Journal of Education and Learning; Vol. 2, No. 4; 2013
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

Teacher Judgment of Reading Achievement: Cross-Sectional and Longitudinal Perspective

Alfred Valdez[1]

[1] College of Education, New Mexico State University, New Mexico, USA

Correspondence: Alfred Valdez, College of Education, Special Education/Communication Disorders Department, MSC 3SPE, New Mexico State University, Las Cruces New Mexico, 88003-8001, USA. Tel: 1-575-646-7607. E-mail: valdez1@nmsu.edu

Received: July 12, 2013 Accepted: August 12, 2013 Online Published: November 18, 2013

doi:10.5539/jel.v2n4p186 URL: http://dx.doi.org/10.5539/jel.v2n4p186

Abstract

Analyses were conducted, using the Early Childhood Longitudinal Study, Kindergarten Class (ECLS-K) database, to compare teachers' judgment of reading skill with direct measures of reading performance for kindergarten, first grade, and third grade students. Teacher judgments of kindergarten students' reading skill significantly predicted first and third grade performance on direct reading measures. In addition, concurrent validity quotients were moderate ranging from r = .58 to .71. These concurrent relationships were further investigated to determine whether socioeconomic status (SES) or teacher experience significantly moderated the concurrent validity estimates. While teacher experience did not significantly moderate the relationship between teacher judgment and direct measures of reading, a small but significant moderation effect was found for SES. That is, SES appeared to differentially bias estimates of reading skill. Judgment bias due to SES level appeared to be greatest when teacher judgments were higher on the rating scale (i.e., proficient) for kindergarten and first grade students. However, for third grade students, judgment bias due to SES was greatest for students lower on the rating scale (i.e., skill not present).

Keywords: teacher judgment, validity, interaction regression

1. Introduction

1.1 Background

Throughout the course of instruction, teachers make judgments of their students' knowledge. Such judgments allow teachers to offer appropriate feedback or to alter instructional goals (McCormick & Pressley, 1997; Ormrod, 2006; Woolfolk, 2004). Teacher judgment of student knowledge offers an efficient, cost effective means for the assessment of academic progress and the early identification of children at risk for later academic failure (Fletcher & Satz, 1984; Kenny & Chekaluk, 1993). However, some question the validity of teacher judgments of students' general academic knowledge. According to Hogue and Coladarci (1989) an implicit assumption exists among researchers and policy makers that teacher judgment of students' academic knowledge is biased. In fact, teachers themselves do not appear to possess a great deal of confidence in their own judgment of students' general academic knowledge (Eggen & Kauchak, 2004; Hoge & Butcher, 1984). In spite of the belief that teacher judgment contains bias, reviewers agree that there is substantial evidence supporting the validity of teacher judgment of general academic skill (Perry and Meisels, 1996; Harlen, 2005). Reviewers describe moderate to large correlations between various teacher-judgment measures and criterion achievement tests.

As it relates to reading instruction, teacher judgment of student performance offers crucial information that guides instructional decisions and identifies students in need of specialized instruction (Goodman & Webb, 2006; Graney, 2008). However, assessing levels of reading ability is not a straightforward process. While we may measure concrete objects directly by their physical attributes (e.g., height in inches, weight in pounds, etc.), reading ability is a construct that must be determined indirectly. Therefore, a perceived judgment of reading ability may be biased. Consequently, teachers' judgment of student reading proficiency may not validly capture the construct of reading and decisions derived from these judgments may be erroneous. For example, teachers may overestimate their students' knowledge and therefore advance students to the next level of reading instruction before they are capable of understanding the material. Teachers may also underestimate student

reading performance and keep students at a given level for too long a time. Both of these consequences may have a negative effect on student engagement, effort, and learning (Bandura, 1986; Carver & Scheier, 1998). Given the significance of teacher-based judgment, it is reasonable to investigate the extent to which one may rely on such judgment as a valid estimate of reading performance.

1.2 Teacher Judgment of Reading Ability

Researchers investigating the validity of teacher judgment have typically gathered criterion-related evidence of measurement validity, with the criterion of choice often being a norm-referenced standardized measure of reading performance. With little exception, teacher judgments of reading performance have moderately correlated with standardized measures of reading. For example, Hopkins, George and Williams (1985) asked teachers to rate first and second grade students' performance on the reading and language arts subtests of the Comprehensive Test of Basic Skills (CTBS, 1985) and then compared teacher ratings with the CTBS measures. They found large and positive correlations between the ratings and the CTBS subtests for reading (r = .74) and language arts (r = .73). Bates and Nettelbeck (2001) compared teacher estimates of students' percentile ranks on two subtests of the Neale Analysis of Reading Ability-Revised (NARA-R, Neale, 1988) with scores from the same measure. Again, findings revealed large and positive correlations between ratings and standardized tests for both reading accuracy (r = .77) and reading comprehension (r = .62). Finally, Beswick, Willms, and Sloat (2005) compared teacher judgments from a standardized rating scale, the Teacher Rating Scale-Literacy (TRS, Flynn, 1997), with test scores from the Wechsler Individual Achievement Test –2nd Edition (WIAT-2, Wechsler, 2002) Word Reading subtest, and found a large and significant correlation (r = .67) between the ratings and the test scores.

Studies investigating the predictive validity of teacher judgment have found moderate correlations between teacher judgment measures and a later measured criterion test. Stevenson, Parker, Wilkinson, Hegion, & Fish (1976) compared teacher ratings in second grade with reading achievement scores in third grade and found a moderate correlation for boys (r = .65) and a somewhat smaller correlation for girls (r = .45). Freeman (1993) explored concurrent (same grade) predictions of achievement based on teacher ratings and found moderate correlations for grades four, five and six (r = .72, .74. & .71 respectively). Hecht & Greenfield (2001) asked teachers to rate first grade children's academic competence and found that teacher ratings for first grade children significantly predicted third grade performance on selected reading measures (Letter Word Identification r = .71, Passage Comprehension r = .70).

