The predictive validity of kindergarten readiness judgments: Lessons from one state:

Article *in* The Journal of Educational Research · July 2016

DOI: 10.1080/00220671.2015.1039111

CITATIONS
18

READS
422

3 authors, including:

D. Betsy Mccoach
University of Connecticut
172 PUBLICATIONS **7,225** CITATIONS

SEE PROFILE

Huihui Yu
Yale University
16 PUBLICATIONS **77** CITATIONS

SEE PROFILE

The Journal of Educational Research

ISSN: 0022-0671 (Print) 1940-0675 (Online) Journal homepage: http://www.tandfonline.com/loi/vjer20

The predictive validity of kindergarten readiness judgments: Lessons from one state

Jessica Goldstein, D. Betsy McCoach & HuiHui Yu

To cite this article: Jessica Goldstein, D. Betsy McCoach & HuiHui Yu (2017) The predictive validity of kindergarten readiness judgments: Lessons from one state, The Journal of Educational Research, 110:1, 50-60, DOI: 10.1080/00220671.2015.1039111

To link to this article: http://dx.doi.org/10.1080/00220671.2015.1039111

Published online: 28 Jul 2016.

THE JOURNAL OF EDUCATIONAL RESEARCH
2017, VOL. 110, NO. 1, 50–60
http://dx.doi.org/10.1080/00220671.2015.1039111

Routledge
Taylor & Francis Group

The predictive validity of kindergarten readiness judgments: Lessons from one state

Jessica Goldstein, D. Betsy McCoach, and HuiHui Yu

Department of Educational Psychology, University of Connecticut, Storrs, Connecticut, USA

ABSTRACT

Recent federal investments in early childhood assessment systems are the result of a national need for developmentally appropriate, psychometrically sound instruments to monitor young children and evaluate the effectiveness of their learning programs. In this paper, we examined the association between teachers' perceptions of their students at the start of kindergarten and academic achievement in Grade 3 with hierarchical linear modeling using state-level data from nearly 30,000 students. The analyses showed that such an association exists even after accounting for student-level and school-level demographic variables and is moderated by the percentage of free-lunch-eligible students in a given school. Implications of these findings related screening and assessment at kindergarten entry are discussed.

ARTICLE HISTORY
Received 10 July 2014
Revised 12 March 2015
Accepted 3 April 2015

KEYWORDS
Assessment; early childhood; validity

Early childhood education in the United States is on the brink of great change. As of January 2014, over $1 billion in federal Race to the Top Early Learning Challenge grants were awarded to 20 states for the development and enhancement of comprehensive early childhood assessment systems (CECAS). The National Research Council (2008) defined a CECAS as a network of developmental screening measures, formative assessments, measures of environmental quality, measures of the quality of adult-child interactions, and a kindergarten entry assessment. In September 2013, the U.S. Department of Education awarded more than $15 million in Enhanced Assessment Grants (EAGs) to three state education agencies for the development or enhancement of kindergarten entry assessments. These federal investments are the result of a national need for developmentally appropriate, psychometrically sound instruments to monitor young children and evaluate the effectiveness of their early childhood learning programs. The purpose of this paper is to explore the predictive validity of a single indicator of kindergarten readiness using existing state data. Specifically, we examine the association between teacher judgments of their students at kindergarten entry and the Grade 3 achievement of those same students as measured by one state's summative assessment.

A focus on kindergarten entry

Current policy

Kindergarten marks a transition point for children as they move from early learning and development settings to the K–12 system. Data collected at kindergarten entry offer a cumulative glimpse into children's early experiences and offer both a baseline for kindergarten instruction and for measuring future progress. Kindergarten entry assessments (KEA) are designed to be administered by classroom teachers as they begin to understand their students. Federal funding competitions were recently designed to ensure a level of uniformity to KEAs across the nation. Specifically, the Race to the Top Early Learning Challenge required that each submitting state employ a KEA that: "is aligned with the State's Early Learning and Development Standards and covers all Essential Domains of School Readiness… [and] is valid, reliable, and appropriate for the target population and for the purpose for which it will be used, including for English learners and children with disabilities" (U.S. Department of Education, 2011, p. 68). The federal government also required the KEAs to address the essential domains of school readiness, which are defined as language and literacy development, cognition and general knowledge (including early mathematics and early scientific development), approaches toward learning, physical well-being and motor development, and social and emotional development (U.S. Department of Education, n.d.).

Before the Race to the Top Early Learning Challenge, 25 states required universal assessments of kindergarten students (Stedron & Berger, 2010). Definitions of the knowledge and skills measured at kindergarten entry varied within that set of states. Eleven states required schools to use specific instruments to address all five essential domains of school readiness. Slightly less than half of the states required specific instruments specifically for the evaluation of early literacy. The remaining states required neither a multidimensional measurement of readiness nor the use of a particular instrument. Also noteworthy, several of the states designed custom readiness measures based on teachers' observations of children's skills and abilities across multiple domains for the purpose of reporting aggregate data. One such readiness measure based on teacher observation is the foundation of this study.

CONTACT Jessica Goldstein jessica.goldstein@uconn.edu Department of Educational Psychology, University of Connecticut, 249 Glenbrook Road, Storrs, CT 06269.

Early learning skills as predictors of later outcomes

The unprecedented federal investment in CECAS and kindergarten entry assessments is a response to a growing body of research that learning and development in the early years is the foundation for future educational achievement (Alexander & Entwisle, 1988, 1996; Bowman, Donovan, & Burns, 2001; Duncan et al., 2007; Neuman & Dickinson, 2001; Saluja, Scott-Little, & Clifford, 2000). Differences in academic performance begin early and persist, and perhaps worsen, over time. One early study found that 88% of students identified as poor readers in Grade 1 were also considered poor readers in fourth grade (Juel, 1988). More recently, Sabol and Pianta (2012) identified associations between social-emotional skills at 54 months of age and achievement in fifth grade in one study of 944 children. Other studies have found that early reading problems persisted into high school (Francis, Fletcher, Shaywitz, Shaywitz, & Rourke, 1996; Shaywitz et al., 1999). These patterns exist in mathematics as well. An achievement gap in mathematics can be seen in children as young as age three (Case & Griffin, 1990; Jordan, Huttenlocher, & Levine, 1992), and the effects of this gap linger not only through kindergarten and Grade 1, but can persist into middle and high school (Berkner & Chavez, 1997; Braswell et al., 2001; Denton & West, 2002; Entwisle & Alexander, 1993; West, Denton, & Reaney, 2001). Gaps in the academic skills of Black and Hispanic children as compared to White children are evident even before children enter elementary school (Fryer & Levitt, 2006; Haskins & Rouse, 2005; Loeb, Bridges, Bassok, Fuller, & Rumberger, 2007). At the start of kindergarten, Hispanic children have been found to be less ready for school than White or Black students (Duncan & Magnuson, 2005; Fryer & Levitt, 2004; Reardon, 2003; Rumberger & Arellano, 2004; Zill, Collins, West, & Hausken, 1995). More recent data from the federal Early Childhood Longitudinal Study—Kindergarten Class of 2010–2011 showed that Asian and White first-time kindergarteners had higher reading and mathematics scores than Black, Hispanic, Native Hawaiian/Pacific Islander, and American Indian/Alaska Native students (Mulligan, Hastedt, Carroll, & Carlivati, 2012). Several studies have examined the predictive associations between skills at kindergarten entry and later achievement with large data sets. Hair, Halle, Terry-Humen, Lavelle, and Calkins (2006) used Early Childhood Longitudinal Study—Kindergarten Class of 1998–1999 (ECLS-K) to examine how patterns of school readiness predict first-grade outcomes. The analyses showed that children with social and emotional or health risks performed worse on Grade 1 academic outcomes while children with more positive developmental profiles had better outcomes. In a similar study, Halle, Hair, Burchinal, Anderson, and Zaslow (2012) analyzed data from the National Institute of Child Health and Human Development's Study of Early Child Care and Youth Development (NICHD SECCYD) and the ECLS-K to identify associations between skills at school entry and later school success. They found that children who entered with stronger school readiness skills tended to maintain their advantage over time, while children who entered with lower school readiness skills tended to maintain their relative disadvantage over time. Further, they found a domain-related association for the predictions. Specifically, mathematics skills at kindergarten entry were most predictive of subsequent mathematics skills and entry-level social skills provided the best prediction of later social skills.

