Article

Journal of Psychoeducational Assessment
2017, Vol. 35(4) 410–423
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0734282916639195
journals.sagepub.com/home/jpa
SAGE

Factors Associated With Accuracy in Prekindergarten Teacher Ratings of Students' Mathematics Skills

Emily C. Furnari[1], Jessica Whittaker[1], Mable Kinzie[1], and Jamie DeCoster[1]

Abstract

The No Child Left Behind Act requires that 95% of students in all public elementary and secondary schools are assessed in mathematics. Unfortunately, direct assessments of young students can be timely, costly, and challenging to administer. Therefore, policy makers have looked to indirect forms of assessment, such as teachers' ratings of student skills, as a substitute. However, prekindergarten teachers' ratings of students' mathematical knowledge and skills are only correlated with direct assessments at the .50 level. Little is known about factors that influence accuracy in teacher ratings. In this study, we examine the influence of student and teacher characteristics on prekindergarten teachers' ratings of students' mathematical skills, controlling for direct assessment of these skills. Results indicate that students' race/ethnicity and social competency, as well as teachers' self-efficacy, are significantly related to prekindergarten teachers' ratings of students' mathematical skills.

Keywords

teachers, prekindergarten, validity, mathematics

Students' early mathematical skills are foundational for later skill development and are among the best predictors of later school success (Bowman, Donovan, & Burns, 2001; Duncan et al., 2007). According to recent reports of national education progress, all states, with the exception of New Hampshire, have recognized the importance of early mathematics by including it in their early learning standards, with 22 of those states requiring school readiness assessments for all students entering kindergarten (NCES, 2013). Between 2011 and 2015, 43 states were approved for Elementary and Secondary Education Act Flexibility, allowing them to design their own plans to monitor and improve educational outcomes for all students ("NCLB ESEA Flexibility," 2015). Due to the costs of administering individual direct assessments and the rising numbers of enrolled early childhood students (NCES, 2013), many of these states (e.g., Connecticut, Louisiana, Wyoming) as well as a nationwide assessment program are using teacher ratings as a means of assessment of young students' knowledge, skills, and overall school readiness (Kim, Lambert, & Burts, 2013).

[1]University of Virginia, Charlottesville, VA, USA

Corresponding Author:
Emily C. Furnari, University of Virginia, 1605 Keith Valley Rd., Charlottesville, VA 22901, USA.
Email: ecf2w@virginia.edu

Research examining the concordance between teacher ratings and direct assessments of students' knowledge and skills suggests that the correlations are modest (.50-.63), and that as much as 70% of the variance in preschool teacher ratings may be unrelated to students' actual abilities (Kilday, Kinzie, Mashburn, & Whittaker, 2012; Mashburn & Henry, 2004; Südkamp, Kaiser, & Möller, 2012). These findings raise questions about the factors that may influence variability in teachers' ratings. In elementary mathematics, teachers' ratings of student behavior in first, third, and fifth grades have been found to be associated with their ratings of students' mathematics and literacy skills (Hinnant, O'Brien, & Ghazarian, 2009). And in preschool literacy, students' sex, age, and socioeconomic backgrounds have been found to be associated with teachers' ratings of their literacy skills and school readiness (Mashburn & Henry, 2004). There is still a gap in our understanding of factors that influence variability in teacher ratings in the domain of preschool mathematics. In this study, we examine the alignment between prekindergarten (Pre-K) teachers' ratings and direct assessments of students' mathematical skills at the end of the Pre-K year. We also examine the construct-irrelevant variation, or bias, in teachers' ratings, including the influence of construct-irrelevant factors such as students' demographic characteristics and social-emotional competence, as well as teachers' education, experience, and self-efficacy. To further explain variation in students' mathematical skills, we also examine the alignment between direct assessments of students' mathematical skills and the construct-irrelevant factors.