While these findings showed a strong relationship between teacher judgments of reading performance and students' later performance on standardized measures, in general, they failed to specifically consider other variables that may moderate (or alter) this relationship. There is one exception. Hinnant, O'Brien, & Ghazarian (2009) investigated whether certain demographic variables (sex, ethnicity, family income, and child social skills) moderate the relationship between teacher expectation and student achievement in mathematics and reading. Teacher expectation was operationalized as the discrepancy between teachers' judgment of students' reading and mathematics achievement and students' actual performance on the mathematics and reading subtests of a norm-referenced achievement test. In terms of reading achievement, the authors found a marginally significant interaction (p = .08) where student's sex and minority status moderated the relationship between first grade teacher expectation and third grade reading achievement. This finding suggested that teacher expectation, or bias, in predicting later reading achievement was strongest for minority males. Unfortunately, no other interactions were shown. This may have been due to the extreme attrition rate in their longitudinal study (1,364 at start to 955 fifth grade students).

In summary, studies investigating teacher judgment of reading proficiency reveal moderate levels of concurrent and predictive validity. However, there is a great deal of variability in the judgment/criterion correlations for individual teachers. Hoge and Coladarci (1989) found that judgment/criterion concurrent correlations varied by as much as 64 points (r = .28 to .92). Coladarci's (1986) predictive accuracy estimates, while demonstrating moderate judgment accuracy on the average, showed a great deal of variation (41 to 98 percent accurate) among individual teachers. In fact, regardless of the manner of judgment applied by teachers (e.g., rating, ranking, etc.), a high level of variability in judgment/criterion correlations was shown. This high level of variability of judgment/criterion correlations among individual teachers suggests that other variables may moderate the judgment/criterion relationship.

1.3 Moderator Variable Teacher Experience

In spite of the interest voiced by researchers that teacher experience may moderate the relationship between teacher judgment and a criterion test (Coladarci, 1986; Freeman, 1993; Hecht & Greenfield, 2002; Hoge &

Butcher, 1984), few studies have investigated this relationship. Those that have investigated this relationship have failed to show conclusive outcomes. For example, Webb, Diana, Luft, Brooks, and Brennan (1997) investigated teachers' ability to judge student comprehension from their non-verbal behaviors (i.e., hesitation, impulsive response, etc.). They hypothesized that more experienced teachers would be better able to notice and accurately interpret students' non-verbal cues than less experienced teachers. They found, however, that while more experienced teachers had greater success at correctly predicting students' comprehension than less experienced teachers, the difference was not significant. Freeman (1993) investigated whether years of teacher experience influenced teachers' accuracy at judging reading performance outcomes. Teachers were asked to estimate the number of correct responses students would achieve on the Gates-MacGintie Reading Test (MacGintie, Kamons, Kowalski, MacGintie, & MacKay, 1980) and these estimates were compared with actual correct test responses. Moderate judgment/criterion correlations where found. The authors then investigated how teacher experience and gender might moderate this relationship. Multiple regression analysis revealed that the combined teacher experience and gender variables significantly predicted the judgment errors ($R^2 = .10$) showing that both variables moderated teacher judgment. Unfortunately the authors failed to report the unique contribution of teacher experience. Mulholland and Berliner (1992) also investigated how teacher experience influenced judgment accuracy. They asked teachers to rank order the expected performance of their students on the reading and mathematics subtests of the Iowa Test of Basic Skills (ITBS; Hoover, Hieronymus, Frisbie,& Dunbar, 1990) and compared these ratings to the actual scores students had achieved on this measure. Years of teaching experience was categorized into five groups: (a) preservice teachers, (b) one to five years experience, (c) six to ten years experience, (d) eleven to fifteen years experience, and (e) more than fifteen years experience. Researchers expected to see teacher predictions improve as years of experience increased. Instead, the relationship between teacher experience and prediction accuracy for the reading measure was not significant ($r = -.02$, $p < .88$). Finally, Martin and Shapiro (2011) investigated the relationship between years of teaching experience and teacher judgment accuracy of kindergarten and first-grade students' scores on the Nonsense Word Fluency (NWF) and Phoneme Segmentation Fluency (PSF) measures of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002). Students' discrepancy scores were calculated by subtracting students' actual scores on the NWF and PWF from the predicted scores that were estimated by their teachers. The mean judgment discrepancies scores were compared across three categories of years of teacher experience (1-10 years, 11-20 years, & 21 – 30 years) using an analysis of variance (ANOVA) for each dependent measure (NWF & PWF). A significant relationship was shown for the NWF measure but no significant difference across categories of years of experience was shown for the PSF measure.

1.4 Moderator Variable Student Socioeconomic Status (SES)

Hart and Risley (1995) described the powerful relationship that exists between the early literacy supports that children receive at home (i.e., parent/child language interactions, books in the home, etc.) and their later literacy success at school. These home literacy supports are less likely to be present in low SES households (Evans, 2004) and teachers sometimes make early judgments of children's literacy potential based on what they perceive as SES factors (Feiler and Webster, 1999). Does SES influence teacher judgment accuracy? Elhowleris (2008) investigated how SES affects teacher referral and recommendation decisions for student placement in the gifted and talented program and found no significant relationship between student SES status and either referral or placement decisions. However, other researchers have demonstrated a relationship between SES and teacher judgment. For example, Alvidrez & Weinstein (1999) found that teachers tended to significantly underestimate the predicted performance of low SES students and to significantly overestimate the predicted performance of high SES students. Beswick, Willms, & Sloat (2005) compared teacher ratings of kindergarten students' literacy skill with a standardized reading test criterion measure. The judgment/criterion correlation was large ($r = .67$) and SES appeared to moderate this relationship. Similar to the findings of Alvidrez and Weinstein, Beswick et al. found that teachers significantly underestimated reading performance for students who possessed low socioeconomic indicators. These errors in judgment are important because teachers tend to hold on to their initial estimates of student potential and they tend to disregard evidence that is contradictory to their initial judgment (Feiler & Webster). Thus initial errors in teacher judgment can have lasting impact.

In summary, studies comparing aggregate teacher judgment of reading skill with a standardized criterion test have revealed moderate correlations. However, the strength of this relationship varies widely among individual teacher judgments suggesting that a moderator variable may influence the association that exists between teacher judgment and a criterion test. While researchers have suggested that teacher experience and student SES may moderate the relationship between teacher judgment and a criterion measure, it remains unclear whether this is actually the case.

Unlike the majority of studies cited, this study utilized longitudinal data, rather than cohort comparisons, to more accurately investigate predictor/criterion relationships in terms of actual growth. In addition, the sample from this study was constructed to model United States population estimates and the statistical analysis used in this study considered the complex nature of this sampling method by applying cohort specific and longitudinal weights to arrive at more accurate standard error estimates. Finally, this study utilized a sample of nearly 7,000 participants in order to investigate the relationship that exists between teacher judgment and standardized reading measures. In addition, this study will investigate how SES and teacher experience might moderate this relationship.

