

# Accuracy of Teacher Assessments of Second-Language Students at Risk for Reading Disability

Marjolaine M. Limbos and Esther Geva

## Abstract

This study examined the accuracy of teacher assessments in screening for reading disabilities among students of English as a second language (ESL) and as a first language (L1). Academic and oral language tests were administered to 369 children (249 ESL, 120 L1) at the beginning of Grade 1 and at the end of Grade 2. Concurrently, 51 teachers nominated children at risk for reading failure and completed rating scales assessing academic and oral language skills. Scholastic records were reviewed for notation of concern or referral. The criterion measure was a standardized reading score based on phonological awareness, rapid naming, and word recognition. Results indicated that teacher rating scales and nominations had low sensitivity in identifying ESL and L1 students at risk for reading disability at the 1-year mark. Relative to other forms of screening, teacher-expressed concern had lower sensitivity. Finally, oral language proficiency contributed to misclassifications in the ESL group.

Recently, major interest has developed regarding the early identification of learning disabilities among second-language learners. Historically, minority children in certain immigrant groups have been discriminated against and regarded as educationally inferior and have been much more likely to be identified with learning disabilities 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. Furthermore, 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 children learning English as a second language (ESL) with learning disabilities for up to 4 or 5 years in order to allow time for proficiency in the language to develop. It has been argued that only then can such difficulties reliably be attributed to a learning difficulty. Although this

growing awareness of the language needs of ESL children indicates an increased sensitivity by professionals, it has also led to a tendency to ignore the possibility that ESL learners are having difficulty not only due to insufficient oral language proficiency. Instead, word decoding or language processing problems typical of a learning disability may be present. Delaying assessment for these children takes away preventive opportunities and possibilities of instituting remediation strategies. Consequently, professionals are reconsidering the early identification of reading disabilities among second-language students.

The process of early identification, however, is highly complex, with major concerns related to measurement, prediction, and 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).

There has been much research in the first-language literature regarding the processes involved in reading disabilities. Phonological processing refers to an individual's mental operations that use the phonological or sound structure of oral language when he or she is learning to decode written language. It is well established that there is a causal relationship between phonological awareness and reading, particularly from kindergarten through the second grade (Elbro, Borstrom, & Petersen, 1998; Stanovich, 1988; Wagner & Torgesen, 1987). However, other research has suggested that there may be other processes involved in reading disabilities in addition to phonological awareness. For example, Wolf and Bowers' double-deficit hypothesis (Bowers, 1995; Bowers & Wolf, 1993; Wolf & Bowers, 1999) has proposed that there are deficits associated with the timing of various reading subprocesses, such as letter recognition and serial scanning of print, that are independent of

deficits in phonological skill. There is research that suggests that there may be subgroups of children with specific deficits in either naming speed or phonological awareness and in both naming speed and phonological awareness (Bowers, 1995; Wolf & Bowers, 1999). Although the degree to which phonological awareness and rapid serial naming contribute unique or shared variance is unclear, these factors consistently emerge as two of the most powerful predictors of first grade reading achievement (Bowers, 1995; Felton & Wood, 1989; Meyer, Wood, Hart, & Felton, 1998; Torgesen, Wagner & Rashotte, 1994; Wolf, 1997).

With the establishment of a robust relationship between phonological skills, rapid naming, and reading in first-language learners (L1), researchers are beginning to examine whether similar predictors apply to second-language learners. Recent studies with bilingual and second-language learners have suggested that universal cognitive and linguistic factors, such as phonological processing, working memory, orthographic knowledge, and speed of lexical access, are involved in reading skills acquisition for both L1 and ESL children (Bruck & Genesee, 1995; Chitiri, Sun, Willows, & Taylor, 1992; Durgunoglu, Nagy, & Hancin-Bhatt, 1993; Geva & Clifton, 1993; Geva & Siegel, 2000; Geva, Wade-Wooley, & Shany, 1993; Geva & Yaghoub-Zadeh, in press; Gholamain & Geva, 1999).

Given the strong data indicating the role of phonological awareness and naming speed in predicting subsequent reading abilities found in the first-language literature and the 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 students with learn-

ing disabilities 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 with learning disabilities also meet the same criteria (Epps, Ysseldyke, & Algozzine, 1983), and many typical 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 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 and rapid naming, the current research considers these to be two of the best early indicators of reading difficulty.

In practice, not all children are formally assessed for adequate development of phonological awareness or naming speed 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 is through teacher assessments, as obtained through verbal nominations or rating 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 has been only one well-designed study of teacher screening for reading disability, which was conducted by Salvesen and Undheim (1994). These authors found that the degree of correspondence between teacher assessments of low achievement and 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) 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 used correlation coefficients. More recently, epidemiologists and researchers investigating early detection and intervention have emphasized the utility of other indices of accuracy, such as sensitivity, specificity, likelihood ratio, and the kappa statistic (Jaeschke, Guyatt, & Sackett, 1994; Mantzicopoulos & Morrison, 1994; Meisels, 1988; Sackett, Haynes, & Tugwell, 1985). These indices offer an enhanced interpretation of the results of a screening test, extending beyond the relatively nonspecific 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 quantify the diagnostic ability of the test and have important clinical value. The sensitivity of a test provides information regarding the ability of the test to identify people at risk for a diagnosis, and the specificity of a test confirms the presence of the diagnosis. Screening tests with a high sensitivity give valuable information on the importance of a negative screen-

ing result; if a student is said not to be at risk, there is a high likelihood that he or 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, because many students could still be falsely positive. That is, they could be identified as at risk when, in fact, they are not.

To better define the usefulness of a positive screening result, 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 children. 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) indicated 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. Mantziopoulos and Morrison (1994) warned, 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 with falsely negative results 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 has 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 students. This subgroup presents several unique challenges to the assessment process, as the correct iden-

tification 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 as 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 only when students have mastered the basic oral communication skills they will be 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 has indicated 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, 1997). For example, oral language skills such as narrative and communicative adequacy have not been found to correlate significantly with prereading variables such as phonemic awareness, print production, and decoding (Dickinson & Snow, 1987). Other research has indicated that such oral language skills as vocabulary and grammatical knowledge were either marginally related or not related to word-identification performance (Durgunoglu et al., 1993; Geva & Siegel, 2000; Gholamain & Geva, 1999). 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). Although syntactic awareness may contribute to decoding, it appears to be secondary to phonological awareness (Gottardo et al., 1996). The difficulties in assessing oral language and in understanding the association between oral language and reading in ESL students may account for these children being either overlooked or 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 students 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 of the teacher assessment of reading achievement between 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.