Assessment Approaches

Students' academic skills and abilities can be measured with direct or indirect assessments. In Pre-K mathematics, direct assessments involve the student performing hands-on tasks and responding to questions to demonstrate knowledge and skills. For instance, to assess skills and knowledge related to the subdomain of number sense, students are asked to count and compare groups of objects (e.g., in the Tools for Early Assessment in Mathematics [TEAM], Clements, Sarama, & Wolfe, 2011, students are asked, "How many are there?" and "Which group has more?"). To assess skills in geometry, students are asked to identify and describe shapes—for example, in the Test of Early Mathematics Ability–Third Edition (TEMA-3; Ginsburg & Baroody, 2003), students are shown a picture of a square and asked, "what shape is this?" and "How do you know it is a square?". Because direct assessments are administered outside of the classroom, often by objective assessors, and involve concrete sets of tasks or questions with pre-defined criteria, there is less risk of measurement error attributable to the assessor or the classroom context (Braun, 1976). However, direct assessments can be costly to implement with young students, are sometimes lengthy, and require a trained assessor for administration (La Paro & Pianta, 2000). Direct assessments also present more risk of child-related measurement error, as some students can become fatigued, distracted, or uncomfortable with strangers (Vacc & Ritter, 1995).

In contrast, indirect assessments involve teachers rating students' proficiencies (typically on a Likert-type scale from *emerging* to *mastered*) based on their observations of behavioral markers and indicators of student knowledge and skills. Indirect assessments require as little as 5 min per student, typically have low associated costs, and allow the students' performance on many occasions across time to be considered. However, teachers' ratings are prone to systematic errors, such as routinely scoring some students more leniently than others, overusing of the central or average rating category, or failing to discriminate students' performance on distinct skills (Ferguson, 2003; Martin & Shapiro, 2011; Martínez, Stecher, & Borko, 2009; Ready & Wright, 2011).

The average correlation between teacher ratings and directly assessed mathematical skills across Grades 1 to 4 has been found to be .58 (Xiang & Schweinhart, 2002), which is slightly lower than the average correlation for K-12 general academic direct and indirect assessments (.63; Südkamp et al., 2012). Specifically in Pre-K math, and using the same set of teacher and student participants as in the current study, Kilday and colleagues found only a .50 correlation between teacher ratings and directly assessed mathematical skills at the beginning of the school year, suggesting that other factors may be responsible for some of the variability in ratings (Kilday et al., 2012).

Variability in Teacher Ratings

In measuring students' skills and abilities, there exists a true score and measurement error (Spearman, 1904). Despite the potential shortcoming of direct assessments, scholars who have examined validity and variance in teacher ratings have typically used direct assessments as the true score to which teacher ratings are compared (e.g., Ferguson, 2003; Mashburn & Henry, 2004; Ready & Wright, 2011; Südkamp et al., 2012). Teacher ratings that are not significantly different from directly assessed scores are considered valid or accurate, whereas teacher ratings significantly above or below those directly assessed scores represent measurement error (Ferguson, 2003). Measurement error can either be random or it can be systematic, following patterns based on the characteristics of students, teachers, or classrooms. Systematic error represents trends in teacher beliefs that are expressed in their ratings of students' skills.

Concerning ratings of students' general kindergarten readiness, approximately 70% of the variance in Pre-K teacher ratings—and 50% in kindergarten teacher ratings—is attributable to construct-irrelevant factors, such as student and teacher characteristics (Mashburn & Henry, 2004). Some evidence suggests that students' characteristics influence teachers' ratings, independent of students' directly assessed abilities. For instance, in two different studies, Pre-K and kindergarten teachers rated literacy skills and school readiness of girls higher than boys, and older students higher than younger students; teachers also rated literacy skills higher for students from higher socioeconomic backgrounds. Compared with Caucasian students, teachers gave lower ratings to African American, Hispanic, and Asian students. Teachers also rated English Language Learners (ELLs) lower than their non-ELL peers (Mashburn & Henry, 2004; Ready & Wright, 2011).

There is some evidence that teachers also use students' social skills as a factor in rating their academic competencies. In studies with elementary aged students, teachers rated reading and mathematical skills higher for students they perceived to have higher social competence (Hinnant et al., 2009). We found no studies examining whether students' social-emotional competence influences early childhood teachers' ratings of academic skills, but there is some evidence that the quality of teachers' perceived relationships with Pre-K students (e.g., conflict, closeness) is associated with students' academic skills (Sabol & Pianta, 2012).