1.5 Purpose of the Study

This research will focus on teacher-based judgment accuracy as it pertains to reading achievement in the early elementary grades (kindergarten through third grade). The goals of this study are the following:

Examine the concurrent validity of teacher ratings of kindergarten, first grade and third grade students' literacy skills by investigating how well they agree with direct reading measures given in the same year.

Determine the predictive validity of kindergarten teacher-based judgments of student literacy skills by investigating how well they predict performance on direct reading measures given at first and third grades.

Identify how SES and teacher experience moderate the concurrent relationship between teachers' judgment of reading skill and students' performance on standardized tests of reading performance.

2. Method

2.1 Data Source

Data were obtained from the National Center for Educational Statistics (NCES) Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) public-use data file (Department of Education, Institute of Educational Sciences, 2004). This data source was created to investigate children's growth and development from kindergarten through third grade and includes direct cognitive measures from children as well as indirect measures from parents and teachers of the children included in this data set. The data set is comprised of a nationally representative sample of kindergarten children attending both public and private school in the fall of 1998. These same children were followed through the spring of 2002 as most had progressed though their third grade year of elementary school. The full ECLS-K base year sample consists of 22,782 children enrolled in 1,277 schools during the 1998-99 school year.

This study is concerned with children who were not previously identified at-risk for academic failure so students who had repeated kindergarten, those previously placed in a special needs program and those who could not be assessed due to non-compliance or limited knowledge of English were excluded. In addition, children who changed schools at any time from kindergarten through third grade were excluded in order to simplify interpretation of teacher-based judgments and standardized objective measures longitudinally. Finally, children who did not have both objective and indirect measures for all four waves of data collection (fall-kindergarten, spring-kindergarten, spring-first grade, and spring-third grade) were excluded leaving 6,924 students for this analysis.

2.2 Measures

2.2.1 Teacher Judgment

The teacher judgment variable consisted of scores from the Language/Literacy portion of the Academic Rating Scale (ARS, see Pollack, Atkins-Burnett, Rock & Weis, 2005; Rock & Pollack, 2002). The ARS consists of three self-administered teacher questionnaires that were given during the four waves of data collection. The questionnaires asked teachers to apply a 5-item rating (not yet, beginning, in-progress, intermediate, proficient) to questions designed to measure three learning domains (language & literacy, mathematics, and general knowledge). This analysis used Part C of the ARS, the Language/Literacy-ARS, to represent teacher judged reading performance. The Language/Literacy-Academic Rating Scale (LL-ARS) was designed to measure a common language-literacy construct with items and performance criteria adjusted so that they pertain to the appropriate grade-level. The kindergarten items pertained to childrens' use of complex sentences, ability to interpret a story read to them, production of rhyming words, ability to predict story outcomes, demonstration of early writing behaviors, understanding of conventions of print, and computer use. First grade items pertained to childrens' skill at contributing to classroom discussion, comprehension of a narrative that was read aloud, skill at reading single words (regular and irregular vowels), skill at reading and comprehending grade level text, reading fluency, narrative composition, understanding of print conventions, and computer use. Third grade items

pertained to students' skill at conveying ideas when speaking, their strategy use for information seeking, their reading fluency, skill at reading and comprehension of grade level text (narrative and expository), skill at written composition, skill at editing and improving written composition, and computer use. Each item offered an example that described behaviors linked to each question. A one-parameter logistic model (see Nunnally & Bernstein, 1994, Rausch Model, pp. 398-404) was used to construct the scale. Scores ranged from 1 to 5 (1 meaning "not yet proficient" to 5 meaning "proficient"). Grade appropriate performance criteria were used to determine proficiency levels for each grade level. Pearson reliability estimates obtained from the ECLS-K data were .91, .94, and .95 for kindergarten, first and third grade respectively (Pollack, Atkins-Burnett, Rock, & Weiss; Rock & Pollack).

2.2.2 Reading Measure

The criterion reading measure consisted of the Item-Response Theory-based (IRT) scale scores of the ECLS-K reading assessment, the Reading IRT scale (see Pollack, Atkins-Burnett, Rock & Weis, 2005; Rock & Pollack, 2002). The Reading IRT scale included measures of basic literacy, vocabulary and reading comprehension. The IRT scale scores were developed to measure longitudinal change rather than to compare individuals with same-aged peers. Reliability estimates (Thetas) were .93, .95, .97 and .94 for the four waves of data collection respectively (Pollack et al.; Rock & Pollack).

2.2.3 Moderator Variables

Moderator variables included socio-economic status (SES) and years of teaching experience. SES was measured using the continuous SES scale. The SES scale was derived from variables that included caregivers' income, education, and labor work force status. Each of the variables were converted to z-scores and then aggregated into a single SES scale with values in the sampling frame that ranged from –4.75 to 2.75. Teaching experience was defined as years of teaching experience that pertained to a given analysis. For example, questions comparing kindergarten teacher-based judgments and later grade level criterion measures would operationalize teaching experience as years of experience teaching kindergarten.

2.3 Analytic Approach

The data used in this study (ECLS-K) was collected using a dual-frame, multi-stage sampling design. Statistical analysis must consider the complexity of this sampling design in order to arrive at correct inferential decisions and to calculate appropriate parameter estimates. Two analytic strategies, design-based or model-based approach, may be utilized when dealing with complex sampling designs (Hahs-Vaughn, 2005; Thomas & Heck, 2001; West, 2008). Given that the majority of questions of interest pertained to student-level variables, a designed-based approach was utilized for all analyses. The ECLS-K data set longitudinal weights appropriate for the four waves of data collection were used to derive accurate population estimates for both cross sectional and longitudinal analyses. The ECLS-K implemented a dual-frame, multistage sample, therefore, the jackknife repeated replication (JRR2) method, using AM Statistical Software 6.02 Beta (Cohen, 2004), was used to estimate standard errors and calculate p values for all analyses. The AM software was specifically designed to consider and adjust estimated standard errors given complex sampling designs. Concurrent and predictive validity estimates were obtained by correlating the Reading IRT scale with the Language/Literacy-ARS for spring kindergarten, fall kindergarten, spring first grade, and spring third grade.

