In the L1 sample, the ethnic origins included South-East Asian (17.70%), Portuguese (8.00%), Italian (4.42%), Carribean

(16.81%), Philipino (1.77%) and other (4.42%). Of the entire sample, 278 (72.77%) were born in Canada, 21 (5.50%) in India, 16 (4.19%) in Sri Lanka, 2 (0.52%) in Pakistan and 65 (17.02%) in other countries.

Insert Table 1 about here

Teachers Forty-eight teachers completed the interviews and rating scales and 4 teachers refused to participate. Total years of teaching ranged from 1 to 35 (M=15.83, SD=9.83) and years of experience in teaching grade 1 ranged from 1 to 25 (M=6.95, SD=5.82). Years of teaching ESL children ranged from 1 to 27 years (M=8.97, SD=6.67). Six teachers (12.50%) had a Master's degree in addition to their teacher education, 15 (31.25%) had special education certification while 14 (29.17%) had ESL teacher certification. Of 31 teachers of children in Cohort 2 who filled out a questionnaire examining their reading philosophy, thirteen (41.89%) rated themselves as using a "whole word" as opposed to a "phonics" method in reading instruction, and 18 (58.10%) described themselves as using both a phonics and whole word method to reading instruction.

Instruments and Procedure

The present study forms part of a larger research project investigating oral language and literacy skills development among young ESL learners. The project involves a longitudinal design in which children are tested in the fall and spring of grades 1 to 3 on a variety of oral and written language indices. As part of a longitudinal study of ESL children, a large battery of oral language and reading tasks was administered. The battery of tests was administered individually during two, half-hour sessions. There were two cohorts of grade 1 children

tested one year apart (Cohort 1 and Cohort 2). Although the entire sample (N= 375) was tested at Time 1, data on Time 2 (i.e., fall of grade 2) is available so far only for Cohort 1 (N=144). Testing at Time 1 for both cohorts occurred concurrently with teacher interviews and rating scales. Of relevance to the current investigation are the following measures of oral proficiency, phonological awareness, word recognition, and spelling:

A. Oral proficiency.

In order to assess oral proficiency skills four measures were administered.

(1) Syntactic Awareness. To assess participants' grammatical and oral language skills, an adaptation of Johnson and Newport's (1991) test was administered. In this test, children listen to tape recorded sentences and have to repeat each sentence as it is heard. Various grammatical structures are manipulated (e.g., number, tense, relative pronouns, auxiliary). Phonological and morphophonological errors are not counted in the total scores.

(2) Expressive Vocabulary. The Expressive One-Word Picture Vocabulary Test-Revised (EOWPT; Gardner, 1990) was used. In this test, children are asked to name pictures they are shown and a total score is given based on the number correctly identified.

(3) Receptive Vocabulary. The Peabody Picture Vocabulary Test-Revised (PPVT-R; Dunn & Dunn, 1981) was used as a measure of receptive vocabulary and verbal comprehension. Participants are asked to point to the picture corresponding to the vocabulary word given.

B. Phonological Awareness

(1) Pseudoword decoding. The Word Attack subtest of the Woodcock Reading Mastery Test-Revised (WRMT-R; Woodcock, 1987) was used to assess children's ability to employ phonological skills to decode pseudowords. This test consists of 50 pseudowords that

comply with English phonology. Children read these pseudowords one at a time, and testing is discontinued when the child makes 5 consecutive errors. Pseudoword decoding is taken to be a reliable measure of phonological skills.

(2) Rate of access for phonological information. Rate of retrieval of phonological codes from long-term memory was measured with the Rapid Automatized Naming task (RAN; Denckla & Rudel, 1976). In this continuous naming task, children are asked to name as fast as they can a series of 5 letters that they had first been successful in naming during a trial. There are ten rows each containing 10 letters, making a total of 100 items. The time (in seconds) to complete the naming task was recorded.

C. Word recognition. To assess children's ability to read words in English, the word recognition subtest of the Wide Range Achievement Test-3 (WRAT3; Wilkinson, 1993) was used. This test consists of 42 unrelated words. The words range from being highly frequent in the child's environment (e.g., cat) to being relatively uncommon (e.g., egregious). When the child makes 10 consecutive errors testing is discontinued.

D. Spelling. A variation of the Developmental Spelling Test (Ferroli & Shanahan, 1987) was used to assess children's development and knowledge of word elements in English, as revealed in their spelling and error patterns. This list consists of sixteen simple and highly frequent words which have been included on the basis of orthographic representations of specific morphological and phonological features (e.g., long vowels, morphophonological endings). The experimenter pronounced each word, gave a short phrase to specify the meaning of the word, and then repeated the word again before asking the child to write it down (e.g., cats).

Definitions of At-Risk

Teacher Identification of At-Risk Status. Three sources of information were utilized to yield three types of teacher at-risk classifications. First, teacher interviews and Individual Scholastic Records (ISR) provided information on concern expressed by the teacher. Any time teachers indicated in the interview that a child had been referred to an in-school review committee for academic or other concerns or when a referral for assessment, below average academic performance or concern was written in the ISR, this information was noted (RECORDS). Second, during the winter and spring of the first year of primary school, teachers assessed pupils on various academic domains. A performance scale from 1 to 7 (1 = very poor; 4 = medium; 7 = very high) for each domain was utilized. Teachers were asked to rate each student according to his/her expectations at this level and/or in comparison with all other children in the classroom in each of the domains of spelling, reading, arithmetic, oral expression, vocabulary, writing, reading comprehension, oral/listening comprehension, and grammatical sentence structure. For the current study the outcome of interest was reading performance. Thus, any reading score less than or equal to 2 was classified as at-risk on the Teacher Rating Scale - Reading (TRATING), while scores above 2 were classified as not at-risk. Thirdly, during semi-structured interviews, teachers were asked to nominate children they felt were at-risk for a significant reading difficulty or disability. Any pupil the teacher "nominated" as being at-risk for the development of a learning difficulty or disability in the interview was classified as Teacher Nomination (TNOM) at-risk. Only children nominated by the teachers as being at-risk for developing a reading difficulty were designated as at-risk; most children were described by teachers as having other difficulties in addition to reading (e.g., oral language, mathematics). It should

be noted that 2 children who were nominated as at-risk, were eliminated from the sample because they were described by the teachers as having social, emotional or behavioral difficulties and not reading difficulties per se.

