International Electronic Journal of Elementary Education, 2013, 5(2), 125-142.

Using First-Grade Teacher Ratings to Predict Third-Grade English Language Arts and Mathematics Achievement on a High-Stakes Statewide Assessment

Dorinda J. GALLANT *

The Ohio State University, Columbus, OH, USA

Received: 2 August 2012 / Revised: 25 December 2012 / Accepted: 10 January 2013

Abstract

Early childhood professional organizations support teachers as the best assessors of students' academic, social, emotional, and physical development. This study investigates the predictive nature of teacher ratings of first-grade students' performance on a standards-based curriculum-embedded performance assessment within the context of a state accountability system. The sample includes 4292 elementary school students cross-classified by 131 first-grade and 137 third-grade schools attended. This study uses extant statewide assessment data for students located in a state in the southeastern part of the United States. Controlling for student and school demographic variables in cross-classified random effects multilevel models, first-grade teacher ratings—as reflected by domain scores on a performance assessment—are found to positively and significantly correlate with students' third-grade academic achievement.

Keywords: Teacher ratings, Predictive validity, Curriculum-embedded Performance assessment

Introduction

Research on the use of teacher-based judgment measures (e.g., measures that use teacher ratings or rankings to assess students' knowledge or skills in specific academic content areas) to assess students' academic achievement in core academic areas (i.e., math, reading, science, or social studies) span over four decades. One of the most comprehensive scholarly reviews of the use of teacher-based judgment measures to assess student achievement was completed by Hoge and Coladarci (1989). The authors presented a thorough review of 16 empirical studies from 1962-1988. With these studies, they examined the association between concurrently administered direct and indirect teacher-based judgment measures in which

* Dorinda Gallant, *The Ohio State University*, Department of Educational Studies, College of Education and Human Ecology, 211B Ramseyer Hall, 29 West Woodruff Avenue, Columbus, OH 43210, United States, telephone: 614-247-8860. E-mail: gallant.32@osu.edu

ISSN:1307-9298

www.iejee.com

teachers used ratings, rankings, grade equivalence, number correct, and item responses in reading, mathematics, social studies, and science, and norm-referenced measures of academic achievement. Overall, Hoge and Coladarci found a moderate to strong association (Mdn *r* = .66) between teachers' judgments of students' academic performance and their actual performance on standardized norm-referenced achievement tests.

Specific to the predictive value of teachers' ratings, several studies have investigated the longitudinal nature of teacher ratings in relation to students' performance on norm-referenced tests (e.g., Hecht & Greenfield, 2001; Meisels, Liaw, Dorfman, & Nelson, 1995; Quay & Steele, 1998; Stevenson, Parker, Wilkinson, Hegion, & Fish, 1976). Stevenson et al. (1976), followed a cohort of students from kindergarten through third grade and found moderate to strong correlations (*r*s ranged from .41 to .71) between teacher ratings (i.e., instructions, vocabulary, reflective, retention, learning, independence, or attention) and students' reading achievement, as measured by the Wide Range Achievement Test (Jastak, Bijou, & Jastak, 1965), across four time periods (i.e., prior to kindergarten and in the spring of grades 1, 2, and 3), and moderate to strong correlations (*r*s ranged from .37 to .65) between teacher ratings (i.e., instructions, learning, vocabulary, retention, hardworking, independence, or reflective) and students' arithmetic achievement across the same four time periods.

Using a sample of kindergartners in three Michigan school districts to present predictive validity evidence for a performance assessment for young children, Meisels, Liaw, Dorfman, and Nelson (1995) reported high correlations between teacher ratings and students' performance on a norm-referenced measure within a one-year period. Specifically, the correlations between kindergarteners' total scores on the Work Sampling System (WSS; Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 1994) checklists—an authentic assessment measure in which students are observed in their natural settings to determine the extent to which they can demonstrate proficiency on defined local, state, or national curriculum standards—and a total score on the Kindergarten Achievement Battery of the Woodcock-Johnson Psychoeducational Battery–Revised (WJ-R; Woodcock & Johnson, 1989), were .76. The fall/spring and winter/spring correlations between individual WSS checklists and WJ-R total scores were similar (ranging from .38 to .78 for fall/spring and ranging from .39 to .78 for winter/spring). The highest correlations were found between the concept and number and the language and literacy subscales of the WSS and the WJ-R (*r*s = .75 and .78, respectively, for fall/spring and *r*s = .78 for winter/spring).

In 2001, Hecht and Greenfield found that first-grade teacher ratings, collected using the academic competence subscale of the Social Skills Rating System (Gresham & Elliott, 1990) in a large urban public school system, explained approximately 50% of the variance in third-grade letter-word identification and passage comprehension for a predominantly minority population of children (i.e., 73% African American, 18% Hispanic, and 2% Asian American) exposed to poverty. The authors also reported that first-grade teacher ratings accurately classified third-grade students into good or impaired reader ability groups with an accuracy rate of at least 73%. Hence, the finding provided support for the use of teacher ratings to predict students' later performance.