2.4 Moderator Analysis

A moderator variable "affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable" (Baron & Kenny, 1986, p. 1174). To investigate whether SES influences the relationship between the Language/Literacy-ARS score and the Reading IRT scale score, a series of hierarchical multiple regression analyses were conducted using the grade appropriate Reading IRT scale score as the criterion measure, the Language/Literacy-ARS score as the predictor variable, one of two variables (SES or teacher years of experience) as the moderator variable, and a cross product term. The cross product terms were constructed by first centering the predictor and moderator variables and then multiplying their transformed values by one another. Centering the variables was accomplished to reduce multicollinearity and to ease interpretation. Various sources (Aiken & West, 1991; Jaccard, Turrisi, & Wan, 1990; Keith, 2006) describe this analytic approach in depth.

2.5 Hierarchical Regression Analysis

Three waves of data were included in this analysis: spring kindergarten, spring first grade, and spring third grade. The predictor and moderator variables were entered in step one of the analysis and the cross product term was entered in step two. A statistically significant R^2 increase in step two indicates significant moderation of the

relationship between the predictor and the criterion variables. That is, the relationship between teacher judgment and the criterion measure are conditional depending on the values of the moderator variable. A significant interaction was followed by post hoc analyses of simple slopes that further investigated the nature of the moderation.

2.6 Simple Slopes Analysis

Simple slopes analysis investigated the relationship between the Reading IRT scale score and the Language/Literacy-ARS score at high (+ 1 SD), medium (mean), and low (-1 SD) values of the moderator variable. Raw beta coefficients of the regression of Language/Literacy-ARS scores on Reading IRT scale scores at high, medium and low values of SES were calculated using the following formulas:

Simple slope of Language/Literacy-ARS variable = (b1 + (b3 * SES value)

Simple intercept = (b2 * SES value) + constant

Where b1 = regression of Language/Literacy ARS on Reading IRT scale score, b2 = regression of SES on Reading IRT scale score, and b3 = the coefficient for the cross product term (Language/Literacy ARS * SES).

2.7 Standardized Solutions

As suggested by Aiken and West (1991, p. 40) Friedrich's (1982) procedure was used to calculate the standardized solutions for all regression coefficients reported in the moderator analyses. This procedure essentially involves conducting a z-score transformation on the criterion variable and on the uncentered predictor and moderator variables. The cross product term was then constructed by cross multiplying the z-score transformed predictor and z-score transformed moderator variables. Regression analyses were performed using the z-score transformed variables and the unstandardized solution from this analysis becomes the appropriate standardized solution for use with predictor, moderator and cross-product regression coefficients (also see Jaccard, Turrisi, & Wan, 1990, pp. 33-34). These standardized solutions were used to calculate standardized beta coefficients for the simple slope analyses (see Aiken & West, 1991, p. 44). The standardized beta coefficients for the regression of the Reading IRT scale scores on the Language/Literacy-ARS scores at conditional SES values are partial correlations of the two variables controlling for SES, and thus allow one to examine the extent of influence SES has on the relationship between the Language/Literacy-ARS score and the Reading IRT scale score.

3. Results

3.1 Statistical Assumptions

Statistical assumptions for the correlational analyses were evaluated and met by visual inspection of bivariate scatter-plots. Statistical assumptions for linear regression analyses were also evaluated. The dataset contained no missing data and data entries were screened for extreme data values. Extreme data were noted for less than 30 cases. However, Cook's distance calculated for all data indicated none would significantly alter the result of the regression coefficient. Investigation of histograms revealed residual variances as normally distributed. However, investigation of scatter-plots (standardized residuals and predicted values) revealed possible departure of the homogeneity of variances. All other statistical model assumptions were reasonably met. An alpha level of .05 was used to evaluate statistical significance for all statistical tests.

3.2 Descriptive Analysis

Table 1. Means and standard deviations for reading irt scale scores and literature/literature academic rating scale scores, and socio-economic status (ses) for four data collection waves

Measure	Fall Kindergarten	Spring Kindergarten	Spring First Grade	Spring Third Grade
IRT Scale Score	27.81	39.70	71.00	110.71
	(9.85)	(13.00)	(19.22)	(18.45)
LL- ARS Score	2.60	3.49	3.53	3.38
	(.70)	(.75)	(.88)	(.85)

Note. N = 6,924

IRT scale score = Item response theory based scale score for reading

LL-ARS Score = Language/Literacy Academic Rating Scale mean rating

Table 1 shows the means and standard deviations for the Reading IRT scale scores and the Language/Literacy-ARS scores for the four waves of data collection. Mean Reading IRT scale scores showed moderate increase across the four waves of data collection. The fall kindergarten Language/Literacy-ARS scores revealed that children on the average were beginning proficiency (category 2) and in progress (category 3) at the three later grade levels.

3.3 Research Question One: How Well Do Teacher-Based Judgments Correspond with Direct Reading Measures Taken at the Same Point in Time (Concurrent Validity)?

Table 2. Correlations for reading irt scores and language/literacy ars scores for four waves of data collection

	RIRT Fall K	RIRT Spring K	RIRT Spring 1st	RIRT Spring 3rd	LL-ARS Fall K	LL-ARS Spring K	LL-ARS Spring 1st	LL-ARS Spring 3rd
RIRT Fall K	1.00							
RIRT Spring K	.83	1.00						
RIRT Spring 1st	.66	.74	1.00					
RIRT Spring 3rd	.51	.55	.69	1.00				
LL-ARS Fall K	.58	.51	.49	.46	1.00			
LL-ARS Spring K	.57	.62	.57	.49	.63	1.00		
LL-ARS Spring 1st	.51	.58	.71	.61	.47	.58	1.00	
LL-ARS Spring 3rd	.45	.50	.61	.64	.42	.47	.60	1.00

Note. $p < .001$ for all correlations

K = Kindergarten, 1st = First Grade, 3rd = Third Grade

RIRT = Reading Item Response Theory scale score

LL-ARS = Language/Literacy Academic Rating Scale

Table 2 shows correlations among the Reading IRT scale scores and the Language/Literacy-ARS scores for the four waves of data collection. All correlation coefficients were significant, $p < .0001$. Concurrent validity quotients for teacher ratings were moderate for all four waves of data collection ($r = .58$, .62, .71, and .64 for fall kindergarten, spring kindergarten, spring first grade, and spring third grade respectively).

3.4 Question Two: How Well Do Teacher-Based Judgments Correspond with Direct Reading Measures Taken at a Later Grade (Predictive Validity)?