## Method

### *Participants and Procedure*

**Students.** The battery of cognitive, linguistic, and reading tasks was administered to a total sample of 422 children. Three cohorts of Grade 1 children were tested at 1-year intervals. Concurrently, teacher interviews were conducted and rating scales were administered (Time 1). Fifty-three students were excluded from the analysis because of incomplete reading data or unclear language designation, leaving a total of 369. There were 120 (32.5%) L1 and 249 (67.4%) ESL students. The

same battery was administered 1½ years later, in the spring of Grade 2 (Time 2), to 291 children (87 L1 and 204 ESL). Seventy-eight of the initial sample did not complete second testing, equivalent to an attrition rate of 21.14 %.

The demographic characteristics of all the participants are summarized in Table 1. At Time 1, children were between 5 years 8 months and 7 years 3 months old ( $M = 77.23$ ,  $SD = 4.03$  months) attending Grade 1 in 12 general education schools in three different areas of a large metropolitan multi-ethnic city. At Time 2, children were between the ages of 7 years 4 months and 8 years 5 months old ( $M = 93.48$ ,  $SD = 3.51$  months). There was no significant difference in the scores of the ESL and L1 students on the nonverbal intelligence test (Matrix Analogies Test-Expanded Form; Naglieri, 1989). The mean standard scores for the ESL and L1 groups were 99.39 ( $SD = 11.61$ ) and 99.52 ( $SD = 10.52$ ), respectively.

An ESL student was defined as a student whose first spoken language was not English. It should be pointed out that many of the participants 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 (62.72%), followed by Portuguese (16.03%), Cantonese (16.03%), and other languages (5.23%). Although the first language of L1 students was English, many of these students had an ethnic origin other than Canadian. In the L1 group, the ethnic origins included White (41.7%), South-East Asian (15.70%), Portuguese (8.00%), Italian (3.90%), Caribbean (15.00%), Filipino (1.60%), and Other (4.00%). Of the entire sample, 317 (76.57%) were born in Canada, 21 (5%) in India, 16 (3.63%) in Sri Lanka, 2 (0.48%) in Pakistan, and 58 (14.01%) in other countries.

**Teachers.** Fifty-one 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 (11.76%) had a master's degree in addition to their teacher education, 17 (33.33%) had special education certification, and 17 (33.33%) had ESL teacher certification.

### Instruments

As part of a larger longitudinal research project investigating oral language and literacy development among ESL learners, students completed a variety of oral language, reading, and written expression indices. The battery of tests was administered individually during two half-hour sessions. Of relevance to the current investigation are the following measures of oral proficiency, phonological awareness, rapid naming, word recognition, spelling, and nonverbal intelligence:

**Oral proficiency.** To assess oral proficiency skills, the following three 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 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-R; 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 of pictures 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 comprehen-

**TABLE 1**  
Demographic Information for the ESL and L1 Children Assessed  
at Time 1 and Time 2

	<i>n</i>	%
Gender		
Male	198	47.83
Female	216	52.17
Country of Birth		
Canada	317	76.57
India	21	5.00
Sri Lanka	16	3.63
Pakistan	2	0.48
Portugal	2	0.48
Other	56	13.53
First Language of ESL Group <sup>a</sup>		
Punjabi (or dialect of)	180	62.72
Cantonese	46	16.03
Portuguese	46	16.03
Vietnamese	3	1.05
Other	12	4.18

Note. ESL = English as a second language. L1 = English as a first language.

<sup>a</sup>percentages reported for this item are of the ESL group only rather than of the total sample.

sion. Participants in this test are asked to point to the picture corresponding to the vocabulary word given.

**Phonological Awareness.** 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 five consecutive errors. Pseudoword decoding is taken to be a reliable measure of phonological skills.

**Rapid Naming.** As a measure of rapid naming, the letter version of the Rapid Automatized Naming task (RAN; Denckla & Rudel, 1976) was administered. In this continuous naming task, children are asked to name as fast as they can a series of 5 letters that they have first been successful in naming during a trial. There are five rows, each containing 10 letters, making a total of 50 items. The time (in seconds) to complete the naming task was recorded.

**Word Recognition.** To assess children's ability to read words in English, the Word Recognition subtest of the Wide Range Achievement Test-3 (WRAT-3; Wilkinson, 1993) was used. This test consists of 42 unrelated words, ranging 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.

**Spelling.** A variation of the Developmental Spelling Test (Ferrolli & 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 16 simple and highly frequent words that 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 Status

**Teacher Identification of At-Risk Status.** Three sources of information were used to yield three types of teacher at-risk classifications. First, teacher interviews and individual scholastic records (ISR) provided information on concerns 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 teacher 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 ranging from 1 to 7 (1 = *very poor*; 4 = *medium*; 7 = *very high*) for each domain was used. Teachers were asked to rate each student according to his or her expectations at this level and 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), whereas scores above 2 were classified as not at risk.

Third, during semistructured interviews, teachers were asked to nominate children they felt to be at risk for a significant reading difficulty or disability. Teachers were asked specifically whether any particular student was at risk for a learning disability or

for more global intellectual deficits; only the former students were included in the study. Any pupil the teacher nominated as being at risk for the development of a reading difficulty or disability in the interview was classified as Teacher Nomination (TNOM) at risk. Children identified with difficulties other than in reading (e.g., oral language, mathematics) were not excluded from the study. The two children who were eliminated from the sample were described as being at risk for social, emotional, or behavioral reasons, and not reading difficulties.

**Objective Determination of At-Risk Status.** In order to identify at-risk status objectively, a combined standardized reading score (CSRS) based on the WRAT-3, Word Attack, and RAN-Letter 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 at or below the 10th percentile. These standardized scores were calculated at Time 1 (CSRS-1) and then again at Time 2 (CSRS-2). The CSRS-2 at-risk designation was considered the objective "gold standard" assessment of at-risk status for the purpose of this study. This is in keeping with the recommendation of Stanovich (1999) that a cutoff at the 10th or 15th percentile, on at least one of the reading tests, be used.