However, in Quay and Steele's 1998 study in which the authors followed a cohort of students from pre-kindergarten to grade 2, the magnitude of the association between pre-kindergarten teacher ratings and students' later achievement varied according to when the ratings occurred. The researchers found that pre-kindergarten ratings on the Developmental Profile II (DPII; Alpern, Boll, & Shearer, 1986) had only a small significant association with first-grade teacher reports of reading comprehension (*r* = .16), but pre-kindergarten ratings on the

Developmental Rating Scale (DRS) had significant moderate associations with first-grade teacher reports of reading comprehension (r = .42) and math achievement (r = .38) and significant small associations with second-grade teacher reports of reading comprehension (r = .24) and math achievement (r = .27). Yet, the associations between pre-kindergarten ratings on either the DRS or the DPII and students' second-grade Iowa Tests of Basic Skills (ITBS) total reading or math were small (rs ranged from -.02 to .12). Moreover, the first-grade ratings on the DPII were small significant predicators of first-grade teacher reports of reading comprehension and math achievement and second-grade teacher report of reading comprehension (rs ranged from .18 to .29). Conversely, first-grade ratings on the DRS were moderate to high predictors of first- and second-grade teacher reports of reading comprehension and math achievement, and second-grade ITBS total reading and math (rs ranged from .40 to .78)

Significance of This Study

Although much research has been conducted on the use of teacher ratings, this study aims to contribute to research on teacher ratings specifically in two areas: (a) examining the predictive nature of teacher ratings within the context of a state's education accountability system and (b) using multilevel methodology to account for both the nested nature of students within schools and the cross-classification that can occur in longitudinal studies when students do not attend the same schools at both time periods of investigation. There is minimal research literature on the connection between teacher ratings on a continuously administered performance-based assessment and a high-stakes group-administered assessment within the context of an education accountability system. Meisels, Atkins-Burnett, Xue, Nicholson, Bickel, and Son (2003) proposed that instructional and high-stakes assessments can potentially be connected to create an education accountability system that relies on both information obtained in the classroom and tests on student achievement. This team of researchers found that in a large urban public school system in which at least 70% of the sample were African American and at least 87% received free or reduced-price lunch, students who had been exposed to the WSS (Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 1994, 2001) for at least three years, prior to being administered the ITBS, showed a greater change in mean grade 3 and grade 4 ITBS Developmental Standard Scores for reading and mathematics compared to a demographically matched comparison group and all other students within the school district. These findings suggest that, as teachers obtained continuous information on students' academic performance, they could potentially use the information to improve students' learning over time. This is especially true when viewing accountability as a "system" instead of as a stand-alone test. Hence, as school districts and state departments of education attempt to meet national accountability mandates for students in grades 3 through 8 that are found in the No Child Left Behind legislation of 2001, it is imperative that researchers investigate the use of teacher ratings in the early grades to predict students' later performance on high-stakes tests.

Furthermore, from a methodological perspective, this study explores the predictive nature of teacher ratings from a multilevel modeling perspective, accounting for the variance associated with the cross-classification of students within schools. Previous studies have primarily used correlations, linear regression analyses, or correlations and regression analyses (e.g., Hecht & Greenfield, 2001; Meisels et al., 1995; Stevenson et al., 1976) to investigate the predictive nature of teacher ratings. However, the nested nature of students within schools or classrooms has either been ignored or limited in teacher-rating studies. Gallant (2005, 2009) explored the predictive nature of first-grade teacher ratings on a state's standards-based performance assessment for children in a large urban public school district using a two-level

multilevel regression model. The author found a positive significant association between first-grade teacher ratings and students' third-grade performance in mathematics and English language arts when partitioning the between-classroom and the within-classroom variances. However, the study did not account for situations in which students may have attended different schools at the two time points (i.e., grades 1 and 3) because of school configurations (e.g., K-2, K-3, K-5) and because the sample was limited to one urban school district within the state.

Purpose of this Study

Therefore, this study aimed to expand the work of Meisels et al. (1995) and Gallant (2005, 2009) on the predictive nature of teacher ratings in the early grades to determine students' later academic performance by including a larger representative sample of students nested within schools. Furthermore, this study accounts for the cross-classification of students within schools. Hence, the purpose of this study is to use a statewide, standards-based assessment program for students in grades 1 and 3 to investigate the predictive nature of first-grade teacher ratings, taking into consideration the nested nature of the data and the cross-classification of students across schools. Specifically, this study investigated the association between first-grade teacher ratings, as reflected by domain scores, on the language and literacy and mathematical thinking domains of a standards-based curriculum-embedded performance assessment based on the developmental guidelines and checklist of Work Sampling System and third-grade students' achievement in mathematics and English language arts, as indicated by scale scores, on a high-stakes standards-based criterion-referenced test. The research questions of interest were the following:

1. How well do first-grade teachers' ratings of students' language and literacy performance on a standards-based curriculum-embedded performance assessment predict students' English language arts third-grade performance on a standards-based criterion-referenced test?

2. How well do first-grade teachers' ratings of students' mathematical thinking performance on a standards-based curriculum-embedded performance assessment predict students' third-grade mathematics performance on a standards-based criterion-referenced test?

Context of this Study

In a state located in the southeastern part of the United States, an education accountability act was passed in 1998. The accountability act introduced comprehensive accountability measures to schools and school districts and further created an education oversight committee to monitor compliance with the legislation (Education Accountability Act, 1998). Among the accountability components included in the section of the legislation on academic standards and assessments, three major components relevant to this study were (a) adoption of grade-level specific educational standards in English language arts, mathematics, science, and social studies for kindergarten through grade 12; (b) development or adoption of a statewide assessment program to measure student performance on state standards in English language arts, mathematics, science, and social studies for grades 3 through 8; and (c) development, adoption, or selection of separate tests linked to adopted first- and second-grade academic standards. While student performance in grades 3 through 12 were used as an accountability measure at the state level, first- and second-grade readiness tests were not.

However, the second-grade readiness test was to serve as the baseline for the third-grade assessment.