Objective Determination of At-Risk Status

In order to identify at-risk status objectively, a combined standardized reading score (CSRS) based on the WRAT3, Word Attack and RAN raw scores for the entire sample was generated. Because reading norms do not exist for second language speaking children low readers were identified by a CSRS below -1.5 SD. These standardized scores were calculated at Time 1 (CSRS-1) and then again at Time 2 based on reading scores from Cohorts 1 and 2 (CSRS-2). The CSRS at-risk designation was considered the objective "gold standard" assessment of at-risk for the purpose of this study.

Data Analysis

Intercorrelations were calculated amongst the various academic areas (i.e., oral expression and reading) that teachers rated. Teacher ratings were also correlated with the test results at Time 1 for the entire group and at Time 2 for Cohort 1 only. Pearson product-moment correlations were performed on the entire group, as well as the ESL and L1 groups separately.

The three forms of teacher screening assessments (TRATING, TNOM, RECORDS) were examined to determine their respective accuracy for screening of reading disability. Two by two tables were constructed sequentially using teacher screening of at-risk or not at-risk vertically and objective (or gold standard) assessment (i.e., CSRS-1, CSRS-2) of at-risk or not at-risk horizontally, as show below.

Objective Determination (CSRS)

		At-Risk	Not At-Risk
Teacher Screening (TRATING, TNOM or RECORDS)	**At-Risk**	True positive	False positive
	Not At-Risk	False negative	True negative

Using this table, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated. These screening characteristics were calculated at both Time 1 and Time 2, for the total group, ESL and L1 groups separately and for each form of teacher assessment method (i.e., TNOM, TRATING, and RECORDS). *Sensitivity* is the proportion of true positives that are correctly classified by the objective test (i.e. the ability of the test to identify those who are at-risk) and is calculated as the number of true positives divided by the sum of true positives and false negatives. *Specificity* is the proportion of true negatives that are correctly identified by the test (i.e. the ability of a test to identify correctly those who are not at-risk) and is calculated as the number of true negatives divided by the sum of true negatives and the false positive. *Likelihood ratios* (LR) were calculated as the sensitivity divided by (1 – specificity).

Accuracy results are presented as percentage with 95% confidence intervals (95% CI) in brackets. The desired levels of specificity and sensitivity of a screening test vary depending on the goals of the test. As a guide, the American Psychological Association (1985) has suggested that 80% sensitivity and 90% specificity are preferable levels for psychological tests. Further, as a guideline for interpretation, a LR of 1-2 is considered to alter pre-test probability minimally. A LR of 2-5 results in small changes in probability; a LR of 5-10 results in moderate shifts in probability; and a LR of greater than 10 causes large

and often conclusive changes in probability (Jaeschke et al., 1994). The χ^2 and Fisher exact test were used for comparison of proportions (false positive and false negative rates as well as sensitivity and specificity) between L1 and ESL students for the various screening methods (TNOM, TRATING, RECORDS), as well as between the various screening methods for a given language group.

Comparisons of the efficiency of the screening tests used was also determined using the kappa index, which is the ratio of observed accuracy beyond chance to the maximum achievable accuracy beyond chance (Sackett et al., 1985). The kappa index has a maximum value of +1. Values close to 1 reflect almost perfect agreement between tests, values between .2 and .8 indicate a fairly good agreement, values near zero reflect agreement only by chance, and values below zero show disagreement. The kappa index for agreement between the teacher nominations of at-risk for a learning difficulty/disability and low reading achievement, and between teacher assessment of reading (cutoff point at score-value ≤ 3) and low reading achievement were calculated. The kappa index was calculated separately for ESL and L1 pupils at Time 1 and Time 2.

Data analyses were also conducted to examine hypotheses regarding factors that could account for differing accuracy rates. Mean test scores are shown as mean with standard deviations in brackets. One-way analysis of variance (ANOVAs) were calculated using oral proficiency measures (i.e., EOWPT, PPVT-R, sentence repetition) as the dependent measures and the presence of false positives versus the true negatives and subsequently the false negatives as compared to the true positives as the independent variables to examine the impact of oral speaking skills on the false positive rate. These

analyses were conducted separately for the L1 and ESL groups. Results were considered statistically significant at $p < 0.05$.

Results

Correlations Among Teacher Rating Scales

Teacher assessments given on 7-point Likert-type rating scales showed normal distributions, with median scores of 4. The means and standard deviations of the teacher ratings for the L1 and ESL group are provided in Table 2. There were significant differences between the L1 and ESL groups in terms of teacher rating scales of oral comprehension [$F(1, 313)=22.28$, $p<.001$], oral expression [$F(1,313)=52.00$, $p<.001$], reading comprehension [$F(1, 306)=9.75$, $p<.01$], vocabulary [$F(1, 313)=51.23$, $p<.001$] and grammatical sentence structure ratings [$F(1, 312)=31.41$, $p<.001$]. Even though the means on teacher ratings of reading, spelling, writing, grammar, arithmetic were somewhat lower in the ESL than in the L1 group, there were no significant differences between the two groups in these domains.

Insert Table 2 about here

Teacher assessments of academic performance were moderately to highly correlated with each other (range of $r = .62$ to .91). The intercorrelations amongst the various teacher ratings were high to very high for the entire group, as well as for the L1 and ESL groups. For both language groups, the correlations between the oral expression ratings and academic subjects (i.e., math, $r = .55$; reading, $r = .62$; and spelling, $r = .60$) were lower than those with other oral measures (i.e., oral comprehension, $r = .77$; vocabulary, $r = .91$), but continued to be in the high range. Further, oral proficiency ratings (i.e., oral expression and comprehension)

correlated highly with the child's overall academic rating for the L1 (r =.66 and .70, respectively) and the ESL groups (r =.73 and .84, respectively).

Correlations Among Objective Measures

Examination of the correlations among the standardized academic and oral proficiency scores provide a different profile from correlations between similar variables on the teachers' ratings. Correlations among the receptive and expressive oral language measure (i.e., EOWPT, PPVT-R, and sentence repetition) were high and significant (range = .69 - .76) and those between the reading measures, Word Attack and the WRAT3, were also high and significant (r =.77). However, contrary to the pattern of correlations among the teacher ratings, the correlations between the oral comprehension and expression measures and the reading scores were low to moderate (range = .16 - .34). Thus, the correlations between the objective measures of oral and reading skills were lower than teacher ratings of these domains.