The mean scores on the Matrix Analogies Test-Expanded Form were similar between at-risk and not-at-risk students at Time 1 ( $M = 96.15$ ,  $SD = 11.65$ , and  $M = 99.89$ ,  $SD = 11.15$ , respectively), and Time 2 ( $M = 97.73$ ,  $SD = 11.61$ , and  $M = 99.99$ ,  $SD = 11.00$ , respectively).

### Data Analysis

Intercorrelations were calculated among the various academic areas (e.g., oral expression and reading) that teachers rated. Teacher ratings were also correlated with the test results at Time 1 and at Time 2 for the entire group and for each subgroup. Pearson product-

moment correlations were performed on the entire group as well as on 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 matrices were constructed sequentially using teacher screening of at-risk or not-at-risk status and objective (or gold standard) assessment (i.e., CSRS-1, CSRS-2) of at-risk or not-at-risk status, as shown in Figure 1.

From these matrices, sensitivity, specificity, and likelihood ratios (LR) were calculated. These screening characteristics were calculated at Time 1 and Time 2, for the total group, for ESL and L1 groups separately, and for each form of teacher assessment method (i.e., TNOM, TRATING, and RECORDS). *Sensitivity* was defined as the proportion of true positives that were correctly identified by the objective test (i.e., the ability of the test to identify those who are at risk) and was calculated as the number of true positives divided by the sum of true positives and false negatives. *Specificity* was defined as the proportion of true negatives that were correctly identified by the test (i.e., the ability of the test to identify correctly those who are not at risk) and was calculated as the number of true negatives divided by the sum of true negatives and false positives. *Likelihood ratios* (LR) were calculated as sensitivity divided by (1 - specificity).

Accuracy results are presented as percentages with 95% confidence intervals

(95% CI) in brackets. As a guide to the desired levels of specificity and sensitivity of a screening test, the American Psychological Association (1985) has suggested that 80% sensitivity and 90% specificity are preferable levels for psychological tests. For the purposes of interpretation in this study, values below 70 are considered low, 70 to 85 moderate, and above 85 high.

Furthermore, as a guideline for interpretation, a LR of 1 to 2 is considered to alter pretest probability minimally; a LR of 2 to 5 results in small changes in probability; a LR of 5 to 10 results in moderate shifts in probability; and a LR greater than 10 causes large and often conclusive changes in probability (Jaeschke et al., 1994). The  $\chi^2$  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, as well as between the various screening methods for a given language group.

Comparisons of the efficiency of the screening tests used were 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 below zero show disagreement; near zero reflect agreement only by chance; between .2 and .4 indicate fair agreement, between .4 and .6 moderate agreement; between .6 and .8 substantial agreement, and close to 1 reflect almost perfect agreement be-

tween tests (Sackett, 1992). The kappa index for agreement between the teacher nominations of at-risk status for a learning difficulty and low reading achievement, and between teacher assessment of reading (cutoff point at score value  $\leq 2$ ) and low reading achievement were calculated. The kappa index was calculated separately for ESL and L1 students at Time 1 and Time 2.

Data analyses were also conducted to examine hypotheses regarding factors that could account for differing accuracy rates. To test the hypothesis that oral language may be a factor involved in the accuracy of the identification of at-risk students, 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. First, 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 the inaccurate readers who were incorrectly identified as not at risk according to teacher nominations (false negative).

A series of one-way analyses of variance (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). There were 10 true positives, 3 false negatives, 23 false positives, and between 122 and 125 true negatives depending on the dependent measure used. The small number of false negatives limited the number of comparisons somewhat. Results were considered significant at  $p < .05$ .

**FIGURE 1**  
Evaluation Matrix for Data Analysis

Teacher Screening Measure <sup>a</sup>	Objective Determination <sup>b</sup>	
	At risk	Not at risk
At risk	True positive	False positive
Not at risk	False negative	True negative

<sup>a</sup>Teacher Nomination (TNOM), Teacher Rating Scale-Reading (TRATING), or Scholastic Records (RECORDS). <sup>b</sup>Combined standardized reading score at Time 1 (CSRS-1) or at Time 2 (CSRS-2).

**TABLE 2**  
Means and Standard Deviations  
of the Teacher Rating Scale for the  
L1 and ESL Groups

Item	L1	ESL
Oral Comprehension		
M	5.00	4.06**
SD	1.49	1.53
Oral Expression		
M	5.17	3.88**
SD	1.40	1.49
Vocabulary		
M	5.11	3.84**
SD	1.38	1.45
Reading Comprehension		
M	4.48	3.76**
SD	1.68	1.58
Reading		
M	4.32	3.95
SD	1.82	1.73
Spelling		
M	4.22	3.85*
SD	1.75	1.66
Writing		
M	4.02	3.73
SD	1.65	1.62
Grammar		
M	4.51	3.51**
SD	1.52	1.50
Arithmetic		
M	4.42	4.46
SD	1.44	1.44
Overall <sup>a</sup>		
M	4.48	3.97*
SD	1.58	1.60

Note. L1 = English as a first language group,  $n = 12$ .  
ESL = English as a second language group,  $n = 236$ .  
<sup>a</sup>L1,  $n = 70$ . ESL,  $n = 122$ .  
\* $p < .05$ . \*\* $p < .01$ .

## Results

### Correlations

**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 groups are provided in Table 2. There were significant differences between the L1 and ESL groups

in terms of teacher rating scales of oral comprehension,  $F(1, 354) = 30.80$ ,  $p < .001$ , oral expression,  $F(1, 354) = 62.27$ ,  $p < .001$ , reading comprehension,  $F(1, 347) = 15.05$ ,  $p < .001$ , vocabulary,  $F(1, 354) = 62.36$ ,  $p < .001$ , grammatical sentence structure ratings,  $F(1, 353) = 34.77$ ,  $p < .001$ , spelling,  $F(1, 353) = 3.78$ ,  $p < .05$ , and overall performance,  $F(1, 190) = 4.60$ ,  $p < .05$ . Although the means on teacher ratings of reading, writing, and arithmetic were somewhat lower in the ESL than in the L1 group, there were no significant differences between the two groups in these domains.