There is also evidence suggesting that teacher characteristics may influence their ratings of students (Kilday et al., 2012). Teachers with lower levels of education (Mashburn & Henry, 2004) and teachers with less than 3 years of experience (Ready & Wright, 2011) have been found to give higher ratings of students' general academic and literacy skills. Teacher self-efficacy (i.e., teachers' judgments of their abilities to bring about desired outcomes of student engagement and learning; Bandura, 1977) is another characteristic of teachers that has been linked to teachers' levels of education and experience. It has not yet been studied in association with accuracy in teachers' ratings, but has been positively linked to direct assessments of students' academic skills, such that students of teachers with higher self-efficacy perform better on direct assessments of academic skills (Anderson, Greene, & Loewen, 1988; Ashton & Webb, 1986; Ross, 1992). Teachers' self-efficacy has also been viewed as a motivational construct that influences or guides the goals teachers set, and their effort toward meeting those goals (Fives & Buehl, 2012). It may be that teachers with higher self-efficacy set higher goals for their students' achievement, and are more motivated to put forth effort and persist in helping children reach those goals. In this way, teachers' self-efficacy may also be associated with teachers' ratings of students' mathematical skills.

Present Study

The data for this study come from a field trial of MyTeachingPartner–Math/Science (MTP-M/S), which examined the impacts of Pre-K mathematics curriculum and science curriculum and associated professional development on the quality of teacher–child interactions (Whittaker, Kinzie, Williford, & DeCoster, 2016) and children's mathematics and science knowledge and skills (Kinzie et al., 2014). In our analyses, we include the intervention as a covariate, to remove any potential variation in teachers' ratings associated with their use of the curricula and/or involvement in professional development. In a previous study, using data from the larger intervention, we examined associations in the beginning of the school year between teacher ratings and direct assessments, and found a .50 correlation between teacher ratings of Pre-K students' mathematical skills and direct assessments of those skills (Kilday et al., 2012). In this follow-up study, we examine the association between teacher ratings and direct assessments of Pre-K students' mathematical skills in the spring, after teachers have had the chance to work with students for the school year. We also examine the student and teacher factors that may contribute to construct-irrelevant variance in teachers' ratings of students' mathematical skills. Specifically, we explore the following research questions:

> **Research Question 1:** What is the association between direct assessments and teacher ratings of Pre-K students' mathematical skills at the end of the school year?
>
> **Research Question 2:** To what extent are Pre-K students' demographic characteristics, teachers' perceptions of students' social-emotional competence, and teachers' education, experience, and self-efficacy related to construct-irrelevant variance in teacher ratings of students' mathematical skills at the end of the school year?

Based on previous findings of the concordance between teacher ratings and direct assessments, we hypothesize that teacher ratings in the spring will be moderately related to concurrent direct assessments of students' mathematical skills, and that this correlation will be stronger than that of previous findings based on assessments in the beginning of the school year. Regarding Research Question 2, we hypothesize that students' sex and age will be associated with construct-irrelevant variance in teacher ratings such that, controlling for direct assessments, we expect that teachers' ratings will be higher for girls as compared with boys, and older students as compared with younger students. We also hypothesize that students' socioeconomic status (SES) and race will be associated with construct-irrelevant variance in teacher ratings such that, controlling for direct assessments, teacher ratings will be lower for students from low-income families and those of minority racial/ethnic status as compared with Caucasian students. We also hypothesize that, controlling for direct assessments, teachers with less education or less experience will rate students higher than teachers with more education and more experience. Because higher teacher self-efficacy has been linked to higher student performance on direct assessments, we hypothesize that it will also be associated with higher ratings of students' skills.

Method

Participants

The sample for the current study includes 42 classrooms from a single school district near a small mid-Atlantic city. The sample included 435 students (51% female), the majority of whom were eligible for kindergarten in the subsequent academic year (99%), with ages ranging from 2.92 to 5.71 (M = 4.60, SD = 0.32) years at the start of the study. The majority of students (66%) were African American, with 25% being Caucasian and 8% from other racial/ethnic backgrounds

Table 1. Descriptive Statistics of Student and Teacher-Level Variables.