Research on teacher judgments of academic skills

Predictive validity of teacher judgments

What does validity look like when the teacher serves as the assessor in a classroom of young children? Assessments of young children require independent assessors or knowledgeable raters such as teachers or parents. Independent assessors often prove too costly for large-scale assessment programs. Teachers have a familiarity with students that facilitates the data collection process and can be enhanced by the process of administering the assessment (Early Childhood Education Assessment-State Collaborative on Assessment and Student Standards, 2011; Scott-Little, Kagan, & Clifford, 2003).

Judgment accuracy is particularly critical in assessments for young children, which are designed to identify and support children who have potential learning problems and who may need special programs (Barnett, Macmann, & Carey, 1992; Lidz, 1983; Tramontana, Hooper, & Selzer, 1988). Though teacher judgment is a critical component of early childhood assessment, the literature lacks large-scale, empirical study of the issue. Several studies of teacher judgments with smaller samples are based on the use of curriculum-based measurement (CBM), which provides both direct and peer-independent estimates of students' skills in basic content areas. Begeny, Eckert, Montarello, and Storie (2008) analyzed data from 10 teachers of 87 first-, second-, and third-grade students from a suburban school in the northeast and found that teacher judgments were accurate predictors for students with strong oral fluency, but teachers had more difficulty when students had average to low oral reading fluency skills. Hamilton and Shinn (2003) specifically found that teachers overestimated the performance of lower achieving students in a study of 66 third-grade students. Other CBM researchers found a similar trends (Bates & Nettleback, 2001; Feinberg & Shapiro, 2003, 2009). Feinberg and Shapiro (2009) also found that teacher estimates of their students' skills were based on the relative standing of individual students, rather than an evaluation against established benchmarks in a study of 74 teachers and 148 students. These studies focus on the primary grades and are based on small sample sizes. Studies of the predictive validity of kindergarten teacher judgments with larger sample sizes are needed.

There is some promising research to support teacher identification of risk in the early years of schooling with other instruments. Demaray and Elliott (1998) studied relationships among data from the Academic Competence Scale of the Social Skills Rating System–Teacher version (Gresham & Elliott, 1990), the Kaufman Test of Educational Achievement, Brief Form (K-TEA; Kaufman & Kaufman, 1985), and a researcher-developed questionnaire with a sample of 12 teachers and 47 students. Similar to research on the CBM measures, the researchers found that while teachers can accurately judge student performance, they tended to be more accurate for high-achieving students. Martin and Shapiro (2011) found that while

teachers were accurate judges of their student performance on a Dynamic Indicators of Basic Early Literacy Skills (DIBELS) measure, teachers were more accurate for lower achieving students. The authors suggest this shift from the earlier literature is the result of "strong momentum driving educators toward more awareness of the importance of early literacy skills" (p. 353). Speece et al. (2011) studied reading accuracy, fluency, growth, and teacher ratings from 257 students in 16 classrooms to develop a screening measure for the identification of first-grade students at risk for reading difficulty. The authors noted that universal screening measures must be both valid and efficient in that they are quick and easy to administer and few are available. Their analyses identified that kindergarten teacher judgment accounted for the largest proportion of the variance in the prediction of reading risk status in Grade 1. Similarly, Teisl, Mazzocco, and Myers (2001) found that kindergarten teacher ratings of concern were associated with Grade 1 academic achievement and recommended that teacher ratings be used to determine which children receive screening measures to enhance the identification students at risk for a learning disability. Though these studies suggest a trend that early childhood educators are accurate in their identification of students at risk, large-scale, empirical study is needed to demonstrate the validity of screening measures based on early childhood educator judgment as programs begin to integrate such measures into their early childhood assessment systems.

Student demographics and teacher judgments

The influence of demographic variables on teacher judgments as young children start school is important and not widely studied. Ready and Wright (2011) presented a complex picture of the influence of context using hierarchical linear modeling with data from the ECLS-K. These authors found that teachers in higher achieving and higher SES classrooms tended to overestimate their students' abilities and teachers in lower achieving and lower SES contexts more often underestimate their students' abilities. Classroom context and teacher accuracy had stronger associations in lower socioeconomic status classrooms. Moreover, the analyses showed that the relationship between the context in which the teacher worked and the accuracy of their judgments was stronger than any association between the teachers' own social background and the accuracy of their judgments. This work is the lone large-scale analysis of teacher perceptions of students' skills in kindergarten at the time of this writing.

Other recent studies suggest contextual effects on teacher judgments. Children of lower socioeconomic status (SES) are more likely to be retained in kindergarten (Burkam, LoGerfo, Ready, & Lee, 2007) and placed into lower level ability groups, even after controlling for sociodemographic background and measured academic ability (Tach & Farkas, 2006). Beswick, Douglas Willms, and Sloat (2005) also found that kindergarten teacher ratings were influenced by gender, maternal education level, and behavior. Mashburn and Henry (2004) compared preschool and kindergarten teacher ratings of children's kindergarten readiness, academic skills, and communication skills with direct assessments of these skills. The authors found that student demographic characteristics were influential of teacher ratings. Specifically, boys and younger children had lower ratings from both preschool and kindergarten teachers. Family characteristics were associated with kindergarten teacher ratings. The authors found that children whose families received welfare had lower ratings than children whose families did not receive welfare. Also, African American children had higher teacher ratings than White children. More studies are needed.

The present study

Large-scale study of the predictive validity of kindergarten teacher ratings is needed. Current early childhood policy points to the importance of screening measures at kindergarten entry. Early identification of academic risk continues to be an issue for educators, and further study of the validity of teacher judgment at the start of formal schooling is needed. The research presented here is part of a larger validation of one state's inventory of students' skills at the start of their kindergarten year (Goldstein, Eastwood, & Behuniak, 2014; Goldstein & McCoach, 2011). The validation data allowed for a large-scale evaluation of association between teacher ratings at the start of kindergarten and achievement on a Grade 3 summative assessment in a multilevel context, answering a call from the National Research Council (2008) for such studies. Specifically, our research question was: Are teacher judgments of students' skills at the start of kindergarten associated with achievement on the Connecticut Mastery Test in Grade 3, accounting for the multilevel nature of the data?