Correlations Between Teacher Assessments and Objective Measures

The means and standard deviations of the scores on the oral proficiency and academic measures were calculated separately for the L1 and ESL group (see Table 3). There were significant differences between the ESL and L1 groups on the Time 1 oral proficiency measures [EOWPT, $F(1, 340) = 103.10$, $p<.001$; PPVT-R, $F(1, 343) = 100.77$, $p<.001$; sentence repetition, $F(1, 339) = 88.01$, $p<.001$], with the L1 group having higher scores in each of these areas. There were no significant differences between the two groups on the three reading measures (i.e., Word Attack, WRAT3, RAN).

Insert Table 3 about here

For the entire group there were moderate to high correlations (range of r =.44 to .71) between teacher ratings and objective test results at Time 1. Although there were no significant differences between the correlations of the L1 and ESL groups at Time 1, the correlations between teachers' assessments of oral language and objective test results tended to be higher for the ESL group. For example, the relationship between teacher assessment of vocabulary and expressive vocabulary (i.e., the EOWPT) was moderate (r =.45) for the L1 group whereas it was high (r =.56) for the ESL group, and the relationship between teachers' assessment of grammar and the sentence repetition task was moderate (r =.32) for the L1 group but it was high for the ESL group (r =.54). In terms of the assessment of reading, all the correlations between teacher reading assessment score and objective test results at Time 1 were high. For example, for the L1 group, the correlations between teacher reading assessment and the RAN, Word Attack and WRAT3 scores were -.66, .61, and .71 and for the ESL group these correlations were -.58, .65, and .71, respectively.

The correlations between teacher ratings and objective test results at Time 2 for Cohort 1 ranged from moderate to high (range of r = .41 to .72), consistent with Time 1 correlations. There were, however, lower correlations between teacher ratings and test results of oral language skills for the L1 group, but these correlations were not significantly smaller than those for the ESL group. Similarly, there were no significant differences between the two groups in terms of the association between teacher ratings and reading test scores.

Accuracy of Teacher Assessments

The CSRS-1 identified a total of 19 students (5.65%) as at-risk for reading disability at Time 1, eleven (57.89%) of whom were ESL. In the accuracy data (see Table 4, 5, 6), however, the number of students identified by the CSRS-1 as at-risk depended upon data being present for both teacher and objective measures. Designation of at-risk by teachers varied with the screening measured used. The number of students designated at-risk by teacher screening at Time 1 was 65 (20.57%), 69 (21.00%), and 24 (7.95%) depending on whether the teacher rating scale, teacher nominations or record data were used, respectively. Again, the numbers of children designated as at-risk in the accuracy data depended on objective data being available for a particular participant.

Time 1

Accuracy data (i.e., sensitivity, specificity) on the TRATING index were calculated from 2 x 2 tables of the number of correctly and incorrectly identified students (see Table 4). At Time 1, the accuracy data of the TRATING for the entire sample, using CSRS-1 as the gold standard, was moderate to high.

Sensitivity and specificity of TRATING scores for the total group were in the low to high 80% range. Subgroup analysis revealed a higher sensitivity for the L1 group than for the ESL group (sensitivity = 100% vs. 70%, respectively), although this difference was not statistically significant. Specificity for both the L1 and ESL groups was comparable to that of the entire sample.

Insert Table 4 about here

As can be seen in Table 5 which provides accuracy data on the TNOM, screening sensitivity fell below 80% for the total group and was somewhat lower than that of the TRATING for the L1 group. These differences were not significant, however. Sensitivity of the TNOM for the ESL group remained low at 70%.

Insert Table 5 about here

Despite the relatively good performance of teacher screening using structured assessment (i.e., TRATING and TNOM), accuracy was much lower when spontaneous expressed concern (i.e., RECORDS) was considered (see Table 6). Using RECORDS, sensitivity for the entire group was rather low at 20%. Analysis of the subgroups showed that most of the sensitivity was in the L1 group (40%), whereas the teacher sensitivity for the ESL group was extremely low at 10%. This number corresponds with a failure to detect 9 of the 10 ESL children designated as at-risk by the gold standard testing. Despite the large differences in sensitivity values in comparison with the other methods of screening, we were unable to detect statistically significant differences between some subgroups because of the small absolute number of children with a designation of at-risk for reading failure. There were, however, statistically significant differences in teacher sensitivity between RECORDS and the TRATING and TNOM for the total group [χ^2 (1, $\underline{N}$=33) = 13.24, $\underline{p}$<.001 and χ^2 (1, $\underline{N}$ = 33) = 10.94, $\underline{p}$ < 0.001, respectively]. Subgroup analyses revealed that there was also a significantly lower sensitivity between the RECORDS and TRATING and TNOM [both 10% vs. 70%, Fisher's exact test (1, $\underline{N}$ = 20) = 7.5, $\underline{p}$<.01] for the ESL group. Similarly, for the L1 group, there was a significantly lower sensitivity between RECORDS

and TRATING [40 % vs. 100%, Fisher's exact test (1, N = 13) = 6.24]. Conversely, there were no significant differences between the sensitivities of the RECORDS and TNOM for this L1 group. Although the sensitivity of RECORDS was very low, specificity scores remained high and they were not different from the specificity of the other two screening methods.

Insert Table 6 about here

There were differences between the L1 and ESL groups in terms of the likelihood ratios for teacher screening (see Tables 4, 5, 6). According to accepted criteria for screening measures, the likelihood ratios of the TRATING and TNOM were moderate (8.5 and 7.7 respectively), but they were low for the ESL group (4.3 and 3.7 respectively). Conversely, the likelihood ratio of the RECORDS for the ESL group was extremely high (17.1), but it continued to be moderate for the L1 group (5.3). This high likelihood ratio for the ESL group may be accounted for by the high specificity combined with the extremely low sensitivity. Therefore, the use of the two teacher screening methods (i.e., TRATING and TNOM) for the L1 group is likely to increase the probability of the child being at-risk for a reading disability by a moderate amount, specifically 7.7 times more than if no such screening measure were utilized. However, similar methods used for ESL children would only increase their probability for being at-risk by approximately 4 times. Further, although the likelihood ratio of the RECORDS for the ESL group was extremely high, the trade-off for this increase was an extremely low sensitivity. Thus, when RECORDS is positive, the child

is highly likely to have a reading disability, but many children (i.e., 9 out of 10) will not be identified.