Legislative compliance for the first-grade readiness test was met when the state board of education, through the state's department of education, adopted the personal and social development, language and literacy, and mathematical thinking developmental guidelines and checklists of the Work Sampling System for students in first grade, with modifications to indicators for alignment with the state's adopted standards for first grade (Huynh, Prior, & Gallant-Taylor, 2002). The Work Sampling System provided an alignment of curriculum and assessment along with compliance with prior legislation that mandated the development or adoption of developmentally appropriate assessment for children pre-kindergarten through grade 3. The first statewide administration of the first-grade readiness test occurred in spring of 2001.

Method

Data Source and Sampling

This study was based on a cluster random sample of 27 school districts with student assessment records consisting of first-grade teacher ratings and third-grade achievement scores. Originally, 30 school districts were randomly selected, but three districts did not have the necessary data. Two extant data files were merged and provided to the author by the state's department of education. The merged data file included first-grade teacher ratings on the personal and social development, language and literacy, and mathematical thinking domains of a standards-based, curriculum-embedded performance assessment administered in spring of 2002; third-grade scale scores on the English language arts and mathematics subscales of a standards-based, criterion-referenced test administered in spring of 2004; student demographic variables; and school and district codes.

Specific to this study, the English language arts and mathematics domains and subscales on the curriculum-embedded performance assessment and the achievement tests, respectively, were selected because of the national focus on reading and math achievement, as reflected in the No Child Left Behind legislation. Furthermore, only spring ratings of first-grade students' performance on the indicators for language and literacy and mathematical thinking were selected to be consistent with the spring administration of the third-grade achievement tests in English language arts and mathematics. That is, using spring ratings instead of fall or winter ratings on the performance assessment represented one year of academic growth of students in language and literacy and mathematical thinking. A brief description of each assessment measure as it relates to this study follows.

First-grade performance assessment. Teacher ratings were obtained using a low-stakes, standards-based, curriculum-embedded performance assessment (hereafter referred to as performance assessment) based on the Work Sampling System. The language and literacy domain consisted of 12 indicators and the mathematical thinking domain consisted of 14 indicators (Huynh, Prior, & Gallant-Taylor, 2002). Specific examples of each indicator can be found in *Using Work Sampling Guidelines and Checklists: An Observational Assessment* (Dichtelmiller, Jablon, Meisels, Marsden, & Dorfman, 1998).

Using the developmental guidelines and checklists for each domain, teachers observed and rated students' performance on the language and literacy and mathematical thinking indicators as: 1 = *not yet*, 2 = *in process*, and 3 = *proficient*. Ratings of *not yet* indicated that the skill, knowledge, or behavior had not been demonstrated; ratings of *in process* indicated that the skill, knowledge, or behavior was emergent and was not demonstrated consistently; and

ratings of *proficient* indicated that the skill, knowledge, or behavior was firmly within the child's range of performance (Dichtelmiller et al., 1998). Training sessions were provided to teachers at the district and the regional levels in 2001, and additional training sessions were conducted by individual school districts in subsequent years as needed (Huynh, Prior, & Gallant-Taylor, 2002). The internal consistency—Cronbach's alphas—of .98 were reported for the language and literacy and mathematical thinking domain scores for the spring observation period for all first-grade students in 2001. Domain scores were computed by summing teachers' ratings of each indicator on the language and literacy and mathematical-thinking domains. Thus, performance assessment domain scores ranged from 12 to 36 on the language and literacy domain and from 14 to 42 on the mathematical thinking domain. For this study, Cronbach's alpha of teachers' ratings was .93 for both the language and literacy domain and for the mathematical thinking domain.

Third-grade achievement tests. The third-grade achievement tests were untimed high-stakes standards-based criterion-referenced assessments administered to students in grade 3 through a statewide assessment program. The tests assessed mathematics and English language arts using multiple-choice and constructed-response items in the spring of an academic year. The English language arts subtest also consists of an extended writing item. Test results were reported as total scale scores (i.e., theoretical minimum and maximum scores are computed as grade level times 100, plus or minus 64) and performance levels (i.e., *Below Basic*, *Basic*, *Proficient*, and *Advanced*) for each of the academic areas (Huynh, Meyer, & Barton, 2000). Reliability indices (i.e., Cronbach's alpha and KR-21) for the 1999 administration of the English language arts and mathematics achievement tests were larger than .85 for all students and across gender and ethnicity (Huynh, Meyer, & Barton, 2000).

Sample. Student records meeting the following criteria were included in the study: (a) records contained both first-grade performance assessment domain scores for 2002 and third-grade achievement scale scores in English language arts and mathematics for 2004, (b) records indicated that students were tested at grade level for the achievement tests in English language arts and mathematics (i.e., students in third grade took the grade 3 version of the achievement test), and (c) performance assessment domain scores ranged from 12 to 36 for language and literacy and from 14 to 42 for mathematical thinking (the range of scores was the natural minimum and maximum for the performance assessment, based on the numerical values assigned for rating categories). Hence, the sample of records for this study consisted of 4292 student records representative of 131 first-grade schools and 137 third-grade schools. Classroom-level data were not included in the data file.

Descriptive characteristics of students, including descriptive statistics for performance assessment domain scores and achievement tests scores, are presented in Table 1. As reflected in the table, the sample of records represents 50% female, 49% non-White, 58% eligible for subsidized lunch, and 11% on an individualized education plan. Achievement tests were administered to third-grade students in spring of 2004.

Table 1. *Descriptive Characteristics of Students* (*n* = 4292)

Characteristic	*n*	%	*M*	*SD*	*Min*	*Max*
Gender						
Female	2146	50.0				
Race						
Non-White	2104	49.0				

Individual Education Plan	472	11.0
Eligible for free or reduced-price lunch	2489	58.0
Eligible for free or reduced-price lunch	2489	58.0

Table 1 (Continue). *Descriptive Characteristics of Students* (*n* = 4292)

Characteristic	*n*	%	*M*	*SD*	*Min*	*Max*
Performance Assessment						
Mathematical Thinking	4292		37.61	5.28	14.00	42.00
Language and Literacy	4292		31.93	4.76	12.00	36.00
Achievement Tests						
Mathematics	4292		308.71	11.94	265.00	344.00
English language arts	4292		311.35	14.26	257.00	352.00

Note. Performance assessments were administered to first-grade students in spring of 2002.