Correlational analyses were conducted to evaluate the relationship between earlier Language/Literacy-ARS (LL-ARS) scores and later Reading IRT (RIRT) scale scores. As shown in Table 2, fall kindergarten LL-ARS scores were significantly and moderately correlated with later RIRT scale scores ($r = .51$, .49, & .46 for spring kindergarten, spring first grade, and spring third grade respectively). Spring kindergarten LL-ARS scores were also significantly and moderately correlated with later RIRT scale scores ($r = .57$ & .49 for spring first grade, and spring third grade respectively).

In addition, Table 2 shows that nearer predictions (i.e., fall kindergarten LL-ARS predicting spring kindergarten RIRT and spring kindergarten LL-ARS predicting first grade RIRT) showed greater association ($r = .51$ & .57 respectively) than far predictions (i.e., fall kindergarten LL-ARS predicting third grade RIRT, $r = .46$). Predictive validity quotients were smaller when fall kindergarten Language/Literacy-ARS scores were used to predict spring first and spring third grade reading IRT scale scores ($r = .52$ and .45 respectively) than when spring kindergarten ratings were used to make the same predictions ($r = .58$ and .50 respectively).

3.5 Question Three: Does SES Significantly Moderate the Concurrent Relationships between Teacher-Based Judgments of Literacy Achievement and Direct Reading Measures for Spring Kindergarten, Spring First Grade, and Spring Third Grade?

A hierarchical multiple regression analysis was conducted to answer question three. In this analysis, two predictor variables, Teacher Judgment (LL-ARS) and Student Socio-economic level (SES) were entered first.

Table 3. Summary of hierarchical regression analysis for teacher rating (tr), socio-economic status (ses), and cross-product (tr x ses) predicting reading irt scores (rirt) for spring kindergarten (n =6924), spring first grade (n= 6629), and spring third-grade students (n = 6333)

Variable	R^2	ΔR^2	B	SE B	β (Z score)
Kindergarten Step 1	.415				
Constant			39.741	.208	
Teacher Rating			9.899	.322	.453***
SES			2.863	.286	.174***
Kindergarten Step 2	.418	.003***			
Constant			39.534	.203	
Teacher Rating			9.997	.325	.588***
SES			2.793	.271	.164***
Cross-Product SES X Teacher Rating			1.277	.34	.057***
First Grade Step 1	.532				
Constant			71.789	.347	
Teacher Rating			14.342	.286	.747***
SES			4.310	.310	.174***
First Grade Step 2	.536	.004***			
Constant			71.463	.350	
Teacher Rating			14.650	.296	.667***
SES			4.230	.307	.171***
Cross-Product SES X Teacher Rating			1.728	.280	.061***
Third Grade Step 1	.460				
Constant			111.631	.372	
Teacher Rating			12.100	.250	.557***
SES			5.705	.422	.239***
Third Grade Step 2	.464	.004***			
Constant			111.995	.378	
Teacher Rating			11..829	.260	.544***
SES			5.863	.413	.246***
Cross-Product SES X Teacher Rating			-1.773	.344	-.063***

As Table 3 shows, LL-ARS and SES significantly predicted variation on the criterion measure, Reading IRT scores (RIRT), for all three grade levels: a.) Spring Kindergarten, $F (2, 6921) = 2410.62$, $p < .001$, $R2 = .42$; b.) Spring First Grade, $F (2, 6626) = 3676.54$, $p < .001$, $R2 = .53$; and c.) Spring Third Grade, $F (2, 6330) = 2634.55$, $p < .001$, $R2 = .41$. The analysis indicated that the combined variables, LL-ARS and SES, predicted nearly half of the variation of the criterion reading measure (RIRT).

However, to determine the conditional relationship (or moderation) of the two predictor variables (LL-ARS and SES) on the criterion measure (RIRT) for the three grade levels, a third variable, the cross product of LL-ARS and SES, was entered at the second level of the analysis for each of the three grade levels. The analysis revealed a small but significant interaction for all three grade levels: a.) Spring Kindergarten, $F (3, 6920) = 1628.21$, $p < .001$, R2 change = .003; b.) Spring First Grade, $F (3, 6625) = 2488.38$, $p < .001$, R2 change = .004; and c.) Spring Third Grade, $F (3, 6329) = 1784.22$, $p < .001$, R2 change = .004. This finding suggested that Teacher Rating conditionally predicts the criterion reading measure (RIRT) depending on the student's SES level. Post hoc analyses of simple slopes were conducted to examine the nature of this interaction.

3.6 Question Three Follow-up (Simple Slopes Post-Hoc Analysis): How Does SES Significantly Moderate the Concurrent Relationships between Teacher-Based Judgments of Literacy Achievement and Direct Reading Measures for Spring Kindergarten, Spring First Grade, and Spring Third Grade?

Table 4. Simple slopes regressing reading irt scale score on the language/literacy-ars score at selected ses levels for spring kindergarten (n = 6924), spring first grade (n = 6629), and spring third grade (n = 6333)

Kindergarten	Intercept	b (slope)	SE b	β	t
High SES	41.63	10.95	.42	.63	26.33***
Medium SES	39.53	10.00	.33	.59	30.72***
Low SES	37.44	9.04	.41	.55	22.00***
First Grade					
High SES	75.23	16.19	.45	.71	36.36***
Medium SES	71.46	14.65	.30	.67	49.43***
Low SES	67.70	13.11	.32	.62	41.14***
Third Grade					
High SES	116.98	10.32	.44	.50	23.58***
Medium SES	112.00	11.83	.26	.54	45.47***
Low SES	107.01	13.34	.34	.59	39.38***

Table 4 shows the intercepts and slopes for spring kindergarten, spring first and spring third grades at three SES levels; high (+ 1 SD), medium (mean), and low (-1 SD). Of particular interest in this analysis are the calculated slopes. The slopes indicate the predicted criterion score (RIRT) given a unit measure of teacher judgment (LL-ARS) at three discrete levels of SES. For example, Table 4 shows spring kindergarten students at medium SES level with an intercept of 39.53 and a slope of 10.00. Given these values, the regression model-predicted mean RIRT score for students at the mean for SES and LL-ARS values is the intercept, 39.53. This value roughly corresponds to the sample mean of 30.70 shown in Table 1. The slope, 10.00, predicts a ten point increase in RIRT score for every unit increase in LL-ARS score for students at typical SES. However, as shown by the slopes for kindergarten students, the unit increase in RIRT score is greater for high SES and less for low SES values. This is the nature of the interaction of SES and LL-ARS when predicting the criterion measure (RIRT). More simply, equal slopes across the three SES categories would suggest no interaction of SES and LL-ARS when predicting RIRT scores and unequal slopes across the SES categories, as shown for the three grade levels, suggests that teacher's prediction of student reading level (as measured by the RIRT criterion measure) is influenced by the student's SES.