Time 2

Similar analyses were conducted one year later, in the fall of grade 2, on teacher data available for Cohort 1 using CSRS-2 as the gold standard (see Tables 4, 5, 6). The number of students from Cohort 1 for whom there was teacher and objective measure data at Time 2 was 142 for TRATING and RECORDS, and 144 for TNOM. The number of students that were designated as at-risk according to the CSRS-2 was 8 (4.3%). There was the same number of Cohort 1 at-risk students at Time 1 as at Time 2. However, using the Time 2 CSRS as the gold standard and the TRATING as the screening measure, 1 student designated as false negative at Time 1 was a true negative at Time 2. Also, 1 student designated as false positive at Time 1 was classified as true positive at Time 2. Similarly, with respect to TNOM screening, two students originally designated at Time 1 as false positives were changed to a designation of true positive. Conversely, 2 students designated as false negative at Time 1 were designated as true negative at Time 2. Thus, although the same overall number of children were at-risk by the gold-standard measure, the initial teacher assessments improved with respect to their correct designation of students as at-risk or not at-risk in the long term.

Sensitivity and specificity of teacher assessment at Time 2 using both TRATING and TNOM improved or stayed the same level for the total group as well as the ESL and L1 subgroups, although the differences were not statistically significant. As at Time 1, sensitivity for the ESL group was lower as compared to that of the L1 group (75% versus 100% for TRATING and 50% versus 100% for RECORDS, respectively). This trend was

not noted for the TNOM where sensitivity was 100% for both ESL and L1. Furthermore, there was a trend towards lower sensitivity of the RECORDS as compared with the TRATING (75% versus 87.5%) and TNOM (75% versus 100%), but this difference was not statistically significant. Similarly, although there were trends toward a lower sensitivity of RECORDS as compared with TRATING (50% versus 75%) and TNOM (50% versus 100%) in the ESL group, these differences were also not statistically significant. For the L1 group, RECORDS was just as sensitive and specific as the TRATING and TNOM (all being 100%) at Time 2.

The likelihood ratios at Time 2 were higher for the total and subgroups on all three screeners. At this second time period, using any screener increased the probability of identifying reading difficulty by at least 5 times. It is important to note that while the likelihood ratios were higher at Time 2 for all groups, this figure continued to be lower for the ESL as compared to the L1 group. So, while using the TRATING or TNOM increased the probability of detecting individuals with the reading difficulty (as opposed to those without it) by 10-fold in the L1 group, these same screeners increased the likelihood by a moderate 5 and 6 times in the ESL group. Again, the higher likelihood ratios of the RECORDS relative to the other screeners must be considered in combination with the concurrent reduced sensitivity. Although using RECORDS increases the odds of detecting individuals with (as opposed to those without) a positive objective test, this screener is also more likely to miss individuals.

Relationship of Oral Proficiency to Accuracy of Teacher Assessments

In order to test the hypothesis that oral language may be a factor involved in accurate and inaccurate identifications of at-risk students (by TRATING and TNOM), we examined if

there were differences among the group designations (false positive, false negative, true positive, true negative) with respect to their oral proficiency scores. Two comparisons were of particular interest. Firstly, accurate readers who were correctly classified (true negative) were compared with those who were incorrectly classified (false positive). Both groups of children were in the not at-risk reading group in the second grade. The second intended comparison was between the inaccurate readers who were correctly classified as at-risk by teachers (true positive) and inaccurate readers who were incorrectly identified as not at-risk for reading failure according to teacher nominations (false negative). However, due to the fact that there were no false negative TNOMs in this subsample and only 1 false negative TRATING, such analyses could not be completed.

First, a series of one-way ANOVAs were performed separately for ESL and L1 students using accuracy data (i.e., TNOM and TRATING) as the independent variable and oral proficiency measures (EOWPT, PPVT-R, sentence repetition) as the dependent variable. There were four levels or groups of each independent variable (i.e., true positives, true negatives, false positives and false negatives). Overall significant differences among the TNOM groups emerged for the ESL group but not for the L1 group. There were significant differences among the TNOM groups on the EOWPT [$F(2, 94) = 5.85$, $p<.01$], PPVT-R, [$F(2, 97) = 6.87$, $p<.01$] and sentence repetition scores [$F(2,95) = 7.6$, $p<.001$]. Specifically, post-hoc Sheffé analyses on the ESL group revealed that there were significant differences between the false positive group and the true negative group on sentence repetition (M=10.89 vs. 18.79, $p<.05$) and PPVT-R (M=19.0 vs. 77.0, $p<.05$) scores, but not on the EOWPT scores. These findings indicate that incorrectly identified adequate ESL readers, compared to their correctly identified peers, had significantly lower means on two of the three oral

proficiency measures. Similar analyses conducted with respect to the TRATING groups confirmed these findings. There were significant differences among the accuracy groups on all three oral proficiency scores: EOWPT, F (3, 91) = 5.36, $p<.01$; PPVT-R, F (3, 94) = 4.04, $p<.01$ and sentence repetition F (3, 92) = 3.40, $p<.05$, and this difference occurred only for the ESL group. Post hoc tests could not be performed on the TRATING data, to indicate more specifically which groups differed, as there were too few cases.

Discussion

The purpose of the present study was to examine the accuracy of various teacher assessment methods for screening children for reading disability. A second related purpose was to examine differences in the accuracy of teacher screening of reading disability for L1 as compared to ESL children. A third purpose was to determine the influence of oral language factors on the misclassification (i.e., false positive and false negative rates) of students. For the total sample of children, teacher sensitivity and specificity, using semi-structured interviews and rating scales for screening, were in the moderate to high range. The accuracy of teacher nominations was high when children were reassessed after 1 year, suggesting that teachers are able to predict relatively well students who might have long-term difficulties as opposed to students who only have transient difficulties in reading. Despite the favorable performance of teacher screening in the total sample, the teachers' sensitivity for identifying ESL children was substantially lower than that of the L1 children, regardless of the screening method used or the time of assessment. Of particular concern was the marked disparity between the accuracy of the structured screening methods as compared to the spontaneous expressed concerns of teachers (as indicated through interviews and scholastic records) at Time1, particularly for ESL students. Sensitivity of teachers'

spontaneous expressed concerns for the ESL group was only 10% indicating that 9 of 10 children at-risk for reading disability would not have been identified; this sensitivity was significantly different from the two other screening methods. Finally, errors in teacher assessment for ESL students were at least partly explained by over-reliance on oral proficiency in the assessment of reading performance.