Selected descriptive characteristics of first- and third-grade schools students attended, available on the state's report card Web site, are presented in Table 2. The author hypothesized that the selected school-level demographic variables influenced student achievement. Due to the varied organizational structure of schools in this state (e.g., K-2, K-3, K-5), approximately 23% of students did not attend the same elementary school for grades 1 and 3. As reflected in the table, the mean student-teacher ratio and attendance rates for students and teachers were similar for first- and third-grade schools attended. However, third-grade schools, on average, spent more dollars per student.

Table 2. *Descriptive characteristics of the First- and Third-Grade Schools*

	First-Grade School (*n* = 131)				Third-Grade School (*n* = 137)			
Characteristic	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Dollars spent per student	5701.47	978.28	3939.0	9461.0	6230.04	1209.29	3601.0	9539.0
Student-teacher ratio	17.74	3.25	3.40	25.60	18.65	2.73	9.70	26.90
Student attendance rate	96.16	1.66	91.00	100.00	96.65	1.22	94.50	99.90
Teacher attendance rate	94.76	1.50	89.20	99.10	94.83	1.41	90.80	99.30

Note. Descriptive characteristics of schools were obtained from school report cards posted on the state's Web site. The characteristics are not exhaustive but reflect a small sample of characteristics hypothesized by the author to affect student achievement.

Statistical Analyses

Statistical analyses included computing the partial correlation coefficients for the association between first-grade performance assessment domain scores and third-grade achievement test scores and cross-classified multilevel regression analyses. For all analyses, the statistical significance level was set at $\alpha = .05$. However, alpha levels of .001 were also reported. An in-depth description of each statistical analysis follows.

Correlations. Correlations between performance assessment domain scores and achievement tests scale scores were computed to determine the amount of shared variance between first-grade teachers' summed ratings on the performance assessments and students' later performance on the third-grade achievement tests, controlling for student demographic variables (i.e., gender, race, individualized education plan status, and eligibility for subsidized meals). The effect of student demographic variables was partitioned out to mitigate the possible confounding that inclusion of the variables would have created in the association between measures. The magnitude of the associations were described using Cohen's conventions (1992) of *small* = .10, *medium* = .30, and *large* = .50.

Cross-classified random effects multilevel regression analyses. Two-level multilevel regression analyses were used to partition the variance associated with the nested nature of the data (i.e., students nested within schools). Specifically, cross-classification random effects models were specified because students were not purely nested within one school over the two-year time period. That is, students did not necessarily attend the same school in first and third grade. Unlike traditional two-level multilevel regression models, cross-classification random effects models do not assume that the nesting is purely hierarchical (Raudenbush & Bryk, 2002; Beretvas, 2008). Thus, the cross-classification random effects models allow the researcher to account for the cross-classification of students across first- and third-grade schools attended.

The cross-classified random effects models used in this study were estimated using maximum likelihood estimation with an independent z-structure and specified to address each research question. Dependent variables were third-grade achievement tests scores in English language arts and mathematics. Independent variables were student demographic variables and language and literacy and mathematical thinking domain scores, computed as the sum of teacher ratings for each indicator on the domain, on the performance assessment. The demographic variables were coded such that race (RACE) was coded 0 = White and 1 = non-White, gender (GENDER) was coded 0 = male and 1 = female, eligibility for subsidized meals (LUNCH) was coded 0 = ineligible for subsidized meals and 1 = eligible for free or reduced-priced lunch, and individual education plan status (IEP) was coded 0 = no IEP and 1 = IEP. The software *HLM 6.08: Hierarchical Linear and Nonlinear Modeling* (Raudenbush, Bryk, & Congdon, 2004) was used for the multilevel analyses.

The unconditional and conditional cross-classification random effects models for this study were specified as follows. The mixed-effects unconditional model for predicting student achievement is:

$$Y_{ijk} = \theta_o + b_{ooj} + c_{ook} + e_{ijk}, \quad (1)$$

where Y_{ijk} is the third-grade mathematics or English language arts achievement test scores of student *i* in grade 1 school *j* and grade 3 school *k*; θ_o is the overall grand-mean achievement score of all students; b_{ooj} is the random effect for grade 1 school attended [$b_{ooj} \sim N(0, \tau_{boo})$]; c_{ook} is the random effect for grade 3 school attended [$c_{ook} \sim N(0, \tau_{coo})$]; and e_{ijk} is the deviation of student *ijk*'s score from the cell mean [$e_{ijk} \sim N(0, \sigma^2)$].