Method

Participant characteristics

This study was an exploration of two assessments: the Kindergarten Entry Instrument (KEI; Connecticut State Department of Education, n.d.) and the Connecticut Mastery Test (CMT; Hendrawan & Wibowo, 2012). Both measures are described in the next section. The data set for the KEI was complete. There were many cases for which there was no match to Grade 3 CMT data in 2011. Reasons for the missing data include student mobility outside the state or to private school settings as well as student retention or early promotion. In addition, there were also some match field issues including flawed state student identification data and name changes (for the cases that had to be matched on name). All cases for which matches of KEI and CMT data could be confirmed were used in the analyses. Of the 37,048 students for whom there were KEI data, Grade 3 CMT data were available for 29,845 students. These students were nested in 571 elementary schools. On average, there were 65 students in each school. Teacher- and program-level data were unavailable for these analyses.

In 2007, demographic data were collected for the kindergarten students. Of the 37,048 students, 48% were girls, 36% were eligible for free or reduced-price lunch, 31% were minority students, 5% were English language learners, and 10% were eligible for special education services. The research team received available 2011 CMT data for all students who were kindergartners in 2007; demographic data were not reissued.

Instrumentation

Two state-level data sets from a single cohort of students were used in this analysis: Connecticut's KEI from 2007 and the

CMT for the same group of students in spring of their Grade 3 year, in 2011. Kindergarten teacher judgments of students' skills were defined by Connecticut's KEI. CMT data were the outcome variable. Both measures are reviewed subsequently.

KEI

In 2005 and 2006, the State of Connecticut passed legislation requiring the implementation of a statewide developmentally-appropriate assessment that "measures a child's level of preparedness for kindergarten." In response to this legislation the Connecticut State Department of Education developed the KEI, which was designed to provide a statewide snapshot of the skills students demonstrate, based on teachers' observations, at the beginning of the kindergarten year. The KEI is a rating form of six domains: Language, literacy, numeracy, physical/motor, creative/aesthetic, and personal/social. Each domain is defined by three to five specific indicators. The content of the inventory was selected to represent the most important skills that students need to demonstrate at the beginning of the kindergarten year based on Connecticut Preschool Curriculum Framework and State Curriculum Standards for language arts and mathematics in use at that time. A group of preschool and kindergarten teachers, representing urban and suburban districts, special education, and English language learners, reviewed the indicators and provided the Department of Education with their recommendations on the appropriateness of the indicators for a measure of this nature. The indicators that were selected for the Inventory are a result of the input from this committee. KEI results are currently collected for every public school kindergarten student through an electronic statewide data collection system. The instrument was first used in the fall of 2007 and is available for review in the Appendix.

By the end of October, each kindergarten teacher is required to classify the students in his/her class(es) into three performance levels by domain (i.e., each teacher assigns ratings to each student on each of six domains). Teachers are asked to assign a rating from 1 to 3 based on the consistency with which the student demonstrates the skills and the level of instructional support required for skill demonstration. A rating of 3 is used for students who consistently demonstrate the skills in the specified domain and require minimal instructional support. Students who receive a rating of 2 inconsistently demonstrate the skills in the specified domain and require some instructional support. A rating of 1 is used for students who demonstrate emerging skills in the specified domain and require a large degree of instructional support.

The joint American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education Standards for Educational and Psychological Testing (1999) guided validation of the KEI. The studies addressed two broad themes: The relationship of the KEI to other measures of academic achievement and the structure of the indicators used to define each domain. The dimensionality of the teacher ratings was also investigated through exploratory and confirmatory factor analyses (Goldstein & Behuniak, 2010; Goldstein & McCoach, 2011) and teacher focus groups (Goldstein & McCoach, 2011). Hierarchical linear modeling was used to demonstrate the association between teacher ratings on the KEI with kindergarten retention (Goldstein, Eastwood, & Behuniak, 2014) and proficiency on the state's summative assessment in Grade 3 (Goldstein, Behuniak, & Eastwood, 2011; Goldstein & Behuniak, 2011). Analyses of the association between the KEI and the state's summative assessment, the CMT, formed the foundation for this study.

The KEI was designed to produce a global snapshot of the state's students as they started kindergarten. Though the teachers render judgments across six domains, the data are unidimensional. For this sample, the data show high correlations among the ratings (see Table 1). Eigenvalues from an exploratory factor analysis of these ratings also suggest a unidimensional measure which accounts for 78.7% of the variance in the language rating, 84.7% of the variance in literature, 85.7% for numeracy, 73.1% for physical/motor, 77.8% for creative/aesthetic, and 70.7% for personal/social. Other model fit indices support good fit for the one-factor model (comparative fit index = .98; Tucker-Lewis index = .97; standardized root mean square residual = .06). Chi-square statistics and root mean square error of approximation were not considered as measures of model fit because of the exceptionally large sample size (n = 37,048). Cronbach's alpha (.91) was calculated as a measure of reliability for the single factor.

In this study, we created a single variable to represent teachers' global perceptions of students as they began kindergarten. The analyses focus on the association between this global perception and Grade 3 reading achievement. A variable named KEISum was created as the sum total of the teacher ratings across the six domains to represent a global measure of teacher judgment of a student's skills at the start of kindergarten. This sum score had a minimum of 6 (which indicates that a student received 1s on each of the six domains) and a maximum of 18 (which indicates that a student received 3s on each of the six domains). While the summary measure leads to some loss of information about strengths and weaknesses across the individual domains, we believe these summed ratings present a greater level of differentiation than the restricted range of the trichotomous rating format of the KEI as designed. At the student level, the mean KEISum was 13.04 ($SD = 3.65$) and the school mean KEISum was 13.00 ($SD = 2.01$). The school mean KEISum was significantly negatively correlated with the percentage of free lunch eligible students in the school ($r = -.48$), the percentage of minority students ($r = -.45$), the percentage of English language learners ($r = -.30$), and the percentage of students eligible for special education services ($r = -.15$).

KEISum was group-mean-centered to allow for comparisons within schools. In this study, the data were grouped by school. In group mean centering, the school's mean is subtracted from the score for each student in that school. As such, the transformed score captures a person's standing relative to his or her school (McCoach, 2010). Then the school aggregate KEI rating was added as a variable at the school level to ensure that the between school

Table 1. Spearman's rho correlations among domain ratings ($n = 37,048$).

Domain	Language	Literacy	Numeracy	Physical	Creative	Personal
Language	1.00					
Literacy	0.71	1.00				
Numeracy	0.70	0.79	1.00			
Physical	0.60	0.57	0.61	1.00		
Creative	0.62	0.56	0.59	0.70	1.00	
Personal	0.67	0.55	0.57	0.61	0.67	1.00

variability in KEI ratings was preserved and to facilitate comparisons among schools. The student-level predictors available to the researchers included gender, eligibility for free or reduced-price lunch (Lunch), minority status (Min), English language learner status (ELL), and eligibility for special education services (SWD). Several variables were re-coded to facilitate interpretations of the model estimates within regression-based analyses. Demographic data were recoded so that male = 1, female = 0; eligible for free or /reduced-price lunch = 1, not eligible = 0; English language learner = 1, non–English language learner = 0; and minority = 1, nonminority = 0. These variables were also aggregated by school to create the set of school-level predictors. The school-level demographic variables represented the proportion of students in the school who were either eligible for free/reduced-price lunch, underrepresented minorities, English language learners, or students with disabilities receiving special education services. All of these school level variables were grand mean centered in all analyses.