The relatively high accuracy of teacher assessment of at-risk status in the overall group is consistent with the findings of previous research. Salvesen and Undheim (1994) studied the correspondence between teacher rating scales and low reading achievement in 603 Norwegian first language speaking grade 2 and 3 children. Consistent with our findings the authors found a high specificity (92%) for teacher assessments. However, the sensitivity in their study was considerably lower at 71%, indicating a high number of false negative assessments. In contrast to our study, Salvesen and Undheim (1994) used a lower cutoff for the teacher rating scale (≤ 3), a procedure that would be expected to increase sensitivity and lower specificity. The divergent findings of this study may be partially accounted for by the fact that Salvesen and Undheim (1994) studied second and third grade students who have more advanced reading decoding and comprehension skills. These added skills make it considerably more difficult for teachers to assess students' reading skills and the differences between children of varying levels of achievement. In contrast, students at the beginning of grade 1 have fewer reading skills and those considered at-risk may show more noticeable difficulties than their average achieving peers. Another factor that may account for the difference in the findings of the two studies is the varying referral practices and definitions of reading disability that exist in Norway and Canada.

Although consistent with previous research, the current study is the first to demonstrate differential accuracy in teacher screening for reading disability between first language and second language speakers. In this study, the sensitivity of teachers in detecting reading difficulties in ESL children was below acceptable levels of screening test accuracy. For example, while TRATING screening detected all L1 students with potential reading disability (100% sensitivity), more than a third of ESL children at-risk for reading disability were not detected by the same method. The sensitivity of a screening test gives primarily information on the usefulness of a negative judgment by teachers (i.e., if teacher screening is not sensitive, a negative screen cannot rule out the possibility of reading disability). However, one is often interested in the certainty of a diagnosis when the screening result is positive. The current study found that the value of a positive teacher screening at increasing certainty for the presence of reading difficulty (i.e., the likelihood ratio) is also lower for ESL children. For example, a teacher nomination of an ESL child being at-risk increases the probability that the child is truly at-risk by only about 4 times, a likelihood ratio considered to change one's probability only mildly (Jaeschke, et al., 1994). In contrast, an L1 student nominated as at-risk by the TNOM would have nearly an 8 fold increase in probability of being at-risk; this change in probability is moderate to high, and in clinical practice would certainly influence the decision regarding the child's true status.

Another important finding was the significant difference between the sensitivity of teacher nominations and rating scales relative to expressed concern (i.e., actual referral rates) regarding reading disabilities for the ESL, L1 children and the total group. The two forms of teacher screening were considerably more accurate than teacher expressed concern, and this discrepancy was more marked in the ESL group. This lower degree of sensitivity of the

teachers' expressed concern could be due to several factors. Firstly, it may represent a reporting bias whereby, despite teacher concerns and perhaps even intervention, these concerns are not recorded in the ISR files. Secondly, with respect to their ESL students who may already be receiving some service, teachers may not feel the necessity for a formal referral despite concerns about the child's skill development. Alternatively, teachers may be waiting for their ESL students to mature or develop oral language skills before making a referral. This may be particularly the case if teachers are less confident of their ability to tease apart oral language and reading skills among ESL children and thus accept some level of difficulty in this population to be normal. Of more concern, however, is the possibility that in actual practice, outside the artificial constraints of a study (i.e., without prompting or use of a structured assessment situation), teachers are not communicating their concerns about some of their students' reading development, particularly their ESL students. This may reflect over-correction for previously reported bias in identifying learning problems among immigrant and minority children (Cummins, 1984). At the same time, such communication is crucial in order to facilitate consultation and to implement pre-referral activities such as monitoring.

The results of the present study also suggest that reliance on oral language skills may account for the lower accuracy of teachers' screening of ESL students. This is evidenced by the finding that, among children who were not at-risk by our objective measure, there was a lower oral language proficiency for those rated falsely positive (children identified as at-risk when they truly were not) as compared to those rated truly negative. These results suggest that teachers incorrectly use oral language proficiency as their gauge for the child's overall academic performance, a practice that may result in over-identification of some children.

This hypothesis is confirmed through a comparison of the correlations among teacher ratings and performance on the objective measures; objective oral language scores were only modcrately associated with children's performance in reading and spelling while teacher-rated oral language ratings were highly correlated to their ratings of academic performance. These results are consistent with previous research findings of a bias or halo effect in teacher's ratings of academic performance (Salvesen & Undheim, 1994). In other words, teacher ratings of their students' performance in certain academic areas affect ratings in other areas, either in a positive or negative direction.

There are several limitations of the current study. Firstly, no gold standard for reading disability is widely accepted and thus differing definitions and objective measures are likely to yield different rates of reading disability designation. However, this study used multiple, standardized and reliable measures combined with relatively conservative criteria to minimize over-classifying of children as at-risk for reading disability. Secondly, reassessment of children at Time 2 revealed that the "gold standard" re-classified children into different at-risk categories. Using Time 2 CSRS as the gold standard and TRATING and TNOM as the screening measures, 2 and 4 students were re-classified, respectively. Although this suggests weaknesses in the reliability of the gold standard at Time 1, it is also quite possible that the study itself served as an intervention, raising teacher's awareness of at-risk children and precipitating pre-referral or remediation activities (i.e., informal educational programs) resulting in changes in at-risk status. Thirdly, follow up data was available for only a portion of the total cohort, resulting in a smaller sample size at Time 2. This factor limited the power of the study to detect significant differences between the subgroups of ESL and L1 on a number of measures. Nevertheless, the consistency in the trends toward a lower

sensitivity in the ESL children across the three teacher assessment methods adds validity to the conclusions. Fourthly, classifying children as ESL was complicated by the fact that a large majority of L1 students had families of non-English speaking backgrounds raising the possibility that they had been exposed to a first language other than English. This, however, is an unavoidable state of affairs reflecting universal demographic trends.