The conditional models were specified the same at level 1 but differently at level 2 for mathematics and English language arts achievement. Differences in the models occurred at level 2 because the intercept variability across first-grade school attended in the mathematics model was not statistically significant ($p > .05$), indicating that the effect be better modeled as fixed rather than random. Student demographic variables and two of the four school demographic variables (i.e., dollars spent per student and teacher-student ratio) were included in the model to potentially account for some of the unexplained variances found in

the unconditional model and to serve as controls in the analyses. Only dollars spent per student and teacher-student ratio were included at the school levels because inclusion of student attendance rates and teacher attendance rates created extremely large standard errors, possibly due to multicollinearity. Level 1 was as follows:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(X)_{ijk} + \pi_{2jk}(Gender)_{ijk} + \pi_{3jk}(IEP)_{ijk} + \pi_{4jk}(Race)_{ijk}$$

$$+ \pi_{5jk}(Lunch)_{ijk} + e_{ijk}, e_{ijk} \sim N(0, \sigma^2), \quad (2)$$

where Y_{ijk} is the third-grade mathematics or English language arts achievement test score of student *i* in grade 1 school *j* and grade 3 school *k*; X is the grand-mean centered mathematical thinking or language and literacy domain score; π_{0jk} is the overall grand-mean math achievement score for students attending the same school both years; $\pi_{1jk}...\pi_{5jk}$ are the regression coefficients relating the predictors to the mathematics achievement; and e_{ijk} is a random effect that represents the deviation of student *ijk*'s outcome from the predicted outcome based on the predictors in the model.

Level 2 (for mathematics):

$$\pi_{0jk} = \theta_0 + c_{00k} + \gamma_{01}(Sch1Dllrs) + \gamma_{02}(Sch1Ratio) + \beta_{03}(Sch3Ratio)$$

$$+ \beta_{04}(Sch3Dllrs) \quad (3)$$

$$\pi_{1jk} = \theta_1 + b_{10j}$$

$$\pi_{2jk} = \theta_2$$

$$\pi_{3jk} = \theta_3$$

$$\pi_{4jk} = \theta_4$$

$$\pi_{5jk} = \theta_5; \text{ and}$$

Level 2 (for English language arts):

$$\pi_{0jk} = \theta_0 + b_{00j} + c_{00k} + \gamma_{01}(Sch1Dllrs) + \gamma_{02}(Sch1Ratio) + \beta_{01}(Sch3Ratio)$$

$$+ \beta_{02}(Sch3Dllrs) \quad (4)$$

$$\pi_{1jk} = \theta_1 + b_{10j}$$

$$\pi_{2jk} = \theta_2$$

$$\pi_{3jk} = \theta_3$$

$$\pi_{4jk} = \theta_4$$

$$\pi_{5jk} = \theta_5,$$

where Sch1Dllrs and Sch1Ratio are the School 1 dollars spent per student and student-teacher ratio, respectively; Sch3Dllrs and Sch3Ratio are the School 3 dollars spent per student and student-teacher ratio, respectively; θ_0 is the grand mean, when all School 3 variables are held constant; $\theta_1...\theta_5$ are the fixed effects of the student-level predictors; γ_{01} and γ_{02} are the fixed effects of the School 1 variables; $\beta_{01}$and β_{02} are the fixed effects of the School 3 variables; b_{00j} and c_{00k} are the residual random effect of School 1 and School 3, respectively, on the overall grand-mean achievement score after taking into account the School 1 and School 3 predictors; and b_{10j} is the residual random effect of School 1 on the grand-mean centered performance assessment domain score.

Mathematical thinking and language and literacy domain scores were grand-mean centered (i.e., centered on the school performance assessment domain score means) to create a meaningful reference point for the predictors in the model. That is, by centering on the school performance domain score means, a score of zero can be interpreted as a student receiving the average performance assessment domain score; whereas if the performance assessment domain scores were not centered, a score of zero would be interpreted as a student receiving a performance assessment domain score of zero is impossible. In addition, standardized coefficients were computed as $\frac{s_X}{s_Y}\theta$ (Snijders & Bosker, 1999) for the multilevel results to provide a scale for effects independent of the measurement units. Hence, results are reported in terms of standard deviations.

Results

Correlations between Performance Assessment Domain Scores and Achievement Test Scores

As presented in Table 3, positive significant associations were found between language and literacy domain scores and English language arts scale scores (r = .37, p < .001) and between mathematical-thinking domain scores and mathematics scale scores (r = .30, p < .001). The magnitude of the associations can be described as *medium*. Based on the partial correlation coefficients, approximately 14% of the variability in language and literacy domain scores overlapped with the variability in English language arts achievement scale scores, whereas approximately 9% of the variability in mathematical thinking domain scores overlapped with the variability in mathematics achievement scale scores.

Table 3. *Correlations among First Grade Performance Assessment Domain Scores and Third Grade Achievement Scale Scores, Controlling for Student Demographic Variables* (n = 4287)

	First-Grade Performance Domains		Third-Grade Achievement Scales	
Variable	Language & literacy	Mathematical thinking	English language arts	Mathematics
Language & literacy	—	.75*	.37*	.29*
Mathematical thinking		—	.32*	.30*
English language arts			—	.62*
Mathematics				—

Note. Partial correlation coefficients were computed for 4,287 records that included all demographic variables.

*p < .001, two-tailed.

Cross-Classified Multilevel Analyses for Mathematics and English Language Arts Achievement

Mathematics achievement. Estimates for the fixed- and random-effects for the unconditional and conditional models for predicting student mathematics achievement are reported in Table 4. For the unconditional model, the mean mathematics achievement score was 308.26, with a total variance of 145.31. Of the total variance, approximately 2% of the variance was between scores of students who attended the same first-grade school but a different third-grade school and approximately 12% of the variance was between scores of students who

attended the same third-grade school but a different first-grade school. Thus, differences in mathematics achievement were attributed more to the third-grade school attended than to the first-grade school attended. Additionally, approximately 86% of the total variance was within cross-classification of the first- and third-grade school attended (i.e., within students who attended the same first- and third-grade school).