	(%) *M* (*SD*)	*n*	Missing
Student-level variables			
Gender		435	9
Male	48.40%		
Female	49.50%		
Age	4.60 (0.32)	410	34
Race/ethnicity		437	7
Caucasian	25.86%		
African American	66.00%		
Other	8.01%		
Hispanic	3.6%		
Asian	1.4%		
Native American	.5%		
Other	2.5%		
Income-to-needs ratio	1.34 (0.98)	383	61
Social competence	3.75 (0.74)	330	114
Problem behaviors	1.30 (0.40)	332	112
TR-Math raw score	4.14 (0.78)	337	107
DA-TEMA raw score	17.72 (8.84)	339	105
DA-GMA raw score	16.37 (5.17)	339	105
Teacher-level variables			
Education		39	3
Bachelor's	41.9%		
Bachelor's plus 1-year coursework	14.6%		
Master's	43.5%		
Years of experience	7.27(6.30)	36	6
Self-efficacy	8.01(0.64)	35	7

Note. TR = teacher rating; DA = direct assessments; TEMA = Test of Early Mathematical Ability; GMA = Geometry and Measurement Assessment.

(3.6% Hispanic, 1.4% Asian, .5% Native American, and 2.5% Others). The standardized income-to-needs ratio for each student's family was predominantly low; the average income-to-needs ratio (computed by taking family income, exclusive of federal aid, and dividing this by the federal poverty threshold for that family) was 1.34 (*SD* = 0.98) with 40% of households having ratios lower than one (below the poverty line) and 78% of families having ratios lower than two. Students' demographic characteristics, mean performance on direct mathematics assessments, and mean teacher ratings of mathematical skills are presented in Table 1.

Participating teachers were predominantly female (97.60%); more than half were Caucasian (53.80%), 43.60% were African American, and 2.60% Others. Their ages ranged from 24 to 65 years (*M* = 45.36 years, *SD* = 10.47). The majority of teachers had a master's degree (43.5%), with 41.9% having a bachelor's degree and 14.6% having a bachelor's plus at least one additional year of coursework. Teachers reported having between 1 and 32 years of experience in teaching Pre-K (*M* = 7.27, *SD* = 6.30). Descriptive information about teachers is presented in Table 1.

There was some teacher attrition throughout the course of the study. A total of seven teachers dropped out of the study: four were pulled to participate in another study, one teacher left because of personal circumstances, and two because of workload; the total attrition rate was 17%. To estimate attrition bias, we compared the 35 classrooms who fully participated with the seven classrooms who left, and found no significant differences (all *p*s > .05) in mean family income-to-needs ratio, teacher education, and teacher years of experience. Attrition of students also occurred. A total of 69 students could not be assessed in the spring due to teacher withdrawal from the study or student withdrawal from the preschool. To compensate for student attrition, we randomly selected additional 28 students for spring assessment from the original pool of consented students. Students who had fall data of any kind were included in the study. We conducted comparative analyses to determine whether there were significant differences between students who had both fall and spring data (n = 317) and students who had only fall or only spring data (n = 127) with regard to race/ethnicity, age, sex, or family income-to-needs ratio. We found that there was a significant difference in the number of African American students in the full sample and the sample with fall or spring only data, such that there was a higher proportion of African American students in the sample with fall or spring data only.

Procedures

At the start of the study, classrooms were randomly assigned to one of three groups: MTP-M/S curricula plus professional development supports (Plus), MTP-M/S curricula only (Basic), or a Control (Business-As-Usual) condition; we include all classrooms and use intervention group as a covariate in our analyses. At the beginning of the school year, teachers completed a survey describing their demographic and professional backgrounds as well as their self-efficacy. Teachers also sent home a consent form and short family demographic survey to all student families. A total of 94% of parents or guardians consented to allow their students to participate in the direct assessment and teacher rating component of the study. From these 529 consented students, an average of 10 students per classroom were randomly selected for participation in both fall and spring assessments. Data collectors were trained to reliability over 2 full days (see Kinzie et al., 2014, for training procedures), then visited classrooms and performed direct assessments. In addition, teachers were asked to complete rating scales on students' mathematics and science knowledge and skills. Students who were reported by teachers as having Individualized Education Plans (6% of students) or limited English proficiency (3% of students) were excluded because there were no valid and reliable mathematics and science assessments for these populations at the time of assessment.

Data for this study were collected at 2 time points: fall and spring. Demographic information on students and teachers was collected only in the fall, whereas teacher self-efficacy, ratings of students, and direct assessments of students' mathematical skills were collected in both fall and spring. This study uses spring data on all measures, with exception of demographic information, which is based on fall data. We chose to analyze spring data, as opposed to fall data, to examine the concordance between ratings and direct assessments after teachers have had multiple opportunities—throughout the school year—to observe students' skills and abilities.