Teacher ratings of academic performance were moderately to highly correlated with each other (range of  $r = .51$  to  $.91$ ) for the entire group, the L1, and the ESL groups. For both language groups, the correlations between the oral expression ratings and academic subjects (i.e., math,  $r = .51$ ; reading,  $r = .65$ ; and spelling,  $r = .65$ ) were lower than those with other oral measures (i.e., oral comprehension,  $r = .80$ ; vocabulary,  $r = .92$ ) but continued to be in the high range. Furthermore, oral proficiency ratings (i.e., oral expression and comprehension) correlated highly with the child's overall academic rating for the L1 ( $r = .69$  and  $.72$ , respectively) and the ESL groups ( $r = .76$  and  $.85$ , respectively).

**Correlations Among Objective Measures.** Examination of the correlations among the standardized academic and oral proficiency scores at Time 1 provided a different profile from correlations between similar variables on the teachers' ratings. Correlations among the receptive and expressive oral language measures (i.e., EOWPT, PPVT-R, and sentence repetition) were high and significant (range =  $.67$ – $.76$ ), and those between the reading measures (i.e., Word Attack and WRAT-3) were also high and significant ( $r = .75$ ). 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

( $r = .24$ – $.34$ ). Thus, the correlations between the objective measures of oral and reading skills were lower than between the 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,  $F(1, 372) = 105.81$ ,  $p < .001$ ; PPVT-R,  $F(1, 380) = 108.65$ ,  $p < .001$ ; sentence repetition,  $F(1, 376) = 100.20$ ,  $p < .001$ ), with the L1 group having higher scores in each of these areas. There were no significant differ-

**TABLE 3**  
Means and Standard Deviations  
on Objective Measures for L1 and  
ESL Groups at Time 1

Raw scores	L1	ESL
Expressive Vocabulary		
M	47.47	32.03***
SD	13.03	14.02
Receptive Vocabulary		
M	66.65	48.13***
SD	14.28	17.13
Sentence Repetition (0–52)		
M	27.45	16.79***
SD	9.08	9.92
Word Attack (0–45)		
M	5.53	5.35
SD	6.69	7.43
WRAT-3 (0–36)		
M	18.73	19.20
SD	4.60	4.77
Spelling (0–16)		
M	2.32	2.43
SD	2.43	3.36
RAN (time in sec)		
M	48.81	45.72
SD	20.93	20.75

Note. L1 = English as a first language group,  $n = 124$ .  
ESL = English as a second language group,  $n = 258$ .  
\*\*\* $p < .001$ .

ences between the two groups on the three reading measures (i.e., Word Attack, WRAT-3, RAN).

For the entire group, there were moderate to high correlations (range of  $r = .44$  to  $.71$ ) between teacher ratings (i.e., TRATING) and objective test results at Time 1 for the respective areas. 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-R) was moderate ( $r = .44$ ) for the L1 group, whereas it was high ( $r = .55$ ) for the ESL group, and the relationship between teachers' assessment of grammar and the sentence repetition task was moderate ( $r = .27$ ) for the L1 group but high for the ESL group ( $r = .55$ ). 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 WRAT-3 scores were  $-.66$ ,  $.59$ , and  $.70$ , and for the ESL group these correlations were  $-.60$ ,  $.62$ , and  $.71$ , respectively.

The correlations between teacher ratings completed at Time 1 and objective test results at Time 2 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 lower 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 37 students (8.4% of the total sample) as at

risk for reading disability at Time 1, 26 (70.27%) of whom were ESL students. In the accuracy data, however, the number of students identified by the CSRS-1 as at-risk depended on data being available (i.e., both teacher and objective measures) for each participant. Designation of at-risk status by teachers varied with the screening measure used. The respective number of students designated at risk by teacher screening at Time 1 was 73 (16.6%), 75 (17.00%), or 24 (5.4%), depending on whether the teacher rating scale (TRATING), teacher nominations (TNOM), or scholastic record data (RECORDS) were used.

**Time 1.** Accuracy data (i.e., sensitivity, specificity) on the TRATING index were calculated from  $2 \times 2$  matrices of the number of correctly and incorrectly identified students (see Figure 2 and Table 4). At Time 1, the specificity values of the TRATING, using CSRS-1 as the criterion, were high for the total, L1, and ESL groups. In terms of the sensitivities, subgroup analysis revealed a higher sensitivity for the L1 group than for the ESL group (sensitivity = 90.9% vs. 70.8%, respectively), although this difference was not statistically significant. The sensitivity of the entire group was in the moderate range (77.1%).

As can be seen in from Figure 3 and Table 5, which provide accuracy data on the TNOM, screening sensitivity fell below 80% for the total group. Similar to the TRATING, sensitivity of the TNOM was high at 90.9% for the L1 group and low at 66.7% for the ESL group. These differences, however, were not statistically significant. Similar to the results for the TRATING, the specificity values of the TNOM were high for all groups.

Compared to the already relatively low sensitivity of teacher screening using structured assessment (i.e., TRATING and TNOM), the accuracy of spontaneously expressed concern (i.e., RECORDS) was even lower (see Figure 4 and Table 6). Using RECORDS, sensitivity for the entire group was extremely low at 21.4%. Analysis of the

subgroups showed that sensitivity was higher in the L1 group (37.5%) compared to the ESL group (15.0%). Further analyses revealed that there were statistically significant differences in teacher sensitivity between RECORDS and TRATING for the total,  $\chi^2(1, N = 96) = 6.53, p < .01$ , and the ESL group,  $\chi^2(1, N = 64) = 5.54, p < .01$ . Similarly, there was a statistically significant difference between the sensitivity of RECORDS and TNOM for the total,  $\chi^2(1, N = 95) = 6.10, p < .01$ , and the ESL group,  $\chi^2(1, N = 63) = 5.04, p < .05$ . Conversely, there were no significant differences between the sensitivities of the RECORDS and TNOM for the L1 group. Although the sensitivity of RECORDS was very low, specificity scores remained high and were not different from the specificity scores of the other two screening methods.