Table 4 shows that slopes increase as SES increases for spring kindergarten and spring first grade. However, for spring third grade, the slopes decrease as SES increases. This pattern of results is shown in Figures 1 and 2 where the interaction of SES and LL-ARS shows a wider discrepancy of predicted RIRT scores at each SES level for higher LL-ARS judgments (i.e., judged as "proficient") for spring kindergarten and spring first grade students. This pattern is reversed for spring third grade students (see Figure 3) where the interaction of SES and LL-ARS shows a wider discrepancy of predicted RIRT scores at each SES level for lower LL-ARS judgments (i.e., judged as "not yet evidenced").

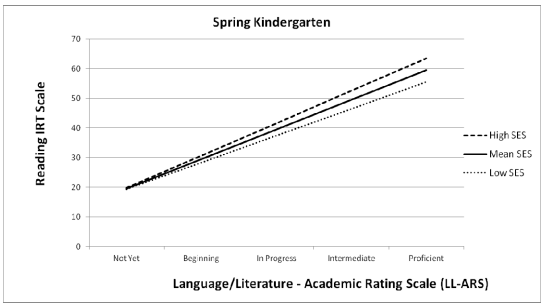

Figure 1. Simple Slopes of Reading Levels (RIRT) regressed on Teacher Judgments of Student Reading (LL-ARS) at three socio-economic levels for spring kindergarten

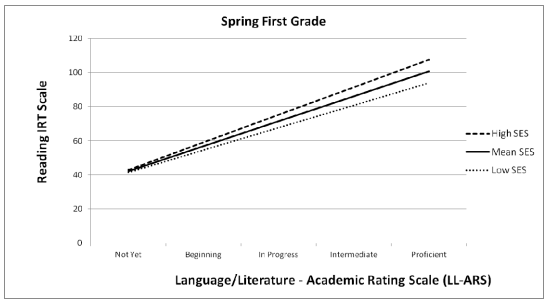

Figure 2. Simple Slopes of Reading Levels (RIRT) regressed on Teacher Judgments of Student Reading (LL-ARS) at three socio-economic levels for spring first grade

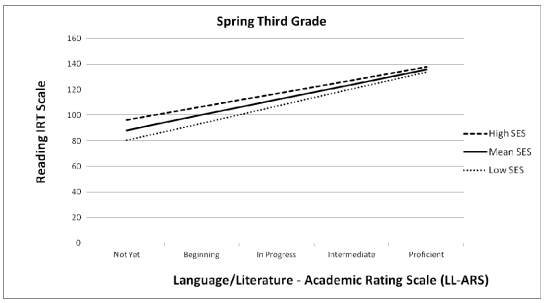

Figure 3. Simple Slopes of Reading Levels (RIRT) regressed on Teacher Judgments of Student Reading (LL-ARS) at three socio-economic levels for spring third grade

3.7 Question Four: Does Teacher Experience Significantly Moderate the Concurrent Relationships between Teacher-Based Judgments of Literacy Achievement and Direct Reading Measures?

Hierarchical regression analyses showed that the relationship between teacher judgment (Language/Literacy-ARS) and the criterion measure (Reading IRT scale) was not significantly moderated by teacher experience for spring kindergarten, $\Delta R^2 = .0000362$, $F(3,6595) = .388$, $p = .761$, spring first grade, $\Delta R^2 = .0001353$, $F(3,6794) = 1.87$, $p = .133$, or spring third grade, $\Delta R^2 = .0001396$, $F(3,6805) = 1.62$, $p = .182$.

4. Discussion

Teacher judgments offer an unobtrusive estimate of student levels of performance and provide information regarding both the product and processes involved in student learning. This study found that teacher judgment of the relative proficiency of students' language and literacy skills were moderately associated with direct reading measures at first, second, and third grades. In addition, teacher judgment of kindergarten students' language and literacy skills were moderately predictive of reading performance in later grades (first and third grades). This

finding is generally consistent with Hoge & Coladarci's (1989) review of 16 studies that showed moderate to large correlations between teacher judgments and standardized reading tests (median r = .66).

While correlational studies have shown a moderate relationship between teacher judgments of student progress in language/literacy and criterion measures of language/literacy progress, less is known about how this relationship might be influenced by other variables such as teacher years of experience or student socio-economic level (SES). Therefore, this study also investigated the moderating influence of SES and teacher years of experience on the relationship between teacher judgment of students' literacy level and students' performance on a criterion measure of literacy.

Moderator analysis showed that teacher experience did not significantly moderate the concurrent relationship between teachers' judgments of language/literacy skills (LL-ARS) and the direct measure of reading (RIRT). This finding may be explained by the Nye, Konstantopoulos, and Hedges (2004) study that pointed out that some teacher characteristics (such as years of teaching) are confounded by other variables such as within-school teacher assignment (i.e., experienced teachers within schools may be placed with more or less advanced students) or across school teacher placement (i.e., less experienced teachers may be placed at lower performing schools), making clear interpretation of this variable difficult without considering and controlling for confounding influences. In addition, years of experience simply may not accurately reflect teachers' ability to engage in reliable judgment. Other teacher background variables such as knowledge of measurement or skill at accurately observing student behavior may help to explain the concurrent relationship between teacher literacy judgments and direct reading assessment. Future studies should continue to investigate teacher characteristics as moderator variables that may influence the relationship between teacher judgments of student performance and actual student performance.

Socio-economic status (SES) has been suggested as an intervening variable that could potentially bias teacher judgments of student performance. For example, using IQ scores as a criterion measure, Alvidrez and Weinstein (1999) found that teacher ratings for children from high SES overestimated later IQ scores while ratings for low SES children underestimated later IQ scores. Unlike the Alvidrez and Weinstein study, this investigation focused on language and literacy rather than cognitive skill as a criterion measure. This investigation found a small but significant moderation effect for SES on the relationship between teacher judgment of language/literacy and the direct measurement of student reading skill using a standardized reading criterion test. However, this moderation effect differed depending on grade level. It appears that for kindergarten and first grade students, the strength of relationship between teachers' judgments of language/literacy skills (LL-ARS) and the direct measure of reading (RIRT) increased as SES increased. In addition, for kindergarten and first-grade students, student SES had the greatest influence on the predicted RIRT score when teacher ratings were at the proficient level. However, for third-grade students, the strength of relationship between teachers' judgments of language/literacy skills (LL-ARS) and the direct measure of reading (RIRT) decreased as SES increased and student SES level had the greatest influence on the predicted RIRT score when teacher ratings were at the not yet evidenced level.