Educational Importance of the Study

The current study has several implications to the practice of school psychology. We recommend the use of screening for the identification of children who may be at-risk for reading disability. In particular, teacher screening of general populations of students, and particularly L1 students, is quite accurate and can be relied upon to correctly determine the at-risk status of primary level students. For ESL children, however, screening accuracy is lower and interpretation of teachers' designations of at-risk should be interpreted with caution. Teachers should consider having a lower threshold for consulting psychologists and monitoring students' performance more closely. For the L1 and ESL groups alike, however, the findings suggest that waiting for teachers to spontaneously express concerns regarding children is not an accurate screening measure as it is likely to overlook most students whom they feel have a problem. Rather, it is suggested that a structured interview or rating scales be used in school settings to elicit teacher nominations of at-risk status to more accurately screen for reading disabilities. When interpreting rating scales, it must be kept in mind that, particularly with ESL students, teachers may be biased and use oral language proficiency as a basis for their ratings of performance in all academic areas. The results of this study suggest that teachers need to be trained on reading development of ESL students and the relationship of oral language to other academic skills.

References

Algozzine, B., & Ysseldyke, J. E. (1986). The future of the LD field: Screening and diagnosis. Journal of Learning Disabilities, 19, 394-398.

Altman, D. G., Bland, J. M. (1994). Diagnostic tests 2: Predictive values. British Medical Journal, 309, 102.

American Psychological Association. (1985). Standards for educational and psychological tests. Washington, D.C. : American Psychological Association.

Bowers, P. G. (1995). Tracing symbol naming speed's unique contributions to reading disabilities over time. Reading and Writing: An Interdisciplinary Journal, 7, 189-216.

Bruck, M. & Genessee, F. (1995). Phonological awareness in young second language learners. Journal of Child Language, 22, 307-324.

Chitiri, H. F., Sun, Y., Willows, D. M., & Taylor, I. (1992). Word recognition in second language reading. In R. J. Harris (Eds.), Cognitive processing in bilinguals. North Holland, Amsterdam: Elsevier Science.

Cummins, J. (1984). Bilingualism and special education: Issues in assessment and pedagogy. Cleveland, England: Multilingual Matters Ltd.

Denckla, M. B., & Rudel, R. G. (1976). Rapid "automatized" naming (R.A.N.): Dyslexia differentiated from other language disabilities. Neuropsychologia, 14, 471-480.

Dickinson, D. K., & Snow, C. E., (1987). Interrelationships among prereading and oral language skills in kindergarten from two social classes. Early Childhood Research Quarterly, 2, 1-25.

Durgunoğlu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J., (1993). Cross-language transfer of Phonological awareness . Journal of Educational Psychology, 85, 453-465.

Dunn, L. M., & Dunn, L. M. (1981). Peabody Picture Vocabulary Test-Revised. Circle Pines, MN: American Guidance Service.

Elbro, C., Borstrom, I., & Petersen, D. K. (1998). Predicting dyslexia from kindergarten: The importance of distinctness of phonological representations of lexical items. Reading Research Quarterly, 33, 36-57.

Epps, S., Ysseldyke, J.E., & Algozzine, B. (1983). Impact of different definitions of learning disabilities on the number of students identified. Journal of Psychoeducational Assessment, 1, 341-352.

Felton, R. H., & Wood, F. B. (1989). Cognitive deficits in reading disability and attention deficit disorder. Journal of Learning Disabilities, 22, 3-13.

Ferroli, L., & Shanahan, T. (1987). Kindergarten spelling: Explaining its relation to first-grade reading. In J.E. Readance & R. S. Baldwin (Eds.), Research in literacy: Merging perspectives (Thirty-sixth Yearbook of the National Reading Conference). Rochester, NY: National Reading Conference.

Gardner, M. F. (1990). Expressive One-Word Picture Vocabulary Test-Revised (EOWPVT-R). California: Academic Therapy Publications.

Geva, E. (in press). Issues in the development of second language reading: Implications for instruction and assessment. To appear in T. Nunes (Ed.), Integrating literacy research and practice. Dordrecht: Kluwer.

Geva, E. (1998, April). Learning to read in a second language (L2) - Does L2 oral proficiency matter? Paper presented at the annual meeting of the Society for the Scientific Studies of Reading, San Diego, CA.

Geva, E., & Siegel, L. S. (in press). Linguistic and cognitive processing in learning to read English and Hebrew. Reading and Writing: An Interdisciplinary Journal.

Geva, E., & Clifton, S. (1993). The development of first and second language reading skills in early French Immersion. Canadian Modern Language Review, 50, 646-667.

Geva, E., Wade-Wooley, L., & Shany, M. (1993). The concurrent development of spelling and decoding in different orthographies. Journal of Reading Behavior, 25, 383-406.

Gholamain, M., & Geva, E. (in press). The role of orthography and cognitive factors in the concurrent development of basic reading skills in bilingual Persian-English children. Language Learning.

Gottardo, A., Stanovich, K. E., & Siegel, L. S. (1996). The relationships between phonological sensitivity, syntactic processing, and verbal working memory in the reading performance of third-grade children. Journal of Experimental Child Psychology, 63, 563-582.

Jaeschke, R., Guyatt, G. H., & Sackett, D. L. (1994). Users' guides to the medical literature III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients. Journal of the American Medical Association, 271, 703-707.

Johnson, J. S., & Newport, E. T. (1991). Critical period effects on universal properties of language: The status of subjacency in the acquisition of second language. Cognition, 39, 215-258.

Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. Journal of Educational Psychology, 78, 243-255.

Lindsay, G. A., & Wedell, K. (1982). The early identification of educationally 'at risk' children revisited. Journal of Learning Disabilities,15, 212-217.

Mantzicopoulos, P. Y., & Morrison, D. (1994). Early prediction of reading achievement: Exploring the relationship of cognitive and noncognitive measures to inaccurate classifications of at-risk status. Remedial and Special Education, 15, 244-251.