There were differences between the L1 and ESL groups in terms of the likelihood ratios for teacher screening (see Tables 4, 5, and 6). According to accepted criteria for screening measures, the likelihood ratios of TRATING and TNOM were high for L1 (10.4 and 10.8, respectively), but they were low to moderate for the ESL group (5.3 and 4.3, respectively). Therefore, a positive result using either of the teacher screening methods (i.e., TRATING or TNOM) for the L1 group is likely to result in a large increase in the likelihood of the child being classified as at risk for a reading disability. However, similar methods used for ESL children would only increase their probability for being classified as at risk by a small to moderate amount.

**Time 2.** Similar analyses were conducted 1½ years later, in the spring of Grade 2, on teacher data available using CSRS-2 as the gold standard (see Tables 4 to 7). The number of students for whom there were teacher and objective measure data at Time 2 was 242 for TRATING, 235 for RECORDS, and 244 for TNOM. The number of students that were designated as at risk according to the CSRS-2 was 29 (6.6% of the total sample), 13 (44.83%) of whom

were L1 students and 16 (55.17%) ESL students.

Sensitivity of teacher assessment at Time 2 using either TRATING or TNOM

stayed at the same low to moderate level for the total and ESL groups, but dropped substantially for the L1 group (69.2% for both groups for TRATING;

69.2% and 76.9%, respectively, for the L1 and ESL groups for TNOM). These differences were not statistically significant, however. The sensitivity values for the ESL group were relatively equivalent to those of the L1 group for TRATING (69.2% for both groups) and TNOM (69.2% and 76.9%, respectively); they were all below recommended standards for screening measures. However, the sensitivity of RECORDS for the L1 group was somewhat higher than for the ESL group (36.4% and 15.4%, respectively), and they were both well below acceptable standards. Furthermore, there was a trend towards lower sensitivity of RECORDS as compared with TRATING (25.0% vs. 69.2%) and TNOM (25.0% vs. 73.1%) for the overall group, but this difference was only significant for TNOM,  $\chi^2(1, N = 75) = 4.00, p < .05$ . Within the respective language groups, there was also a trend toward lower sensitivity of RECORDS as compared

**TABLE 4**  
Accuracy Measures for Teacher Rating Scales-Reading (TRATING) for Total Group and Subgroups at Time 1 and Time 2

Measure	Time 1			Time 2		
	L1	ESL	Total group	L1	ESL	Total group
<i>n</i>	115	225	341	83	159	242
Sensitivity %	90.9	70.8	77.1	69.2	69.2	69.2
CI	57.1-99.5	48.8-86.6	59.4-89.0	38.9-89.6	38.9-89.6	48.1-84.9
Specificity %	91.3	86.6	88.2	91.4	85.6	87.5
CI	83.8-95.7	80.9-90.8	84.0-91.5	81.6-96.5	78.6-90.7	82.2-91.5
LR	10.4	5.3	6.5	8.0	4.8	5.5
Kappa	.6	.4	.5	.6	.3	.4

Note. L1 = English as a first language. ESL = English as a second language. CI = 95% confidence interval. LR = likelihood ratio. Kappa = Kappa index of agreement.

**FIGURE 2**  
Evaluation Matrices for Teacher Rating Scales-Reading (TRATING) for Total Group and Subgroups at Time 1 and Time 2

TOTAL GROUP					
TRATING	CSRS-1		TRATING	CSRS-2	
	at risk	not at risk		at risk	not at risk
at risk	27	36	at risk	18	27
not at risk	8	270	not at risk	8	189
L1 GROUP					
TRATING	CSRS-1		TRATING	CSRS-2	
	at risk	not at risk		at risk	not at risk
at risk	10	9	at risk	9	6
not at risk	1	95	not at risk	4	64
ESL GROUP					
TRATING	CSRS-1		TRATING	CSRS-2	
	at risk	not at risk		at risk	not at risk
at risk	17	27	at risk	9	21
not at risk	7	174	not at risk	4	125

L1 = English as a first language. ESL = English as a second language. CSRS-1 = combined standardized reading score at Time 1. CSRS-2 = combined standardized reading score at Time 2.

with TRATING and TNOM for the ESL (15.4% vs. 69.2%) and TNOM (15.4% vs. 76.9%). The difference between the sensitivities of TRATING and TNOM compared with RECORDS for the ESL

group was statistically significant ( $p < .05$ , Fisher's exact test). For the L1 group, RECORDS was also lower than TRATING (36.4% vs. 69.2%) and TNOM (36.4% vs. 69.2%), but not quite as low as for the ESL group. Specificity of teacher assessment continued to be high for both ESL and L1 students using all forms of teacher assessment methods, including RECORDS.

A further analysis was completed examining the sensitivity of using a combination of TNOM or TRATING. In contrast to the other screening instruments, the accuracy data were high (see Figure 5 and Table 7). For the overall sample, L1, and ESL groups, the sensitivity values were 92.3%, 84.6%, and 100.0%, respectively. The specificities were somewhat lower than for the individual TNOM and TRATING scores, ranging from 66.9% for ESL to 80.0% for L1 students.

The likelihood ratios at Time 2 were comparable to those at Time 1 for the

**TABLE 5**  
Accuracy Measures for Teacher Nominations (TNOM) for Total Group and Subgroups at Time 1 and Time 2

Measure	Time 1			Time 2		
	L1	ESL	Total group	L1	ESL	Total group
<i>n</i>	118	229	349	83	100	244
Sensitivity %	90.9	66.7	74.3	69.2	76.9	73.1
CI	57.1-99.5	44.7-83.6	56.4-86.9	38.9-89.6	46.0-93.8	51.9-87.6
Specificity %	91.6	84.4	86.9	92.9	84.5	87.2
CI	84.2-95.8	78.5-88.9	82.6-90.4	83.4-97.3	77.4-89.7	81.8-91.2
LR	10.8	4.3	5.7	9.7	4.9	5.7
Kappa	.6	.4	.4	.6	.4	.4

Note. L1 = English as a first language. ESL = English as a second language. CI = 95% confidence interval. LR = likelihood ratio. Kappa = Kappa index of agreement.