Connecticut Mastery Test

The CMT, first administered in 1984, is Connecticut's summative assessment of students' skills and knowledge in mathematics, reading, and writing in Grades 3–8, as well as science in Grades 5 and 8. The assessment is based on content that students at each grade level can reasonably be expected to have mastered and the assessment results are used to publicly account for statewide student achievement. Students receive a score from 100 to 400 for each tested content area. Scale scores are based on the raw scores (i.e., number of points earned). These raw scores are converted to scale scores to ensure accurate comparisons of student performance across different forms of the test by adjusting for slight differences in difficulty between test forms. Equating procedures are used to ensure that a given scale score represents the same level of performance within the same grade and content area regardless of the test form. Scale scores are used to define five performance levels for each content area: advanced, goal, proficient, basic and below basic.

CMT reading, mathematics, and writing data were used as separate outcome variables in this study. The CMT reading score is comprised of three elements: the degrees of reading power (DRP) and two test sessions of reading comprehension. The DRP is a holistic, multiple-choice measure of reading ability and measures a student's ability to understand nonfiction English prose on a graduated scale of reading difficulty. The reading comprehension test sessions consist of narrative and informational passages on a variety of topics. Students respond to multiple-choice and open-ended questions after reading each passage. The DRP and the reading comprehension tests are weighted equally in deriving the overall CMT reading score. The mathematics test contains dichotomously-scored multiple-choice items, grid-in response items, and open-ended items. The writing assessment includes multiple choice items to address editing and revising skills and single prompt direct assessment of writing. Descriptive data for the sample are included in Table 2. The KEISum score for students at each performance level are included in Table 3.

The state has an extensive body of research to support the validity of the CMT (Hendrawan & Wibowo, 2012). Thousands of classroom educators were surveyed about the appropriateness of the assessment content in 1984, 1985, and 2000. There were expert and educator reviews of the content and test items for each revision of the assessment. CMT scores were correlated with the Metropolitan Achievement Test subtests in total language, reading comprehension, mathematics concepts, and mathematics procedures. Assessment staff worked with Connecticut educators to establish score boundaries and define score-point examples and training sets. These materials were used to train item readers over several days. Readers had to qualify for assessment scoring by matching several sets of student responses prescored by Connecticut educators. Materials scored by Connecticut educators were also integrated in the assessment scoring process to check for reliability. In addition, 100% of the writing prompts and 20% of the short answer and extended response items in mathematics and reading comprehension were read twice.

Table 2. CMT 2011 data distribution.

		Scale Scores		Performance levels				
Domain	*n*	*M*	*SD*	Below basic	Basic	Proficient	Goal	Advanced
Reading	29,845	243.40	40.90	15%	9%	15%	42%	19%
Mathematics	30,126	261.68	50.53	7%	7%	20%	34%	31%
Writing	30,623	255.05	47.61	7%	10%	19%	41%	22%

Note. CMT data were missing for 18.7%, 19.4%, and 17.3% of the KEI data cases, for mathematics, reading, and writing, respectively.

Analyses

We studied the association between kindergarten teachers' perceptions of readiness and Grade 3 reading achievement with hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002). Hierarchical linear modeling exploits the information contained in cluster samples to explain both the between- and within-cluster variability of an outcome variable of interest. Using multilevel models, predictors at both the student level (level 1), and the school level (level 2) explain the variance in the dependent variable. Further, the relationship between an independent variable and the dependent variable can randomly vary across clusters. If the relationship between a student level predictor and the dependent variable varies across schools, then we can try to explain the variability in this relationship using school-level predictors. Thus, hierarchical linear modeling allows us to simultaneously model the impact of both individual and institutional variables on the dependent variable of interest, as well as to model the cross-level interactions between higher and lower level variables on the outcome of interest (McCoach, 2010).

Before the primary analyses, an empty model was estimated to justify the use of HLM with these data estimate the degree of between school variance in the dependent variable. The

Table 3. Mean KEISum ratings for CMT reading.

CMT performance level	Reading	Mathematics	Writing
Below basic	11.35	10.85	10.64
Basic	12.47	11.76	11.83
Proficient	12.87	12.59	12.63
Goal	14.09	13.83	13.88
Advanced	15.32	14.88	15.13

outcome of interest was CMT scale score for each subject (CMT). The empty model was:

$$CMT_{ij} = \beta_{0j} + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}.$$

The empty model allows for the calculation of the intraclass correlation coefficient (ICC), a measure of between-school variability. The ICC is calculated by partitioning the total variance into within-group and between-group variability (ICC = $\mathrm{Var_{u0}}/(\mathrm{Var_{u0}} + \mathrm{Var_r})$. For these data, 19% of the total variance in the CMT reading scores, 21% of the total variance in CMT mathematics scores, and 18% of the total variance in CMT writing scores is explained by the schools. This level of between-school variability warrants the use of HLM for further analysis of the data.

A contextual model was estimated to answer the research question. The contextual model allows for the estimation of the predictive utility of KEISum over and above school- and student-level covariates. Moreover, this model posited a differential influence of demographics on a relationship between KEISum and CMT. For example, there may be a stronger (or weaker) association between KEISum and CMT for schools with a higher proportion of students who receive free or reduced-price lunch. The slopes of the level 1 independent variables were estimated using random effects to account for potential between-school differences in the relationships among these variables. The model was

$$\begin{aligned}
CMT_{ij} &= \beta_{0j} + \beta_{1j}(\mathrm{KEISum}_{ij} - \mathrm{KEISum}_{\cdot j}) + \beta_{2j}(\mathrm{Lunch}_{ij}) \\
&\quad + \beta_{3j}(\mathrm{Min}_{ij}) + \beta_{4j}(\mathrm{ELL}_{ij}) + \beta_{5j}(\mathrm{SWD}_{ij}) \\
&\quad + \beta_{6j}(\mathrm{Gender}) + r_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}(\mathrm{MKEISum}_j - \mathrm{MKEISum.}) \\
&\quad + \gamma_{02}(\mathrm{Lunch}_j - \mathrm{Lunch.}) + \gamma_{03}(\mathrm{Min}_j - \mathrm{Min.}) \\
&\quad + \gamma_{04}(\mathrm{ELL}_j - \mathrm{ELL.}) + \gamma_{05}(\mathrm{SWD}_j - \mathrm{SWD.}) + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}(\mathrm{MKEISum}_j - \mathrm{MKEISum.}) \\
&\quad + \gamma_{12}(\mathrm{Lunch}_j - \mathrm{Lunch.}) + \gamma_{13}(\mathrm{Min}_j - \mathrm{Min.}) \\
&\quad + \gamma_{14}(\mathrm{ELL}_j - \mathrm{ELL.}) + \gamma_{15}(\mathrm{SWD}_j - \mathrm{SWD.}) + u_{1j} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}(\mathrm{Lunch}_j - \mathrm{Lunch.}) + u_{2j} \\
\beta_{3j} &= \gamma_{30} + \gamma_{31}(\mathrm{Min}_j - \mathrm{Min.}) + u_{3j} \\
\beta_{4j} &= \gamma_{40} + \gamma_{41}(\mathrm{ELL}_j - \mathrm{ELL.}) + u_{4j} \\
\beta_{5j} &= \gamma_{50} + \gamma_{51}(\mathrm{SWD}_j - \mathrm{SWD.}) + u_{5j} \\
\beta_{6j} &= \gamma_{60} + u_{6j}
\end{aligned}$$

Student-level predictors were included at level 1. The KEISum scores were group-mean centered at level 1. The aggregated student-level predictors were included as school-level predictors, which were grand-mean centered. The slopes of the level 1 independent variables were treated as random and the intercept was allowed to vary randomly across schools. Cross-level interactions were also included. Cross-level interactions model the effect of level 2 variables on the relationship between level 1 variables and the outcome variable. Separate models were estimated for each of the three subjects: CMT reading, CMT mathematics, and CMT writing.