In the conditional model, controlling for student demographic variables and school demographic variables, mathematical thinking domain scores were statistically significant predictors of mathematics achievement (θ_1 = .69, p < .001). Thus, each additional standard deviation on the mathematical thinking domain score led, on average, to an increase of a .31 standard deviation in third-grade mathematics achievement. Furthermore, the inclusion of student and school variables in the model decreased the estimated student-level variance from 124.74 to 93.94 (approximately a 25% reduction) and the estimated third-grade school attended variance from 18.05 to 13.88 (approximately a 23% reduction).

Table 4. *Estimation of Fixed and Random Effects for Based on Cross-Classification Models for Mathematics Achievement*

		Unconditional Model			Conditional Model		
Fixed effects		Coefficient	se	t Ratio	Coefficient	se	t Ratio
Intercept, π_0	θ_{000}	308.26**	0.49	625.95	313.66**	5.00	62.71
School 1 Dollars	γ_{01}				-0.00	0.00	-0.09
School 1 Ratio	γ_{02}				0.09	0.10	0.86
School 3 Ratio	β_{01}				-0.07	0.15	-0.47
School 3 Dollars	β_{02}				0.00	0.00	0.02
Domain score, π_1	θ_{100}				0.69**	0.04	17.17
Gender, π_2	θ_{200}				-0.25	0.30	-0.83
IEP, π_3	θ_{300}				-5.24**	0.49	-10.71
Race, π_4	θ_{400}				-4.32**	0.38	-11.35
Lunch, π_5	θ_{500}				-3.63**	0.37	-9.75
Random effects			*Estimate*			*Estimate*	
School 1, b_{00j}	τ_{b00}		2.52**				
Domain score, b_{10j}	τ_{b10}					0.04**	
School 3, c_{00k}	τ_{c00}		18.05**			13.88**	
Students, e_{ijk}	σ^2		124.74			93.94	
Model Deviance			33091.96			31889.28	
Model *df*			4			13	

Note. Ratio = Student-teacher ratio. Dollars = Dollars spent per student. Domain score = First-grade mathematical thinking domain score. School 1 = First-grade school attended. School 3 = Third-grade school attended. For the unconditional model, *df* = 4,291 for fixed effects; *df* = 130 for School 1 random effect; and *df* = 136 for School 3 random effect. For the conditional model, *df* = 4,282 for fixed effects; *df* = 126 for mathematical thinking domain score random effect; and *df* = 132 for School 3 random effect. **$p \leq .001$. *$p < .05$.

English language arts achievement. Estimates for the fixed- and random-effects for the unconditional and conditional models for English language arts achievement are presented in Table 5. The mean English language arts achievement was 310.77 and the total variance was

206.04 for the unconditional model. Of the total variance, approximately 11% was between scores of students who attended the same first-grade school but a different third-grade school and approximately 2% was between scores of students who attended the same third-grade school but a different first-grade school. Unlike mathematics achievement findings, differences in English language arts achievement could be attributed more so to the first-grade school attended than to the third-grade school attended. Similar to the findings for mathematics achievement, approximately 87% of the total variance was due to the within cross-classification of first- and third-grade school attended.

For the conditional model, controlling for student and school demographic variables, language and literacy domain scores were found to be statistically significant predictors of third-grade English language arts achievement ($\theta_1 = 1.12$, $p < .001$). Hence, we can expect, on average, a .37 standard deviation increase in third-grade English language arts achievement for each additional standard deviation increase in the language and literacy domain score. The inclusion of student and school demographic variables in the model decreased the student-level variance from 179.79 to 122.59 (approximately a 32% reduction) and decreased the first-grade school attended variance from 22.91 to 6.57 (approximately a 71% reduction) but increased the third-grade school attended variance from 3.34 to 7.69 (approximately a 130% increase).

Table 5. *Estimation of Fixed and Random Effects for Based on Cross-Classification Models for English Language Arts Achievement*

		Unconditional Model			Conditional Model		
Fixed effects		Coefficient	se	t Ratio	Coefficient	se	t Ratio
Intercept, π_0	θ_{000}	310.77**	0.57	547.03	311.69**	5.29	58.97
School 1 Dollars	γ_{01}				0.00	0.00	0.55
School 1 Ratio	γ_{02}				0.22	0.13	1.71
School 3 Ratio	β_{01}				-0.09	0.15	-0.62
School 3 Dollars	β_{02}				-0.00	0.00	-0.03
Domain score, π_1	θ_{100}				1.12**	0.05	21.19
Gender, π_2	θ_{200}				2.67**	0.35	7.69
IEP, π_3	θ_{300}				-5.11**	0.56	-9.07
Race, π_4	θ_{400}				-5.10**	0.43	-11.76
Lunch, π_5	θ_{500}				-4.32**	0.43	-10.12

Table 5 (Continue). *Estimation of Fixed and Random Effects for Based on Cross-Classification Models for English Language Arts Achievement*

Random effects		*Estimate*	*Estimate*
School 1, b_{00j}	τ_{b00}	22.91**	6.57**
Domain score, b_{10j}	τ_{b10}		0.08**
School 3, c_{00k}	τ_{c00}	3.34*	7.69**
Students, e_{ijk}	σ^2	179.79	122.59
Model Deviance		34648.67	33029.83
Model *df*		4	15

Note. Ratio = Student-teacher ratio. Dollars = Dollars spent per student. Domain score = First-grade mathematical thinking domain score. School 1 = First-grade school attended. School 3 = Third-grade school attended. For the unconditional model, *df* = 4291 for fixed effects; *df* = 130 for School 1 random effect; and *df* = 136 for School 3 random effect. For the conditional model, *df* = 4282 for fixed effects; *df* = 112 for School 1 random effect; *df* = 116 for language and literacy domain score random effect; and *df* = 132 for School 3 random effect. Chi-square statistics were based on 117 of 131 first-grade schools that had sufficient data for computation.
**$p \leq .001$. *$p < .05$.