**FIGURE 3**  
Evaluation Matrices for Teacher Nominations (TNOM) for Total Group and Subgroups at Time 1 and Time 2

TOTAL GROUP					
TNOM	CSRS-1		TNOM	CSRS-2	
	at risk	not at risk		at risk	not at risk
at risk	26	41	at risk	19	28
not at risk	9	273	not at risk	7	190
L1 GROUP					
TNOM	CSRS-1		TNOM	CSRS-2	
	at risk	not at risk		at risk	not at risk
at risk	10	9	at risk	9	5
not at risk	1	98	not at risk	4	65
ESL GROUP					
TNOM	CSRS-1		TNOM	CSRS-2	
	at risk	not at risk		at risk	not at risk
at risk	16	32	at risk	10	23
not at risk	8	173	not at risk	3	125

L1 = English as a first language. ESL = English as a second language. CSRS-1 = combined standardized reading score at Time 1. CSRS-2 = combined standardized reading score at Time 2.

total and subgroups using all three screening methods. The likelihood ratios continued to be lower for the ESL group as compared to the L1 group. So, although using TRATING or TNOM increased the odds of detecting individuals with a reading difficulty 8- to 10-fold in the L1 group, these screening

methods increased the odds 5-fold in the ESL group. The likelihood ratio of RECORDS was low for both the ESL and L1 groups (3.2 vs. 4.0, respectively). Thus, positive screening result on RECORDS results in only small changes in the odds of a student being classified as at risk.

**TABLE 6**  
Accuracy Measures for Spontaneously Expressed Concerns (RECORDS) for Total Group and Subgroups at Time 1 and Time 2

Measure	Time 1			Time 2		
	L1	ESL	Total group	L1	ESL	Total group
<i>n</i>	94	181	275	77	158	235
Sensitivity						
%	37.5	15.0	21.4	36.4	15.4	25.0
CI	10.2-74.1	4.0-38.9	9.0-41.5	12.4-68.4	2.7-46.3	10.6-47.1
Specificity						
%	89.5	95.7	93.5	90.9	95.2	93.8
CI	80.6-94.8	90.9-98.1	89.5-96.1	80.6-96.3	89.9-97.9	89.5-96.5
LR	3.6	3.5	3.3	4.0	3.2	4.0
Kappa	.2	.1	.2	.3	.1	.2

Note. L1 = English as a first language. ESL = English as a second language. CI = 95% confidence interval. LR = likelihood ratio. Kappa = Kappa index of agreement.

### *Relationship of Oral Proficiency to Accuracy of Teacher Assessments*

Overall, ANOVAs revealed that there were significant differences among the TNOM groups for the ESL group but not for the L1 group. There were significant differences among the TNOM groups on the EOWPT-R,  $F(3, 154) = 4.73, p < .01$ , PPVT-R,  $F(3, 157) = 7.61, p < .001$ ; and sentence repetition scores,  $F(3, 155) = 6.68, p < .001$ . Specifically, post hoc Sheffé analyses on the ESL group revealed that there were significant differences between the false

**FIGURE 4**

Evaluation Matrices for Spontaneously Expressed Concerns (RECORDS) for Total Group and Subgroups at Time 1 and Time 2

TOTAL GROUP					
RECORDS	CSRS-1		RECORDS	CSRS-2	
	at risk	not at risk		at risk	not at risk
at risk	6	16	at risk	6	13
not at risk	22	231	not at risk	18	198
L1 GROUP					
RECORDS	CSRS-1		RECORDS	CSRS-2	
	at risk	not at risk		at risk	not at risk
at risk	3	9	at risk	4	6
not at risk	5	77	not at risk	7	60
ESL GROUP					
RECORDS	CSRS-1		RECORDS	CSRS-2	
	at risk	not at risk		at risk	not at risk
at risk	3	7	at risk	2	7
not at risk	17	154	not at risk	11	138

L1 = English as a first language. ESL = English as a second language. CSRS-1 = combined standardized reading score at Time 1. CSRS-2 = combined standardized reading score at Time 2.

positive group and the true negative group on sentence repetition ( $M = 12.13$  vs.  $18.23$ ,  $p < .05$ ) and PPVT-R ( $M = 39.83$  vs.  $51.14$ ,  $p < .05$ ) scores, but not on the EOWPT-R scores. These findings indicated that compared to their correctly identified peers, incorrectly identified adequate ESL readers

had significantly lower means on two of the three oral proficiency measures. Similar analyses conducted on the TRATING groups replicated these findings. There were significant differences between the true negative and false positive ESL groups on all three oral proficiency scores: EOWPT-R,  $F(3, 152) = 6.85$ ,  $p < .001$ ; PPVT-R,  $F(3, 155) = 8.38$ ,  $p < .001$ ; and sentence repetition,  $F(3, 153) = 4.49$ ,  $p < .01$ ; and they occurred only for the ESL group. Post hoc tests revealed that there were differences between the true negative and false positive groups on the PPVT-R ( $M = 51.49$  vs.  $38.81$ ,  $p < .05$ ) and EOWPT-R ( $M = 34.14$  vs.  $23.15$ ,  $p < .01$ ).

**TABLE 7**  
Accuracy Measures for Combined Teacher Screening (TNOM/TRATING) for Total Group and Subgroups at Time 2

Measure	Group		
	L1	ESL	Total
<i>n</i>	83	161	244
Sensitivity			
%	84.6	100.0	92.3
CI	53.7–97.3	71.7–100.0	73.4–98.7
Specificity			
%	80.0	66.9	71.1
CI	68.4–88.3	58.6–74.3	64.5–76.9
LR	4.2	3.0	3.2
Kappa	.5	.2	.3

Note. L1 = English as a first language. ESL = English as a second language. CI = 95% confidence interval. LR = likelihood ratio. Kappa = Kappa index of agreement.