Practically speaking, teachers' judgments of low proficiency kindergarten and first grade students do not appear to be influenced by SES. However, teachers' judgments of reading performance for high proficiency students were more highly influenced by student SES. Martin and Shapiro's (2011) analysis of kindergarten and first-grade teacher judgments of reading performance partially supports this finding. They found that teachers were more accurate at judging performance for low achieving students and less accurate at judging performance for higher achieving students. It may be, as Martin and Shapiro suggest, that current educational policy has encouraged primary grade (kindergarten & first-grade) teachers to be more sensitive to the needs of low achieving readers and therefore more accurate at judging the performance of low achieving students.

However, this moderation effect of SES reversed in third grade in which the strength of relationship between teachers' judgments of language/literacy skills (LL-ARS) and the direct measure of reading (RIRT) decreased as SES increased. In addition, student SES had the greatest influence on the predicted RIRT score when teacher ratings were at the lowest level (i.e., skill not yet evidenced). In practice this finding suggests that third grade teachers who judge a student as a low proficiency reader will be highly influenced by SES but their judgments will be less influenced by SES for students whom they judge as high proficiency readers. A number of studies (Bates & Nettelbeck, 2001; Begeny et al., 2008; Coladarci, 1986; Feinberg & Shapiro, 2009) have shown that teachers were more accurate at judging performance for high achieving students and less accurate at judging performance for lower achieving students. Coladarci explained this result by suggesting that teachers are more inclined to judge their high achieving students as proficient. Given that high achieving students generally perform well on tests there would be little discrepancy between teachers' judgment of student literacy

performance and the same students' actual performance on a direct literacy measure regardless of moderating variables such as SES.

While the explanations offered by Martin and Shapiro (2011) and Coladarci (1986) both seem plausible, neither explanation accounts for the opposing findings. Nor does either explanation make clear the reversal of the SES moderation effect found in this study. What might explain these opposing findings is the fact that participants in the Martin and Shapiro study and the Coladarci study were at very disparate levels of reading skill. The Martin and Shapiro participants were kindergarten and first grade students and the participants in Coladarci's study were third and fifth-grade students. Student skill level in reading is quite different for kindergarten/first grade as opposed to third/fifth grade. Chall (1996) describes kindergarten and first-grade students as demonstrating reading skills at the very beginning stages of reading acquisition and describes third grade students and beyond as moving from reading acquisition to the use of reading to acquire information, or reading to learn. Considering the skill level/stages of reading acquisition, the explanation offered by Martin and Shapiro does seem plausible for kindergarten and first-grade students. That is, teachers may be less prone to judgment error for low proficiency students due to an emphasis in educational policy to identify low proficiency readers at the beginning stages of reading acquisition. This being the case, moderator variables, such as SES, might influence teacher judgment for higher, rather than lower, performing students. Third grade students, on the other hand, have acquired greater skill in reading (i.e., greater reading fluency, consolidation of decoding skills, larger sight vocabulary, etc., Chall, 1996). Students skilled in reading are likely to be skilled in other academic areas. Consistent with Coladarci's explanation, for later elementary grade students, high achieving students generally perform well on tests and therefore test scores and teachers' judgments of performance should correspond well.

In summary, the findings from this study support the validity of teacher judgment of students' literacy skills. In addition, this study found little evidence for the moderation of teacher experience but suggests how SES might bias teacher judgments of students' literacy level differently for early (kindergarten/first-grade) and later (third grade) elementary levels of instruction. However, conclusions were derived from a secondary analysis of this data set and therefore causal interpretations are inappropriate.

Acknowledgements

The U. S. Department of Education, National Center for Educational Statistics (NCES) made the ECLS-K data available and offered training and technical support pertaining to the analysis of the ECLS-K data. The author is grateful for the training and support offered by the NCES. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NCES.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91,* 731-746. http://dx.doi.org/10.1037/0022-0663.91.4.731

Bandura, A. (1986). *Social foundations of thought and action.* Englewood Cliffs, NJ: Prentice Hall. http://dx.doi.org/10.4135/9781446221129.n6

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173-1182. http://dx.doi.org/10.1037/0022-3514.51.6.1173

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgment of reading achievement. *Educational Psychology, 21*(2), 177-187. http://dx.doi.org/10.1080/01443410020043878

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. *School Psychology Quarterly, 23,* 43-55. http://dx.doi.org/10.1037/1045-3830.23.1.43

Beswick, J. F., Willms, J. D., & Sloat, E. A. (2005). A comparative study of teacher ratings of emergent literacy skills and student performance on a standardized measure. *Education, 126*(1), 116-137.

Carver, C. S., & Scheier, M. F. (1998). *On the self-regulation of behavior.* New York: Cambridge University Press.

Chall, J. S. (1996). *Stages of reading development* (2nd ed.). New York: McGraw-Hill.

Cohen, J. (2004). AM Statistical Software (Version 6.02 Beta). [Computer software and manual.] Retrieved July 15, 2005, from http://am.air.org

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78*(2), 141-146. http://dx.doi.org/10.1037/0022-0663.78.2.141

CTBS/McGraw-Hill. (1985). *Comprehensive test of basic skills.* Monterey, CA: CTB/McGraw-Hill.

Department of Education, Institute of Educational Sciences. (2004). ECLS-K Longitudinal Kindergarten-Third Grade Public-Use Data File, [Data file]. Retrieved from http://nces.ed.gov

Eggen, P., & Kauchak, D. (2004). *Educational psychology: Windows on the classroom.* Upper Saddle River, NJ: Pearson.

Elhowleris, H. (Spring 2008). Teacher judgment in identifying gifted/talented students. *Multicultural Education, 15*(3), 25-38.