Discussion

This study applied a cross-classified random effects multilevel approach to determine the extent to which first-grade teacher ratings, as indicated by performance assessment domain scores, predicted third-grade student mathematics and English language arts achievement on a high-stakes, standards-based, criterion-referenced test. Correlations between performance assessment domain scores and achievement subscale scores were also computed to address the research questions. Results from this study provide support for the use of teacher ratings to predict students' later achievement, within the context of a high-stakes accountability system.

Overall, positive medium associations were observed between first-grade performance assessment domain scores and third-grade achievement scale scores, controlling for students' gender, IEP status, race, and eligibility for subsidized meals. The associations found in this study presents about a 14% overlap in the variability among language and literacy and English language arts scores and about a 9% overlap in the variability among mathematical thinking and mathematics scores across a two-year period. These results are slightly smaller than those found by Gallant (2009) and much smaller than those found by Meisels et al. (1995). Gallant (2009) found approximately an 18% overlap in the variability among the English language arts scores and approximately an 11% overlap in the mathematics scores over two years without controlling for student demographic variables. Whereas, Meisels et al. (1995) found within a one-year period at least a 56% overlap in the variability among kindergarteners' fall Work Sampling System (WSS) checklists (i.e., language and literacy and concept and number) scores and the spring Kindergarten Achievement Battery of the

Woodcock-Johnson Psychoeducational Battery-Revised (WJ-R) total score without controlling for student demographic variables. Not controlling for student demographic variables or other extraneous variables may have contributed to the larger percentages of overlap in the Gallant and Meisels et al. studies.

Furthermore, the correlation coefficients obtained in this study are not atypical given the time span of the scores used. Kim and Suen (2003) suggests that a one-year prediction using ratings of early childhood cognitive ability is expected to have the highest predictive validity coefficients in predicting later achievement. Determining predictive validity with time intervals longer than one year between early assessments and later achievement has the potential for poorer predictability compared to shorter time intervals. Hence, the covariance estimates found in this study are reasonable considering the two-year time period between first-grade teacher observations and third-grade student achievement. The presence of overlap in variance among the instruments suggests some alignment of curriculum content standards across grade levels, as was the case for this state, and that similar content are being measured across years.

Moreover, controlling for student and school variables, performance assessment domain scores were statistically significant predictors of third-grade achievement. On average, the mean achievement scores are expected to increase about one-third of a standard deviation for each one-point increase in domain score. This finding was slightly lower than that found by Gallant (2009). The author found about a .34 standard deviation increase in mathematics achievement and about a .43 standard deviation increase in English language arts achievement for each one-point increase in the performance domain score. The smaller standard deviations at the school level are possibly attributed to similarities in student scores across schools due to the cross-classification of students and schools and the fact that about 77% of students attended the same school both years.

Conclusions

The findings from this study provide support for the use of a curriculum-embedded performance assessment, based on the Work Sampling System, within the context of a state accountability system. The potential long-term benefits to students in schools that use developmentally appropriate curriculum and assessments cannot be disregarded. For example, students exposed to a curriculum-embedded performance assessment for at least three years were found to display gains in reading and mathematics achievement from third to fourth grade, as measured by a norm-referenced test, compared to students who were not exposed to a curriculum-embedded performance assessment (Meisels et al., 2003). Hence, the alignment of early childhood practices, curriculum content, and assessment is essential to meeting the overall developmental needs of all children (e.g., NAEYC & NAECS/SDE, 1991; Shephard, Kagan, & Wurtz, 1998) as well as ensuring that students achieve later academic successes.

It is important to remember that within early childhood and elementary school settings, principals and parents often rely on teachers to make decisions regarding the academic performance of students. The use of developmentally appropriate practices—such as using continuous observations instead of group-administered tests to assess students' academic, social, emotional, and physical development in early childhood programs serving children from birth to age 8—has been advocated by early childhood professionals and professional organizations for decades (e.g., National Association for the Education of Young Children [NAEYC], 1988, 1997, 2009; NAEYC & National Association of Early Childhood Specialists in State Departments of Education [NAECS/SDE], 2003; National Association of School

Psychologists, 2005). The use of continuous assessments in the early grades recognizes the continuum on which young children develop. Information obtained from teacher assessments can be used to ensure that students are receiving not only the necessary academic services but also the necessary stimuli to enhance their overall development. Using teachers as evaluators of students' academic achievement is crucial because of teachers' awareness and knowledge of curriculum content standards at the next grade levels (Mashburn & Henry, 2004).

Although the findings in this study were positive, the study was not without limitations. First, the use of domain score ranges as a criterion for inclusion in the study reduced the number of students per school in the study and potentially reduced the within- and between-school variances. Future studies should consider using the mean ratings instead of domain scores to reduce the number of eliminated records. Second, the retrospective nature of this study and the use of extant data made it impossible to determine what interventions, if any, were provided to students in the second grade based on first-grade ratings on the performance assessment. It was also unknown if schools had adopted any special reading or mathematics programs for third-grade students during the academic year and prior to the spring administration of the achievement tests. Hence, the presence of special programs that schools may have implemented was not controlled for in the current study.

▪ ▪ ▪

Acknowledgements

The author would like to thank Theresa Siskind, Shiqi Hao, and Christine Schneider for their assistance in providing the data used in this study, and James L. Moore III and Mark Allen for comments provided on an earlier version of the manuscript.

Dorinda J. GALLANT received the Ph.D. degree in Educational Psychology and Research (Research Track) from University of South Carolina in 2005. She is an associate professor in the Department of Educational Studies, Quantitative Research, Evaluation and Measurement since 2005. Her research interests include applied measurement in elementary, secondary, and postsecondary education within the context of program, product, or personnel evaluation.