**FIGURE 5**  
Evaluation Matrices for Combined Teacher Screening Measures (TNOM/TRATING) for Total Group and Subgroups at Time 2

TOTAL GROUP		
TNOM/TRATING	CSRS-2	
	at risk	not at risk
at risk	24	63
not at risk	2	155
L1 GROUP		
TNOM/TRATING	CSRS-2	
	at risk	not at risk
at risk	11	14
not at risk	2	56
ESL GROUP		
TNOM/TRATING	CSRS-2	
	at risk	not at risk
at risk	13	49
not at risk	0	99

L1 = English as a first language. ESL = English as a second language. CSRS-2 = combined standardized reading score at Time 2.

## Discussion

The purpose of this 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. A third purpose was to determine the influence of oral language factors and cognitive processes on the misclassification (i.e., false positive and false negative) of students. For the total sample of children, teacher specificity, using semistructured interviews and rating scales for screening, was in the moderate to high range, whereas sensitivity was in the low to moderate range. At Time 1, sensitivities of both the teacher nomination (TNOM) and rating scales (TRATING) were higher for the L1 group than for the ESL group, but these differences were reduced at Time 2. Thus, the teachers' sensitivity for identifying ESL students over the long term was comparable to the L1 students regardless of the screening method used. There was no obvious advantage of teacher nominations over teacher ratings of at-risk status in terms of either specificity or sensitivity. Of particular concern, however, was the marked disparity between the accuracy of the structured screening methods and the

spontaneously expressed concerns of teachers, as indicated through scholastic records and rates of referral (RECORDS), at Time 1 and Time 2. The sensitivity of teachers' spontaneously expressed concerns was highly inaccurate and significantly lower than that of the two other teacher screening methods. The strength of teacher assessments was in their specificity, which was at or above 85% at all assessment times, in all groups, and using any screening method. Finally, errors in teacher assessments for ESL and L1 students were explained by different factors. For ESL students only, error in judgment of reading performance was at least partly explained by overreliance on oral language proficiency as an indicator.

The relatively low sensitivity, coupled with a high specificity, 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 Grade 3 children and found a high specificity (92%) but lower sensitivity (71%). In contrast to our study, Salvesen and Undheim used a lower cutoff for their criterion measure (i.e., 1.5 *SD* below the mean), and they studied second- and third-grade first-language speaking students, who are more advanced in their reading decoding and comprehension skills. More stringent cutoff scores and differences in methodology would be expected to increase accuracy (i.e., increase sensitivity), but they did not. The similarity in results suggests that, in fact, students at the beginning of Grade 1 who are at risk for reading failure may show more noticeable difficulties than their average-achieving peers, making them just as noticeable as their older peer group. Together, the studies suggest that teachers may have similar accuracy in screening regardless of the reading level or language group that is being assessed, but their pattern of strengths and weaknesses

may vary depending on the particular attributes of the group.

Perhaps the most important finding of this study is the relatively low sensitivity of the various forms of teacher assessments, a finding that was true for ESL and L1 students alike. The most important quality of a screening measure is its sensitivity, or its ability to identify all students who are at risk. Although a highly sensitive screening measure may result in overreferral, it ensures that all those at risk are referred (or not missed) and, thus, allows for more specific standardized assessment measures to be applied in order to confirm or rule out the existence of a disability. The ideal teacher screening method would have a very high sensitivity without overly compromising specificity. According to the present results, the use of teacher rating scales or teacher nominations alone would result in a failure to identify many potentially at-risk students.

The relative strength of the teacher assessments was their specificity and the moderate to high positive likelihood ratios for all groups across assessment methods. These findings indicate that, when a teacher identifies a child as at risk, there is a high likelihood that he or she indeed has a disability. This certainly has clinical utility, in that all students classified as at risk by teachers are at an increased risk for reading disability compared to those who have not been identified. However, high specificity alone is not sufficient for screening. The cost of this high specificity is that many children thought not to be at risk will indeed have a disability and not be referred. Although still in the moderate range, the likelihood ratios for screening (i.e., TNOM and TRATING) of ESL students were approximately half those of L1 students. There continues to be a need to improve identification measures and techniques for ESL children.

The superior method for teacher screening is one that uses a combination of teacher rating scales and teacher nominations. When both types of screening were combined, sensitiv-

ity in all groups rose to between 85% and 100%. This combined screening method would ensure that most students who are indeed at risk are referred for assessment. Although specificity drops using this method, gold standard testing would allow for correct classification of children initially incorrectly deemed at risk. The use of two assessment methods allows for the collection of convergent data, which likely enhances the reliability of the data. The increased sensitivity could be accounted for by the converging sources of information provided by each assessment tool, one representing the teachers' overall opinion of the severity and prognosis of a difficulty (i.e., TNOM) and the other being the teachers' objective rating of a student's academic performance (i.e., TRATING). Future research operationally defining the key elements of a screening measure would help to clarify this issue further.

Another important finding was the significant difference between the sensitivity of teacher nominations and rating scales and the sensitivity of spontaneously expressed concern for the ESL, L1, and total groups. 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 could be due to several factors. First, it may represent a reporting bias, whereby despite teacher concerns and perhaps even intervention these concerns are not recorded in the ISR files. Second, with respect to their ESL students who may already be receiving some service (i.e., second language support), 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 likely be the case if teachers are less confident in their own ability to distinguish oral language and reading skills among ESL children and, thus, accept some

level of difficulty in this population to be a normal aspect of their development. 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 tool), teachers are not communicating their concerns about some of their students' reading development, particularly their ESL students. This may reflect overcorrection 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 prereferral activities such as monitoring. The findings of this study emphasize the need to elicit teachers' concerns using standardized methods, preferably through the use both of expressed concern obtained in an interview and of teacher rating scales.

When comparing the two language groups, teacher sensitivity was equally poor for ESL and for L1 students (based on the Time 2 data). Intuitively, it would seem that teachers would have less difficulty classifying L1 students with average reading ability, because of their familiarity with reading assessment and development in this group and the complexity of additional language factors that are involved in the assessment of ESL students. Whereas the Time 1 data suggested that this might be the case, the Time 2 results indicated that teachers have equal difficulty identifying ESL and L1 students. It is possible that teachers' overconfidence in their assessment of L1 students is balanced by a tendency to delay referral for their ESL peers, leading to equally high false negative rates (and, as such, lower sensitivity) for both groups. Alternatively, the low sensitivity may be accounted for by different factors in each group.