Results

Model estimates

The analysis was designed to identify an association between teachers' global perceptions of students as they begin kindergarten (KEISum) and Grade 3 achievement in reading, mathematics, and writing, accounting for the influence of student characteristics and school characteristics. The results of the models are included in Table 4. Note that for each subject, Model A does not include KEISum at the student level or school level; the model is included for comparison purposes. The other level 1 predictors and the cross-level interactions were included to control for those variables in the predictions.

The first observation from the estimates in Model B is that KEISum is a significant predictor of Grade 3 reading achievement ($\gamma_{10} = 3.93$), Grade 3 mathematics achievement ($\gamma_{10} = 5.14$), and Grade 3 writing achievement ($\gamma_{10} = 3.94$), even after controlling for student- and school-level demographics. For students who attended schools with average percentages of free lunch eligible students, minority students, English language learners, and students with disabilities, a single point increase on their KEISum rating scale at the start of kindergarten is associated with nearly 4-point increase in Grade 3 reading/writing achievement and a 5-point increase in mathematics achievement. This association translates into a Cohen's *d* effect size of approximately .10 standard deviation units for each point increase on the KEISum rating. The model includes parameter estimates for the effect of the level 2 variables on the slope (γ_{11} through γ_{15}). The percentage of free lunch eligible students has an interesting impact on the association between kindergarten teachers' global perceptions of students and Grade 3 achievement across the three subjects. Specifically, there is a weaker relationship between KEISum and CMT in schools with a larger percentage of free-lunch-eligible students ($\gamma_{12} = -1.91$ for reading; $\gamma_{12} = -2.27$ for mathematics; $\gamma_{12} = -3.22$ for writing). In other words, the positive association between kindergarten teacher judgments and achievement in Grade 3 is weaker in schools with larger percentages of free lunch students and stronger in schools with lower than average percentages of free lunch students. The relationship between kindergarten teacher judgments and Grade 3 achievement across subjects is not affected by the percentage of minority students (γ_{13}), the percentage of English language learners (γ_{14}), or the percentage of students with disabilities in a given school (γ_{15}) after controlling for other variables in the model.

For mathematics achievement only, the school average KEISum influences the association between kindergarten teachers' global perceptions of students and Grade 3 achievement. There is a stronger relationship between KEISum and CMT-Mathematics in schools with a higher overall KEISum ratings ($\gamma_{11} = 0.19$). The positive association between kindergarten teacher judgments and Grade 3 mathematics achievement is stronger in schools with higher than average school mean KEISum scores and weaker in schools with lower school mean KEISum scores.

Also notable, school average KEISum rating helps to predict Grade 3 reading achievement ($\gamma_{01} = 0.53$) even after accounting for student- and school-level demographics.

Table 4. Fixed and variance components for the predicted CMT scores by subject.

		Reading		Mathematics		Writing	
		A	B	A	B	A	B
Fixed components							
For intercept							
Intercept	γ_{00}	254.78	250.84	271.70	266.49	275.12	271.27
MKEISum	γ_{01}		0.53		*(0.40)*		*(0.05)*
L2_Lunch	γ_{02}	−29.29	−29.91	−37.85	−39.46	−32.21	−34.42
L2_Minority	γ_{03}	*(6.95)*	*(6.79)*	9.52	*(8.74)*	*(7.10)*	*(6.67)*
L2_ELL	γ_{04}	10.84	*(6.12)*	19.47	*(12.50)*	*(−0.10)*	*(−5.40)*
L2_SWD	γ_{05}	−15.64	−18.37	*(−15.35)*	*(−20.12)*	−24.99	*(−27.68)*
For slopes							
KEISum	γ_{10}		3.93		5.14		3.94
MKEISum	γ_{11}		*(0.07)*		0.19		*(0.09)*
L2_Lunch	γ_{12}		−1.91		−2.27		−3.22
L2_Min	γ_{13}		*(0.29)*		*(0.45)*		*(0.77)*
L2_ELL	γ_{14}		*(−1.61)*		*(−1.23)*		*(−1.68)*
L2_SWD	γ_{15}		*(1.83)*		*(3.05)*		*(2.26)*
L1_ Lunch	γ_{20}	−14.99	−11.02	−16.44	−11.22	−15.96	−11.93
L2_Lunch	γ_{21}	8.17	*(4.10)*	5.94	*(0.98)*	10.03	5.18
L1_ Minority	γ_{30}	−11.81	−9.64	−17.24	−14.39	−10.19	−8.00
L2_Minority	γ_{31}	*(−4.85)*	−5.95	*(−3.39)*	*(−4.60)*	*(−2.26)*	*(−2.97)*
L1_ELL	γ_{40}	−30.06	−22.09	−24.27	−13.66	−21.02	−13.10
L2_ELL	γ_{41}	22.82	*(11.90)*	*(23.89)*	*(9.83)*	*(5.23)*	*(−9.40)*
L1_SWD	γ_{50}	−41.38	−31.99	−41.59	−29.35	−40.96	−31.31
L2_SWD	γ_{51}	32.43	*(19.78)*	*(22.70)*	*(8.94)*	30.38	*(17.20)*
L1_Gender	γ_{60}	−1.89	0.90	5.04	8.65	−18.05	−15.24
Variance components							
Level 2							
Intercept	u_0	72.29	72.62	146.00	148.60	164.19	160.85
Slope							
KEISum	u_1		1.80		3.02		1.15
L1_Lunch	u_2	41.93	*(32.68)*	*(54.68)*	47.98	*(18.34)*	*(12.94)*
L1_Minority	u_3	60.27	41.17	51.23	*(31.86)*	*(27.67)*	*(19.43)*
L1_ELL	u_4	*(56.37)*	87.79	*(81.71)*	*(101.70)*	*(28.19)*	*(65.44)*
L1_SPED	u_5	191.21	177.54	162.88	139.55	*(80.50)*	*(71.73)*
L1_Gender	u_6	*(3.56)*	*(4.83)*	10.89	15.81	*(19.70)*	*(17.27)*
Level 1	r	1128.77	1014.65	1782.82	*1589.91*	1474.01	*1354.26*
Deviance		290064.68	287240.39	303482.90	300453.53	297760.13	295469.69

Note. Nonsignificant parameters are noted in italics parenthetically.

Schools with higher mean KEISum ratings also tend to have higher Grade 3 reading achievement, even after controlling for the school demographic variables. Kindergarten teacher ratings are associated with Grade 3 reading achievement at the school level as well. This pattern is not evident for Mathematics or Writing.

Comparisons of the Table 4 variance components across Models A and B at level 1 (r) help to quantify the reduction in variance gained by the inclusion of KEI information in the prediction of CMT reading scale scores in Grade 3. This pattern was not significant for Mathematics and Writing. Comparison of Model A to Model B indicates that the inclusion of KEISum information explains approximately 10% of the within-school variability across the subjects, above and beyond student demographics (10.1% for reading; 10.8% for mathematics; 8.1% for writing). Teachers' global perceptions of students as they begin kindergarten are predictive of Grade 3 achievement across subjects.