Meisels, S. J. (1988). Developmental screening in early childhood: The interaction of research and social policy. Annual Review of Public Health, 9, 527-550.

Meyer, M. S., Wood, F. B., Hart, L. S., & Felton, R. H. (1998). Selective predictive value of rapid automatized naming in poor readers. Journal of Learning Disabilities, 31, 106-117.

Petrulius-Wright, J, (1988). The role of language proficiency in the development of L1 and L2 literacy skills in young children. Unpublished master's thesis, Ontario Institute for Studies in Education/University of Toronto, Toronto, Ontario, Canada.

Sackett, D. L., Haynes, R. B. & Tugwell, P. (1985). Clinical epidemiology: A basic science for clinical medicine. Boston: Little, Brown.

Salvesen, K.A., & Undheim, J. O. (1994). Screening for learning disabilities with teacher rating scales. Journal of Learning Disabilities, 27, 60-66.

Schiff-Myers, N. B. (1992). Considering arrested language development and language loss in the assessment of second language learners. Language, Speech and Hearing Services in Schools, 23, 28-33.

Siegel, L. (1993). Phonological processing deficits as the basis of a reading disability. Special issue: Phonological processes and learning disability. Developmental Review, 13, 246-257.

Siegel, L. (1988). Evidence that IQ scores are irrelevant to the definition and analysis of reading disability. Special issue: Child development: When things go wrong. Canadian Journal of Psychology, 42, 201-215.

Snyder L. S., & Downey, D. M. (1997). Developmental differences in the relationship between oral language deficits and reading. Topics in Language Disorders, 17, 27-40.

Stanovich, K. E. (1992). Speculations on the causes and consequences of individual differences in early reading reading acquisition. In P. Gough, L. Ehri, & R. Trieman (Eds.), Reading Acquistion (pp. 307-342). Hillsdale, NJ: Erlbaum Associates.

Stanovich, K.E. (1988). Explaining the differences between the dyslexic and the garden variety poor reader: The phonological-core variable-difference model. Journal of Learning Disabilities, 21, 590-604.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. Reading Research Quarterly, 21, 360-406.

Stanovich, K. E., Cunningham, A. E., & Cramer, B. (1984). Assessing phonological awareness in kindergarten children: Issues of task comparability. Journal of Experimental Child Psychology, 38, 175-190.

Torgeson, J. K., Wagner, R. K. & Rashotte, C. A. (1994). Longitudinal studies of phonological processing and reading. Journal of Learning Disabilities, 27, 276-286.

Tunmer, W. E., & Nesdale, A. R. (1985). Phonemic segmentation skills and beginning reading. Journal of Educational Psychology, 77, 417-427.

Vellutino, F., & Scanlon, D. (1987). Phonological coding, phonological awareness, and reading ability: Evidence from a longitudinal and experimental study. Merrill-Palmer-Quarterly, 33, 321-363.

Wagner, R. K. & Torgeson, J.K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. Psychological Bulletin, 85, 192-212.

Westernoff, F. (1995). L1 loss: Implications for speech and language assessment. The Journal of Speech-Language Pathology and Audiology (in press).

Wilkinson, G. S. (1993). The Wide Range Achievement Test -Revision 3. U.S.A.: Jastak Associates.

Wolf, M. (1997). A provisional, integrative account of phonological naming deficits in dyslexia: Implications for diagnosis and intervention. In B. Blachman (Ed.), Cognitive and linguistic foundations of reading acquisition: Implications for intervention research (pp. 67-92). Hillsdale, NJ: Erlbaum.

Woodcock, R. W. (1987). Woodcock Reading Mastery Test. Circle Press, MN: American Guidance Service.

Ysseldyke, J. E. Algozzine, B., & Epps, S. (1983). A logical and empirical analysis of current practices in classifying students as handicapped. Exceptional Children, 50, 160-166.

Table 1

Demographic Information for 382 Children Assessed at Time1

	Total (N)	% Sample
Sex		
Male	187	48.95
Female	195	51.04
Country of Birth		
Canada	278	72.77
India	21	5.49
Sri Lanka	16	4.18
Pakistan	2	.52
Portugal	1	.26
Other	64	16.75

	Total (N)	% ESL group
First Language of ESL Grou		
Punjabi (or dialect of)	179	68.32
Cantonese/Chinese	46	17.56
Portuguese	21	8.01
Vietnamese	3	1.15
Tagalogue	2	.76
Other	11	4.20

Table 2

Means and Standard Deviations of the Teacher Rating Scale for the L1 and ESL groups (*p<.01)

Item		L1 (N=113)	ESL (N=262)
Oral Comprehension	M	4.90	4.06 *
	SD	1.52	1.48
Oral Expression	M	5.09	3.86 *
	SD	1.42	1.45
Vocabulary	M	5.01	3.81*
	SD	1.40	1.40
Reading Comprehension	M	4.37	3.76 *
	SD	1.72	1.53
Reading	M	4.15	3.98
	SD	1.84	1.69
Spelling	M	4.10	3.88
	SD	1.79	1.61
Writing	M	3.96	3.75
	SD	1.68	1.58
Grammar	M	4.49	3.50
	SD	1.48	1.48
Arithmetic	M	4.33	4.47
	SD	1.45	1.40
Overall	M	4.26	4.00
	SD	1.60	1.55

Table 3

Means and Standard Deviations on the Oral Language and Academic Measures for L1 and ESL Groups at Time 1

Raw scores		L1 (N=113)	ESL (N=262)
Expressive Vocabulary	M	47.33	31.23 *
	SD	13.18	13.98
Receptive Vocabulary	M	66.94	47.74 *
	SD	14.37	17.47
Sentence Repetition (0-52)	M	27.43	16.82 *
	SD	9.39	10.13
Word Attack (0-45)	M	5.31	5.65
	SD	6.51	7.66
WRAT3 (0-36)	M	18.61	19.46
	SD	4.68	4.79
Spelling (0-16)	M	2.32	2.43
	SD	2.44	3.36
RAN (time in sec)	M	49.47	44.99
	SD	21.38	20.71

*p<.001

Table 4

Two by Two Tables of Accuracy of Teacher Rating Scales-Reading (TRATING) Compared to the Combined Standardized Reading Score (CSRS) for Total Group, L1, and ESL Students at Time 1 (T1) and One Year Later (T2)