References

Alpern, G., Boll, T., Shearer, M. (1986). *Developmental Profile II Manual*. Los Angeles: Western Psychological Services.

Beretvas, S.N. (2008). Cross-classified random effects models. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 161-197). Charlotte, NC: Information Age Publishing, Inc.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Dichtelmiller, M.L., Jablon, J.R., Meisels, S.J., Marsden, D.B., & Dorfman, A.B. (1998). *Using Work Sampling guidelines and checklists: An observational assessment*. Ann Arbor, MI: Rebus.

Education Accountability Act of 1998, Chapter 18 (1998).

Gallant, D.J. (2005). *Predictive nature of a curriculum-embedded performance assessment for young children* (Doctoral dissertation). University of South Carolina, Columbia, SC.

Gallant, D.J. (2009). Predictive validity evidence for an assessment program based on the Work Sampling System in mathematics and language and literacy. *Early Childhood Research Quarterly, 24*, 133-141. doi: 10.1016/j.ecresq.2009.03.003

Gresham, F.M., & Elliott, S.N. (1990). *Social Skills Rating System*. Circle Pines, MN: American Guidance Service.

Hecht, S.A., & Greenfield, D.B. (2001). Comparing the predictive validity of the 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.

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

Huynh, H., Meyer, J.P., III, & Barton, K. (2000). *Technical documentation for the 1999 Palmetto Achievement Challenge Tests of English language arts and mathematics, grades three through eight*. Location: State department of education.

Huynh, H., Prior, S.V., & Gallant-Taylor, D.J. (2002). *Technical documentation for the [State] Readiness Assessment of kindergarten and grade one*. Location: State department of education.

Jastak, J.F., Bijou, S.W., & Jastak, S.R. (1965). *Wide Range Achievement Test*. Wilmington, DE: Guidance Association.

Kim, J., & Suen, H.K. (2003). Predicting children's academic achievement from early assessment scores: A validity generalization study. *Early Childhood Research Quarterly, 18*, 547-566.

Mashburn, A.J., & Henry, G.T. (2004). Assessing school readiness: Validity and bias in preschool and kindergarten teachers' ratings. *Educational Measurement: Issues and Practices, 23*(4), 16-30.

Meisels, S.J., Atkins-Burnett, S., Xue, Y., Nicholson, J., Bickel, D.D., & Son, S-H. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives, 11*(9), Retrieved on January 19, 2006, from http://epaa.asu.edu/epaa/v11n9/.

Meisels, S.J., Bickel, D.D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal, 38*, 73-95.

Meisels, S.J., Jablon, J., Marsden, D.B., Dichtelmiller, M.L., & Dorfman, A.B. (1994). *The Work Sampling System: An overview* (3rd ed.). Ann Arbor, MI: Rebus Inc.

Meisels, S.J., Jablon, J., Marsden, D.B., Dichtelmiller, M.L., & Dorfman, A.B. (2001). *The Work Sampling System: An overview* (4th ed.). Ann Arbor, MI: Rebus Inc.

Meisels, S.J., Liaw, F., Dorfman, A., & Nelson, R.F. (1995). The Work Sampling System: Reliability and validity of performance assessment for young children. *Early Childhood Research Quarterly, 10*, 277-296.

National Association for the Education of Young Children. (1997). Developmentally appropriate practice in early childhood programs serving children from birth through age 8 [On-line]. Washington, DC: Authors. Retrieved May 2, 2007 from www.naeyc.org/about/positions.asp.

National Association for the Education of Young Children. (1988). Position statement on standardized testing for young children 3 through 8 years of age. *Young Children, 43*(3), 42-47.

National Association for the Education of Young Children. (2009). Developmentally appropriate practice in early childhood programs serving children from birth through age 8 [On-line]. Washington, DC: Authors. Retrieved May 1, 2009, from http://www.naeyc.org/about/positions.asp.

National Association for the Education of Young Children and National Association of Early Childhood Specialists in State Departments of Education. (1991). Guidelines for appropriate curriculum content and assessment in programs serving children ages 3 through 8. *Young Children, 46*(3), 21-38.

National Association for the Education of Young Children and National Association of Early Childhood Specialists in State Departments of Education. (2003). *Early childhood curriculum, assessment, and program evaluation: Building an effective, accountable system in programs for children birth through age 8* [On-line]. Washington, DC: Authors. Retrieved January 19, 2006, from http://www.naeyc.org/about/positions.asp.

National Association of School Psychologists. (2005). *Position statement on early childhood assessment* [On-line]. Bethesda, MD: Authors. Retrieved May 2, 2007, from http://www.nasponline.org/about_nasp/pospaper_eca.aspx.

Quay, L.C. & Steele, D.C. (1998). Predicting children's achievement from teacher judgements: An alternative to standardized testing. *Early Education & Development, 9*(3), 207-218.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Raudenbush, S.W., Bryk, A.S., & Congdon, R. (2004). *HLM 6.0: Hierarchical linear and nonlinear modeling* [Computer software]. Chicago, IL: Scientific Software International.

Shephard, L., Kagan, S., & Wurtz, E. (1998). *Principles and recommendations for early childhood assessments* [On-line]. Washington, DC: National Education Goals Panel. Retrieved January 19, 2006, from http://govinfo.library.unt.edu/negp/reports/prinrec.pdf.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage Publications Inc.

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*(5), 507-517.

Woodcock, R.W., & Johnson, M.B. (1989). *Woodcock-Johnson psychoeducational battery-Revised*. Allen, TX: DLM Teaching Resources.

This page is intentionally left blank