Model predictions

We created predicted CMT reading scale from the HLM to offer context to these model estimates (see Table 5). At the student level, four circumstances were considered for a hypothetical female student who is neither an English language learner nor receiving special education services:

- Minority and eligible for free or reduced-price lunch;
- Nonminority but eligible for free or reduced-price lunch;
- Minority but not eligible for free or reduced-price; and
- Neither minority nor eligible for free or reduced-price lunch.

Predicted scale scores for male students with this demographic profile would be 0.88 points lower than the values contained within the table. The school-mean-KEISum was held constant at the grand mean for the predictions. Two hypothetical schools were considered. The first school was both high minority (MIN = 68%) and high free lunch eligible students (LUNCH = 72%) and the other was had a low proportion of minority students (MIN = 3%) and free lunch eligible students (LUNCH = 5%). These schools represent proportions of one standard deviation above and below the mean, respectively. For clarity, Table 4 is shaded to represent the cut scores for each proficiency band for the CMT results (below basic = 100–201, basic = 202–216, proficient = 217–234, goal = 235–278, and advanced = 279–400).

There are two points to note in the table. First, the students with low KEISum ratings (KEISum = 6–8) in a hypothetical high minority, high poverty school tend to have higher

Table 5. Predicted CMT reading scale scores for non-special education, non-English language learner girls in two schools.

	High-minority, high-poverty school student demographics				Low-minority, low-poverty school student demographics			
KEI	Free lunch/ minority	No free lunch/minority	Free lunch/ not minority	No free lunch/not minority	Free lunch/ minority	No free lunch/minority	Free lunch/ not minority	No free lunch/not minority
6	134	144	146	155	123	135	131	143
7	147	156	158	168	139	152	147	160
8	159	169	170	180	156	168	164	176
9	171	181	183	193	172	185	180	192
10	184	193	195	205	189	201	197	209
11	196	206	207	217	205	217	213	225
12	208	218	220	230	222	234	229	242
13	221	230	232	242	238	250	246	258
14	233	243	245	254	254	267	262	275
15	245	255	257	267	271	283	279	291
16	258	268	269	279	287	300	295	308
17	270	280	282	291	304	316	312	324
18	283	292	294	304	320	333	328	340

predicted scores than students with the same KEISum rating in a low minority, low poverty school. Alternatively, students with higher KEISum ratings (KEI = 16–18) tend to have much higher predicted scores in a low-minority, low-poverty school than a high-minority, high-poverty school, regardless of their individual demographics. Second, there is overlap in the predicted score distributions when students in the middle of the distribution are compared between schools. Note, for example that a nonminority student who is not eligible for free/reduced-price lunch in a high-needs school with a KEISum rating of 12 has a predicted CMT reading score of 230, while a minority student who is eligible for free/reduced-price lunch in a low-needs school has a predicted score of 222. Finally, predicted CMT reading scale scores for students with lower KEISum ratings (6–9) remained at below basic in both demographic scenarios.

Discussion

HLM analyses with a large-scale data set demonstrated that teachers' perceptions of their students at the start of kindergarten was related to achievement in Grade 3, even after accounting for student- and school-level demographics. Additionally, the analyses showed that this association was weaker in schools with greater numbers of free lunch eligible students.

Predictive validity of kindergarten teacher judgments

These analyses tell the story of one state system in which teacher ratings of students' skills from October of their kindergarten year are associated with performance on the state's summative assessment nearly four years in the future. This association exists across schools even after accounting for differences in student and school demographics. Though several studies question the validity of teacher predictions (Bates & Nettlebeck, 2001; Begeny et al., 2008; Begeny, Krouse, Brown, & Mann, 2011; Hamilton & Shin, 2003), our results suggest a different pattern. This discrepancy may be because teachers in this study were not explicitly asked to define or predict achievement. Rather, teachers were asked for broad ratings of their students' skills, and those ratings were used to define a global measure of teacher judgments at the start of kindergarten. Kindergarten teachers were not asked to categorize or characterize their students' future performance. Further, those often-cited studies were based on very small samples of students. Similar to earlier studies, (Flynn & Rahbar, 1998; Speece et al., 2011; Teisl et al., 2001), this study connected kindergarten teacher judgments to later academic difficulties. Lower ratings by kindergarten teachers were associated with lower scores on the state's summative assessment in Grade 3. The model predictions showed that kindergarten students with the lowest ratings from their kindergarten teachers had Grade 3 reading achievement scores that were below basic. Alternatively, students with the highest ratings had the highest predicted Grade 3 scores. Similarly, Halle et al. (2012) found that children who entered with stronger school readiness skills tended to maintain their advantage over time, while children who entered with lower school readiness skills tended to maintain their relative disadvantage over time.

Demographic variable associations

This study also showed that the association between teacher judgments of their students as they begin kindergarten and Grade 3 achievement may be moderated by the percentage of free lunch eligible students, an indicator of poverty. This association was weaker for schools with greater proportions of free lunch eligible students. Score predictions helped to illustrate these complex relationships for hypothetical students in different types of schools. As an example, the illustration shows two minority students who are eligible for free or reduced-price lunch who both have a KEISum of 12. The student in a high minority, high poverty school has a predicted summative assessment score in the Basic range while the student in the low minority, low poverty school has a predicted score in the proficient range. Alternatively, we can consider two nonminority students who are not eligible for free or reduced-price lunch with a KEISum rating of 12. The predicted score for that child in a high-minority, high-poverty school is proficient, while the predicted score for a similar child in a low-minority, low-poverty school is goal. Ready and Wright (2011) also found that the relationship between kindergarten teacher judgment and a standardized literacy assessment was affected by the socioeconomic context of the classroom. Their hierarchical linear modeling analysis of a national data set showed that teachers in lower SES contexts underestimated their students' abilities in kindergarten. Though our analysis showed a relationship between kindergarten teacher judgments

and Grade 3 assessment data, we also found a moderating influence of socioeconomic context.

The complex influence of classroom context is intriguing, and we can speculate as to why this pattern might exist statewide. It is possible that teacher ratings at the start of kindergarten reflect student experience: students in schools with higher percentages of free lunch eligible students may have less knowledge at the start of school but may grow at faster rates as compared to students in schools with fewer numbers of these students. Alternatively, children who do not live in poverty may have richer early childhood experiences to prepare them for the start of kindergarten. An alternative hypothesis is that teachers in schools with a greater number of students eligible for free or reduced-price lunch are less rigorous in their efforts to complete the KEI at the start of kindergarten. Perhaps these schools are better able to overcome the readiness gap in the early years of schooling. This finding is intriguing, and further study of the variability in this gap is needed.

Implications

This study is especially relevant in light of a new focus on early childhood from the federal government. Teacher judgments of their students' academic abilities are the foundation of the educational process (Eckert & Arbolino, 2005; Salvia & Ysseldyke, 2004; Shapiro & Kratochwill, 2000), and these judgments are especially important at the start of kindergarten. Speece et al. (2011) noted that valid and efficient universal screening measures that are quick and easy to administer are needed in the field. This study suggests there is merit to a simple indicator of concern or a single evaluation of readiness in predicting later academic outcomes. Researchers and psychometricians working toward the development of kindergarten entry assessments should consider the efficacy of a single question about concern for a student's well-being in addition to complex systems of multidimensional developmental screeners, formative assessments, and summative assessments. While global teacher judgments may be accurate, there is evidence that these judgments are shaped by context. Continued efforts to reduce racial and socioeconomic isolation may help limit these associations with teacher ratings. This finding serves as a caution for researchers and policy makers alike. Preservice training and in-service professional development are critical to the success of early childhood assessment systems. Such efforts can focus on assessment fidelity issues including consistent use of assessment rubrics, reduction of bias in student observations, and appropriate assessment scenarios for young children at the start of formal schooling.