T1

Total Group (N=305)

	CSRS-1 +	CSRS-1 -
TRATING +	15	42
TRATING -	3	245

Sens – 83.33% (CI= 57.7 - 95.6)
Spec – 85.37% (CI= 80.6 - 89.1)
LR – 5.69 kappa - .34

L1 Students (N=102)

	CSRS-1 +	CSRS-1 -
TRATING +	8	11
TRATING -	0	83

Sens – 100% (CI = 59.8 – 100.0)
Spec – 88.29% (CI = 79.6 – 93.7)
LR – 8.5 kappa - .54

ESL Students (N=202)

	CSRS-1 +	CSRS-1 -
TRATING +	7	31
TRATING -	3	161

Sens – 70.00% (CI = 35.4 – 91.9)
Spec – 83.85% (CI = 77.7 – 88.6)
LR – 4.3 kappa - .23

T2

(N=142)

	CSRS-2 +	CSRS-2 -
TRATING +	7	16
TRATING -	1	118

Sens – 87.50% (CI= 46.7 - 99.3)
Spec - 88.06% (CI= 81.1 – 92.8)
LR – 7.33 kappa - .40

(N=44)

	CSRS-2 +	CSRS-2 -
TRATING +	4	5
TRATING -	0	35

Sens –100% (CI = 39.6 – 100.0)
Spec – 87.50% (CI = 72.4 – 95.3)
LR – 10.0 kappa - .56

(N=98)

	CSRS-2 +	CSRS-2 -
TRATING +	3	11
TRATING -	1	83

Sens – 75.00% (CI = 21.9 – 98.7)
Spec - 88.29% (CI = 79.6 – 93.7)
LR – 6.4 kappa - .29

Sens=Sensitivity; Spec=Specificity; CI= 95% confidence interval; LR=Likelihood ratio; kappa=kappa statistic for agreement; ESL=English as a Second Language students; L1=First language students; CSRS-1= Time 1 CSRS; CSRS-2=Time 2 CSRS.

Table 5

Two by Two Tables of Accuracy of Teacher Nominations (TNOM) Compared to the Combined Standardized Reading Score (CSRS) for Total Group, L1, and ESL Students at Time 1 (T1) and One Year Later (T2)

T1

Total Group (N=313)

	CSRS-1 +	CSRS-1 -
TNOM +	14	48
TNOM -	4	247

Sens – 77.77% (CI = 51.9 – 92.6)
Spec – 83.73% (CI = 78.9 – 87.7)
LR – 4.8 kappa - .29

L1 Students (N=105)

	CSRS-1 +	CSRS-1 -
TNOM +	7	11
TNOM -	1	86

Sens – 87.50% (CI = 46.7 – 99.3)
Spec – 88.66% (CI = 80.2 – 93.9)
LR – 7.7 kappa - .48

ESL Students (N=206)

	CSRS-1 +	CSRS-1 -
TNOM +	7	37
TNOM -	3	159

Sens – 70.00% (CI = 35.4 – 91.9)
Spec – 81.12% (CI = 74.8 – 86.2)
LR – 3.7 kappa - .20

T2

(N=144)

	CSRS-2 +	CSRS-2 -
TNOM +	8	23
TNOM -	0	113

Sens - 100% (CI = 59.8 – 100.0)
Spec - 83.09% (CI = 75.5 – 88.8)
LR – 5.9 kappa - .35

(N=44)

	CSRS-2 +	CSRS-2 -
TNOM +	4	4
TNOM -	0	36

Sens – 100% (CI = 39.6 – 100.0)
Spec – 90.00% (CI = 75.4 – 96.7)
LR – 10.0 kappa - .62

(N=100)

	CSRS-2 +	CSRS-2 -
TNOM +	4	19
TNOM -	0	77

Sens - 100% (CI = 39.6 – 100.0)
Spec - 80.21% (CI = 70.6 – 87.4)
LR – 5.1 kappa - .25

Sens=Sensitivity; Spec=Specificity; CI = 95% confidence interval; LR=Likelihood ratio; kappa=kappa statistic for agreement; ESL=English as a Second Language students; L1=First language students; CSRS-1= Time 1 CSRS; CSRS-2=Time 2 CSRS.

Table 6

Two by Two Tables of Accuracy of Teacher Expressed Concern (RECORDS) Compared to the Combined Standardized Reading Score (CSRS) for Total Group, L1, and ESL Students at Time 1 (T1) and One Year Later (T2)

T1 T2

Total Group (N=275) (N=142)

	CSRS-1 +	-
RECORDS +	3	19
-	12	241

Sens – 20.00% (CI = 5.3 – 48.6)
Spec – 92.69% (CI = 88.6 – 95.4)
LR – 2.7 kappa - .10

	CSRS-2 +	-
RECORDS +	6	8
-	2	126

Sens – 75.00% (CI = 35.6 – 95.5)
Spec - 94.03% (CI = 88.2 – 97.2)
LR – 12.6 kappa - ..51

L1 Students (N=94) (N=41)

	CSRS-1 +	-
RECORDS +	2	10
-	3	79

Sens – 40.00% (CI = 7.3 – 83.0)
Spec – 88.76% (CI = 79.9 – 94.2)
LR – 5.3 kappa - .17

	CSRS-2 +	-
RECORDS +	4	3
-	0	34

Sens – 100% (CI = 39.6 – 100.0)
Spec – 91.89% (CI = 77.0 – 97.9)
LR – 12..3 kappa - .69

ESL Students (N=181) (N=101)

	CSRS-1 +	-
RECORDS +	1	9
-	9	162

Sens – 10.00% (CI = .5 – 45.9)
Spec – 94.74% (CI = 89.9 – 97.4)
LR – 17.1 kappa - .05

	CSRS-2 +	-
RECORDS +	2	5
-	2	92

Sens – 50.00% (CI = 9.2 – 90.8)
Spec - 94.85% (CI = 87.8 – 98.1)
LR – 9.7 kappa - .33

Sens=Sensitivity; Spec=Specificity; CI = 95% confidence interval; LR=Likelihood ratio; kappa=kappa statistic for agreement; ESL=English as a Second Language students; L1=First language students; CSRS-1=Time 1 CSRS; CSRS-2=Time 2 CSRS.