Evans, G. W. (2004). The environment of poverty. *American Psychologist, 59*(2), 77-92. http://dx.doi.org/10.1037/0003-066X.59.2.77

Feiler, A., & Webster, A. (1999). Teacher predictions of young children's literacy success or failure. *Assessment in Education, 6*(3), 341-356. http://dx.doi.org/10.1080/09695949992784

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher Accuracy: An Examination of teacher-based judgments of students' reading with differing achievement levels. *Journal of Educational Research, 102*(6), 453-462. http://dx.doi.org/10.3200/JOER.102.6.453-462

Fletcher, J. M., & Satz, P. (1984). Test-based versus teacher-based predictions of academic achievement: A three-year longitudinal follow-up. *Journal of Pediatric Psychology, 9*(2), 193-203. http://dx.doi.org/10.1093/jpepsy/9.2.193

Flynn, J. (1997). *Teacher Rating Scale.* LaCrosse, WI: LaCrosse Area Dyslexia Research Institute.

Freeman, J. G. (1993). Two factors contributing to elementary school teachers' predictions of students' scores on the Gates-MacGintie Reading Test. *Level D. Perceptual and Motor Skills, 76,* 563-538. http://dx.doi.org/10.2466/pms.1993.76.2.536

Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science, 26*(4), 797-833. http://dx.doi.org/10.2307/2110973

Good, R. H., III, & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Goodman, G., & Webb, M. A. (2006). Reading disability referrals: Teacher bias and other factors that impact response to intervention. *Learning Disabilities: A Contemporary Journal, 4*(2), 59-70.

Graney, S. B. (2008). General education teacher judgments of their low-performing students' short-term reading progress. *Psychology in the Schools, 25*(6). 537-549. http://dx.doi.org/10.1002/pits.20322

Harlen, W. (2005). Trusting teachers' judgment: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245-270. http://dx.doi.org/10.1080/02671520500193744

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experiences of young American children.* Baltimore: Paul H. Brooks Publishing.

Hecht, S. A., & Greenfield, D. B. (2001). Comparing the predictive validity of first grade teacher ratings and reading-related tests on third grade levels of reading skills in young children exposed to poverty. *School Psychology Review, 30,* 50-69.

Hecht, S. A., & Greenfield, D. B. (2002). Explaining the predictive accuracy of teacher judgments of their students' reading achievement: The role of gender, classroom behavior, and emergent literacy skills in a longitudinal sample of children exposed to poverty. *Reading and Writing: An Interdisciplinary Journal, 15,* 789-809.

Hahs-Vaughn, D. L. (2005). A primer for using and understanding weights with national datasets. *Journal of Experimental Education, 73*(3), 221-248. http://dx.doi.org/10.3200/JEXE.73.3.221-248

Hinnant, J., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology, 101*(3), 662-670. http://dx.doi.org/10.1037/a0014306

Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology, 76,* 777-781. http://dx.doi.org/10.1037//0022-0663.76.5.777

Hoge, R. D., & Coladarci (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research, 59,* 297-313. http://dx.doi.org/10.3102/00346543059003297

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1990). *Iowa Test of Basic Skills (ITBS).* Rolling Meadows, IL: Riverside Publishing Co.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*(3), 177-182. http://dx.doi.org/10.1111/j.1745-3984.1985.tb01056.x

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression.* Newbury Park, CA: Sage.

Keith, T. Z. (2006). *Multiple regression and beyond.* Boston, Allyn & Bacon.

Kenny, D. T., & Chekaluk, E. (1993). Early reading performance: A comparison of teacher-based and test-based assessment. *Journal of Learning Disabilities, 26,* 227-236. http://dx.doi.org/10.1177/002221949302600403

MacGintie, W. H., Kamons, J., Kowalski, R. L., MacGintie, R. K., & MacKay, T. (1980). *Gates-MacGintie Reading Test.* Canadian Edition: Teacher's manual. Toronto, ON: Nelson Canada.

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

McCormick, C. B., & Pressley, M. (1997). *Educational psychology: Learning, Instruction, and Assessment.* New York: Longman.

Mulholland, L. A., & Berliner, D. C. (1992). Teacher experience and the estimation of student achievement. Paper presented at the American Educational Research Association, San Francisco, CA.

Neale, M. D. (1988). *Neale Analysis of Reading Ability-Revised.* Melbourne, ACER.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation Policy and Policy Analysis, 26,* 237-257. http://dx.doi.org/10.3102/01623737026003237

Ormrod, J. E. (2006). *Essentials of educational psychology.* Upper Saddle River, NJ: Pearson.

Perry, N. E., & Meisels, S. J. (1996). *How Accurate are Teacher Judgments of Students' Academic Performance?* U.S. Department of Education, Washington, DC: National Center for Educational Statistics.

Pollack, J., Atkins-Burnett, S., Rock, D., & Weiss, M. (2005). *Early childhood longitudinal study - Kindergarten Class of 1998–99(ECLS-K): Psychometric report for the third grade (NCES 2005–062).* Washington, DC: National Center for Educational Statistics.

Rock, D. A., & Pollack, J. M. (2002). *Early childhood longitudinal study - Kindergarten Class of 1998–99(ECLS-K): Psychometric report for kindergarten through first grade (NCES 2002–005).* Washington, DC: National Center for Educational Statistics.

Stevenson, H. W., Parker, T., Wilkinson, A., Hegion, A., & Fish, E. (1976). Predictive value of teachers' ratings of young children. *Journal of Educational Psychology, 68,* 507-517. http://dx.doi.org/10.1037//0022-0663.68.5.507

Thomas, S. L., & Heck, R. H. (2001). Analysis of Large-Scale Secondary Data in Higher Education Research: Potential Perils Associated with Complex Sampling Designs. *Research In Higher Education, 42*(5), 517-540.

Webb, J. M., Diana, E. M., Luft, P., Brooks, E. W., & Brennan, E. L. (1997). Influence of pedagogical expertise and feedback on assessing student comprehension from nonverbal behavior. *The Journal of Educational Research, 91*(2), 89-97. http://dx.doi.org/10.1080/00220679709597526

Wechsler, D. (2002). *Wechsler Individual Achievement Test* (2nd ed.) (Available from The Psychological Corporation, San Antonio, TX.).

West, B. T. (2008). Statistical and methodological issues in the analysis of complex sample survey data: Practical guidance for trauma researchers. *Journal of Traumatic Stress, 21*(5), 440-447. http://dx.doi.org/10.1002/jts.20356

Woolfolk, A. (2004). *Educational psychology* (9th ed.). Upper Saddle River, NJ: Pearson.