Limitations

Although this research sheds light on an important issue, this study has several limitations. First, the analyses that support this study call into question the use of the KEI as it was designed. Though the instrument was written to represent six domains, there is evidence that the ratings function as a unidimensional measure of students' skills. Though the CMT has been used for nearly 30 years in Connecticut, comprehensive psychometric data to support the quality of the assessment is not publicly available. In addition, teacher-level data would have allowed for more refined conclusions about the use of teacher rating scales. Information on specific programs in place at different schools and more detailed demographic information about the schools would have also contributed to the analyses. Also, the results represent data from one cohort of students and one state's unique assessments. While these issues limits the generalizability of the findings, we believe the results of this study are instructive to the field as early childhood assessment systems begin to expand across the country.

Funding

This research was supported in part from a contract with the Connecticut State Department of Education.

References

Alexander, K. L., & Entwisle, D. R. (1988). Achievement in the first 2 years of school: Patterns and processes. *Monographs of the Society for Research in Child Development*, *53*(2), 67–87.

Alexander, K. L., & Entwisle, D. R. (1996). Schools and children at risk. In A. Booth & J. F. Dunn (Eds.), *Family school links: How do they affect educational outcomes?* (pp. 67–87). Mahwah, NJ: Erlbaum.

Barnett, D. W., Macmann, G. M., Carey, K. T. (1992). Early intervention and the assessment of developmental skills: Challenges and directions. *Topics in Early Childhood Special Education*, *12*, 21–43.

Bates, C., & Nettelbeck, T. (2001) Primary school teachers' judgements of reading achievement. *Educational Psychology: An International Journal of Experimental Educational Psychology*, *21*, 177–187.

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* (1), 43–55.

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

Berkner, L. K., & Chavez, L. (1997). Access to postsecondary education for the 1992 high school graduates. Statistical Analysis Report, NCES 98-105. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Beswick, J. F., Douglas Willms, J., & Sloat, E. A. (2005). A comparative study of teacher ratings of emergent literacy skills and student performance on a standardized measure. *Education*, *26*, 116–139.

Behuniak, P., & Goldstein, J. (2011). *A study of the relationship between Connecticut's Kindergarten Entrance Inventory and the Connecticut Mastery Test*. Hartford, CT: Connecticut State Department of Education.

Bowman, B., Donovan, M. S., & Burns, M. S. (Eds.). (2001). *Eager to learn: Educating our preschoolers*. Washington, DC: National Academy Press.

Braswell, J. S., Lutkus, A. D., Grigg, W. S., Santapau, S. L., Tay-Lim, B. S.-H., & Johnson, M. S. (2001). The nation's report card: Mathematics 2000 (NCES 2001-517). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Burkam, D. T., LoGerfo, L, Ready, D., & Lee, V. E. (2007). The differential effects of repeating kindergarten: Cognitive boost for some, cognitive bust for others. *Journal of Education for Students Placed At Risk*, *12*, 103–136.

Case, R., & Griffin, S. (1990). Child cognitive development: the role of central conceptual structures in the development of scientific and social thought. In C. A. Hauert (Ed.), *Developmental psychology: Cognitive, perceptuo-motor, and neuropsychological perspectives* (pp. 193–230). Amsterdam, the Netherlands: Elsevier Science.

Connecticut State Department of Education (n.d.). *Fall kindergarten entrance inventory*. Available online at: http://www.csde.state.ct.us/public/csde/cedar/assessment/kindergarten/fall.htm

Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly*, *13*, 8–24.

Denton, K., & West, J. (2002). Children's reading and mathematics achievement in kindergarten and first grade (NCES 2002-125). Washington, DC: National Center for Education Statistics.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., … Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*, 1428–1446.

Duncan, G. J., & Magnuson, K. A. (2005). Can family socioeconomic resources account for racial and ethnic test score gaps? *Future of Children*, *15*, 35–54.

Early Childhood Education Assessment-State Collaborative on Assessment and Student Standards. (2011). *Moving forward with kindergarten readiness efforts: A position paper of the of the early childhood education assessment-state collaborative on assessment and student standards*. Washington, DC: Council of Chief State School Officers.

Eckert, T. L., & Arbolino, L. A. (2005). The role of teacher perspectives in diagnostic and program evaluation decision-making. In R. Brown-Chidsey (Ed.), *Beyond labels: Noncategorical individualized assessment methods* (pp. 65–81). New York, NY: Guilford Press.

Entwisle, D. R., & Alexander, K. L. (1993). Entry into schools: The beginning school transition and educational stratification in the United States. *Annual Review in Sociology*, *19*, 401–423.

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

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher -based judgments of students reading with differing achievement levels. *The Journal of Educational Research*, *102*, 453–462.

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

Francis, D. J., Fletcher, J. M., Shaywitz, B. A., Shaywitz, S. E., & Rourke, B. P. (1996). Defining learning and language disabilities: Conceptual and psychometric issues with the use of IQ tests. *Language, Speech, and Hearing Services in Schools*, *27*, 132–143.

Fryer, R. G., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics*, *86*, 447–464.

Fryer, R. G., & Levitt, S. D. (2006). The black-white test score gap through third grade. *American Law and Economics Review*, *8*, 249–281.

Goldstein, J., & Behuniak, P. (2010). *A study of the structure of the Hartford indicator data*. Hartford, CT: Connecticut State Department of Education.

Goldstein, J., & Behuniak, P. (2011). *Developing a framework to define students' skills at kindergarten entry: Focus group summary report*. Hartford, CT: Connecticut State Department of Education.

Goldstein, J., Behuniak, P., & Eastwood, M. (2011). *Understanding patterns of achievement for young learners in Connecticut*. Hartford, CT: Connecticut State Department of Education.

Goldstein, J., Eastwood, M., & Behuniak, P. (2014). Can teacher ratings of students' skills at kindergarten entry predict kindergarten retention? A quantitative analysis. *Journal of Educational Research*, *107*, 217–229.

Goldstein, J., & McCoach, D. B. (2011). The starting line: Developing a structure for teacher ratings of students' skills at kindergarten entry. *Early Childhood Research and Practice 13*(2). Available at: http://ecrp.uiuc.edu/v13n2/goldstein.html

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system manual*. Circle Pines, MN: American Guidance Service.

Hair, E., Halle, T., Terry-Humen, E., Lavelle, B., & Calkins, J. (2006). Children's school readiness in the ECLS-K: Predictions to academic, health, and social outcomes in first grade. *Early Childhood Research Quarterly*, *21*, 431–454.

Halle, T. G., Hair, E. C., Burchinal, M., Anderson, R., & Zaslow, M. (2012). In *the running for successful outcomes: Exploring the evidence for thresholds of school readiness*. Washington, DC: U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation.

Hamilton, C., & Shinn, M. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review*, *32*, 228–240.

Hendrawan, I., & Wibowo, A. (2012). *The Connecticut Mastery Test: Technical report*. Durham, NC: Measurement Incorporated.