The results of this study suggest that reliance on oral language skills may be one factor that accounts for the lower accuracy of teachers' screening of ESL students. This is evidenced by the finding that, among children who were not

classified as 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 inappropriately use oral language proficiency as their gauge for the child's overall academic performance, a practice that may result in overidentification of some children. This hypothesis was confirmed through a comparison of the correlations among teacher ratings and performance on the objective measures; objective oral language scores were only moderately associated with children's performance in reading and spelling, whereas teacher-rated oral language ratings were highly correlated to children's ratings of academic performance. Although this finding could be reflective of a rater effect, whereby multiple ratings from the same person could explain the high correlations, teachers used the full scale in their assessments. These results are consistent with previous research findings of a bias or halo effect in teacher ratings of academic performance, whereby teacher ratings of their students' performance in certain academic areas affect ratings in other areas, either in a positive or a negative direction (Salvesen & Undheim, 1994).

There are several limitations of the current study. First, 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 overclassifying of children as at risk for reading disability. Second, follow-up data were available only for 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 ESL and the L1 subgroups on a number of measures. Nevertheless, the consis-

tency in the trends toward a lower sensitivity in ESL and L1 children across the three teacher assessment methods adds validity to the conclusions. Third, the classification of children as ESL was complicated by the fact that a large majority of L1 students were from non-English speaking backgrounds, which raises the possibility that they were also exposed to another language than English. This, however, is an unavoidable state of affairs reflecting universal demographic trends.

The current study has several implications for the practice of school psychology. We recommend the use of screening for the identification of children who may be at risk for reading disability. Although students who are identified as at risk through teacher screening are highly likely to be at risk, using a single method of teacher screening has low sensitivity and is likely to overlook a large number of at-risk children. Screening with a combination of teacher interviews and objective rating scales is the best method of screening, as it has higher sensitivity (allowing most at-risk children to be referred) and still has acceptable specificity. 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 and will likely prevent the implementation of early intervention reading remediation and language programs. 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 may benefit from training on the reading development of ESL and non-ESL students and on the relationship of such factors as oral language, cognitive processes, and academic skills.

#### ABOUT THE AUTHORS

*Marjolaine M. Limbos, MSW, MA, is a doctoral student in school and child clinical psy-*

chology at the Ontario Institute for Studies in Education of the University of Toronto and a psychoeducational consultant in a children's mental health center. She is interested in the cognitive development, early diagnosis, and epidemiology of learning disabilities among ESL children. Esther Geva, PhD, C Psych., is a professor in the Department of Human Development and Applied Psychology at the Ontario Institute for Studies in Education of the University of Toronto and the head of the School and Child Clinical Program. Her current research focuses on the development of reading and language skills in typically developing children and children with learning disabilities; learning disabilities in bilingual/multicultural contexts; early identification of at-risk ESL children; and research design and evaluation. Address: Esther Geva, Department of Human Development and Applied Psychology, OISE, University of Toronto, 252 Bloor St. West, Toronto, Ontario, Canada, M5S 1V6 (e-mail: egeva@oise.utoronto.ca).

#### AUTHORS' NOTES

1. We wish to thank Dr. David Joyce for his input regarding epidemiological concepts and statistical analyses and Dr. Barbara Schuster for her role in coordinating the project. We also extend our thanks to the teachers, students, and staff who participated in the project.
2. This research was supported by a SSHRC (#410-96-0851) and Block Transfer grant from the Ministry of Education to Dr. Esther Geva.

#### 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, DC: Author.
- 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.
- Bowers, P. G., & Wolf, M. (1993). Theoretical links among naming speed, precise timing mechanisms and orthographic skill in dyslexia. *Reading and Writing: An Interdisciplinary Journal*, 5, 69-85.
- Bruck, M., & Genesee, 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 (Ed.), *Cognitive processing in bilinguals* (pp. 283-297). Amsterdam: Elsevier Science.
- Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. Clevedon, England: Multilingual Matters.
- Denckla, M. B., & Rudel, R. G. (1976). Rapid "automatized" naming (RAN): 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.
- Durgunoglu, 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, 36-60.
- 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.
- Ferrolli, 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* [36th Yearbook of the National Reading Conference]. Rochester, NY: National Reading Conference.
- Gardner, M. F. (1990). *Expressive one-word picture vocabulary test-Revised (EOWPVT-R)*. Novato, CA: Academic Therapy.
- 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., & 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., & Siegel, L. S. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading and Writing: An Interdisciplinary Journal*, 12, 1-30.
- 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.
- Geva, E., Yaghoub-Zadeh, Z., & Schuster, B. (in press). Understanding individual differences in word recognition skills of ESL children. *Annals of Dyslexia*.
- Gholamain, M., & Geva, E. (1999). The role of orthography and cognitive factors in the concurrent development of basic reading skills in bilingual Persian-English children. *Language Learning*, 49, 183-217.
- 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 subadjacency 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.
- Naglieri, S. (1989). *Matrix analogies test*. San Antonio, TX: Psychological Corp.
- Petrucci-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. (1992). A primer on the precision and accuracy of the clinical examination. *Journal of the American Medical Association, 267*, 2638-2644.
- 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. (1988). Evidence that IQ scores are irrelevant to the definition and analysis of reading disability. *Canadian Journal of Psychology, 42*, 201-215.
- Siegel, L. (1993). Phonological processing deficits as the basis of a reading disability. *Developmental Review, 13*, 246-257.
- 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. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-406.
- 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. (1992). Speculations on the causes and consequences of individual differences in early reading acquisition. In P. Gough, L. Ehri, & R. Trieman (Eds.), *Reading acquisition* (pp. 307-342). Hillsdale, NJ: Erlbaum.
- Stanovich, K. E. (1999). The sociopsychometrics of learning disabilities. *Journal of Learning Disabilities, 32*, 350-361.
- 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.
- Torgesen, 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., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 85*, 192-212.
- Wilkinson, G. S. (1993). *The wide range achievement test* (3rd ed.). Wilmington, DE: Wide Range, Inc.
- 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.
- Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology, 91*, 415-438.
- Woodcock, R. W. (1987). *Woodcock reading mastery test*. Circle Pines, 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.