Haskins, R., & Rouse, C. (2005). Closing achievement gaps. *The future of children spring policy brief*. Princeton, NJ: Princeton University and Brookings Institution.

Jordan, N. C, Huttenlocher, J., & Levine, S. C. (1992). Differential calculation abilities in young children from middle- and low-income families. *Developmental Psychology*, *28*, 644–653.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. Journal of Educational Psychology, 80(4), 437–447.

Kaufman, A. S., & Kaufman, N. L. (1985). *Kaufman test of educational achievement- brief form*. Circle Pines, MN: American Guidance Service.

Lidz, C. (1983). Issues in assessing preschool children. In K. Paget & B. Bracken (Eds.), *The psychoeducational assessment of preschool children*. New York, NY: Grune and Stratton.

Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. W. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. *Economics of Education Review*, *26*, 52–66.

Martin, S. D., & Shapiro, E. S. (2011). Examining the accuracy of teachers' judments of DIBELS performance. *Psychology in the Schools*, *48*, 343–356.

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

McCoach, D. B. (2010). Hierarchical linear modeling. In G. O. Hancock & R. O. Mueller (Eds.) *The reviewer's guide to quantitative methods in the social sciences*. (pp 123–140). New York, NY: Routledge.

Mulligan, G. M., Hastedt, S., & Carroll, J. C. (2012). First-time kindergartners in 2010–11: First findings from the kindergarten rounds of the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) (NCES 2012-049). Washington, DC: National Center for Education Statistics.

National Research Council. (2008). *Early childhood Assessment: Why, what, and how*. Washington, DC: National Academies Press.

Neuman, S. B., & D. Dickinson, (Eds.). (2001). *The handbook of early literacy research*. New York, NY: Guilford.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal*, *48*, 335–360.

Reardon, S. F. (2003). *Sources of educational inequality: The growth of racial/ethnic and socioeconomic test score gaps in kindergarten and first grade* (Working Paper 03-05R). University Park, PA: The Pennsylvania State University, Population Research Institute.

Rumberger, R. W., & Arellano, B. (2004). *Understanding and addressing the Latino achievement gap in California*. (Working paper 2004-01). Berkeley, CA: UC Latino Policy Institute.

Sabol, T. J., & Pianta, R. C. (2012). Patterns of school readiness forecast achievement and socioemotional development at the end of elementary school. *Child Development, 83*, 282–299.

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment* (9th ed.). New York, NY: Houghton Mifflin.

Saluja, G., Scott-Little, C., & Clifford, R. M. (2000). Readiness for school: A survey of state policies and definitions. *Early Childhood Research & Practice*, *2*, 1–19.

Scott-Little, C., Kagan, S. L., & Clifford, R. M. (Eds.), (2003). *Assessing the state of state assessments: Perspectives on assessing young children*. Tallahassee, FL: SERVE.

Shapiro, E. S., & Kratochwill, T. R. (2000). Introduction: Conducting a multidimensional behavioral assessment. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Conducting school-based assessments of child and adolescent behavior* (pp. 1–20). New York, NY: Guilford Press.

Shaywitz, S. E., Fletcher, J. M., Holahan, J. M., Schneider, A. E., Marchione, K. E., Stuebing, K. K., … Shaywitz, B. A. (1999). Persistence of dyslexia: The Connecticut Longitudinal Study at adolescence. *Pediatrics*, *104*, 1351–1359.

Speece, D., Schatschneider, C., Silverman, R., Case, L., Jacobs, D., & Cooper, D. (2011). Early identification of reading problems in a response to intervention framework. *Elementary School Journal*, *111*, 585–607.

Stedron, J. M., & Berger, A. (2010). *NCSL Technical Report: State approaches to school readiness assessment*. Washington, DC: National Conference of State Legislatures.

Tach, L., & Farkas, G. (2006). Learning-related behaviors, cognitive skills, and ability grouping when schooling begins. *Social Science Research*, *35*, 1048–1079.

Teisl, J. T., Mazzocco, M. M., & Myers, G. F. (2001). The utility of kindergarten teacher ratings for predicting low academic achievement in first grade. *Journal of Learning Disabilities*, *34*, 286–293.

Tramontana, M. G., Hooper, S. R., & Selzer, S. C. (1988). Research on the preschool prediction of later academic achievement: A review. *Developmental Review*, *8*, 89–146.

U.S. Department of Education. (n.d.). *Definitions*. Retrieved from http://www.ed.gov/early-learning/elc-draft-

U.S. Department of Education. (2011, August 22). *Race to the Top - Early Learning Challenge application for initial funding*. CFDA Number: 84.412. Retrieved from http://www2.ed.gov/programs/racetothetop-earlylearningchallenge/2011-412.doc West, Denton, & Reaney, 2001

West, J., Denton, K., & Reaney, L. (2001). *The kindergarten year* (NCES 2001-023). Washington, DC: National Center for Education Statistics.

Zill, N., Collins, M., West, J., & Hausken, E. G. (1995). *Approaching kindergarten: A look at preschoolers in the United States*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Appendix

Fall Kindergarten Entrance Inventory

The following Performance Level (PL) Literals describe the characteristics of a typical student at each performance level. These will be used to rate each student on each of the six domains.

Performance Level 1: Students at this level demonstrate emerging skills in the specified domain and require a large degree of instructional support.

Performance Level 2: Students at this level inconsistently demonstrate the skills in the specified domain and require some instructional support.

Performance Level 3: Students at this level consistently demonstrate the skills in the specified domain and require minimal instructional support.

Directions: The indicators listed below each domain are examples of the skills a student should be able to demonstrate at the beginning of the kindergarten year; however, these are not the only skills to be considered. Rate each student in your class on each of the six domains. Use the Performance Levels (PL) above and all available and pertinent information when rating a student.

Language skills

At what level does the student:

- Participate in conversations
- Retell information from a story read to him/her
- Follow simple two-step verbal directions
- Speak using sentences of at least 5 words
- Communicate feelings and needs
- Listen attentively to a speaker

Literacy skills

At what level does the student:

- Hold a book and turn pages from the front to the back
- Understand that print conveys meaning
- Explore books independently
- Recognize printed letters, especially in their name and familiar printed words
- Match/connect letters and sounds
- Identify some initial sounds
- Demonstrate emergent writing

Numeracy skills

At what level does the student:

- Count to 10
- Demonstrate one-to-one correspondence while counting (e.g., touches objects as he/she counts)
- Measure objects using a variety of everyday items
- Identify simple shapes such as circles, squares, rectangles, and triangles
- Identify patterns
- Sort and group objects by size, shape, function (use), or other attributes
- Understand sequence of events (e.g., before, after, yesterday, today, or tomorrow)

Physical/motor skills

At what level does the student:

- Run, jump, or balance
- Kick or throw a ball, climb stairs or dance
- Write or draw using writing instruments (e.g., markers, chalk, pencils, etc.)
- Perform tasks, such as completing puzzles, stringing beads, or cutting with scissors

Creative/aesthetic skills

At what level does the student:

- Draw, paint, sculpt, or build to represent experiences
- Participate in pretend play
- Enjoy or participate in musical experiences (e.g., singing, clapping, drumming, or dancing)

Personal/social skills

At what level does the student:

- Engage in self-selected activities
- Interact with peers to play or work cooperatively
- Use words to express own feelings or to identify conflicts
- Seek peer or adult help to resolve a conflict
- Follow classroom routines