Measures

Student and teacher backgrounds. Parents or caregivers completed a survey in the fall, providing information about their child's background including age, sex, race/ethnicity, and SES. As part of a fall survey, teachers reported their professional experience including their level of education (less than a bachelor's degree to graduate degree) and their years of experience in working professionally with Pre-K children.

Direct assessments of students' mathematical skills

Number sense and operations. We used the TEMA-3 (Ginsburg & Baroody, 2003) to directly assess students' number sense and operations knowledge and skills. This standardized, norm-referenced measure is designed for use with students between 3 and 8 years of age, and uses pictures and counting chips to assess skills in counting, ordinality, cardinality, one-to-one correspondence, numeral recognition, and abilities in numerical operations. Ginsburg and Baroody (2003) report that all alphas, test–retest, and alternate form reliability with immediate and delayed administration, exceeded .90 for all subgroups (i.e., different age, sex, ethnicity, and achievement-level groupings), with the exception of test–retest reliability for Form A, which was .82 (Spies, Plake, & Murphy, 2005). TEMA developers have established concurrent validity with both the KeyMath–R Basic Concepts subtest (*r* = .54) and the Young Children's Achievement Test Math Quotient (*r* = .91); these correlations support the claim that the TEMA-3 is measuring concepts similar to those assessed by other related tests (Spies et al., 2005). We found excellent internal reliability in students' TEMA-3 performance (.91 in the fall and .93 in the spring).

Geometry and measurement. To assess students' knowledge of shapes, patterns, measurement, and positional words, we use the Geometry and Measurement Assessment (GMA), which is a derivative of the TEAM (Clements et al., 2011). The GMA includes six original TEAM items, seven extension questions, and 17 new items. All of the extensions and new items were developed to address additional related curricular objectives not assessed in the TEAM and mirrored the format and style of the original TEAM items; for example, to extend a question requiring students to make a triangle and rectangle using coffee stirrers, a related question was added for making a square. The TEAM developers established construct validity (Clements, Sarama, & Liu, 2008), and we found an internal reliability of .82 in the fall and .86 in the spring.

To later compare students' direct assessment scores to teachers' ratings of their overall mathematical skills, we created a composite from the two direct assessments (TEMA-3 and GMA). A correlational analysis showed they were moderately correlated and could be composited (*r* = .66). Giving each measure equal weight, we standardized students' scores on both measures and averaged the two, representing the overall direct assessment mathematics score for each student; the mean is 0 (*SD* = 1) with a range of −2.57 to 2.53.

Teacher ratings of students' mathematical skills. The Academic Rating Scale–Mathematics (ARS-M) was developed by the Early Childhood Longitudinal Study–Kindergarten Cohort (ECLS-K): 1998-1999 (Rock & Pollack, 2002). We added five items (e.g., identifies and understands ordinality, identifies and understands cardinality, uses instruments accurately for measuring) to the original seven items to address additional mathematical competencies covered in the *MTP-M/S* curricula (Kinzie et al., 2014) and the direct assessments we used (TEMA-3; GMA). For each item, the teacher rates the degree to which the student has exhibited a particular skill, on a scale of 1 to 5 (1 = *not yet*, 2 = *beginning*, 3 = *in progress*, 4 = *intermediate*, and 5 = *proficient in the skill*). Teachers were also given the option to mark any skill as "Non-Applicable." Items reflected students' knowledge and skills in the areas of number sense, operations, geometry, and measurement. The ECLS-K Psychometric Report demonstrated high person reliability (.94), which is analogous to Cronbach's alpha, for the ARS-M in the spring of both kindergarten and first grade (Rock & Pollack, 2002). We found excellent internal consistency in our sample, with a Cronbach's alpha of .97. We computed a mean ARS-M score for each student based on their teacher's spring ratings. The average teacher rating of students' mathematical skills in the spring was "intermediate," at 4.14 (*SD* = 0.78) with a range of 1.58 to 5.00 (see Table 1 for descriptive information on teacher ratings).

Students' social-emotional competence. The Teacher-Child Rating Scale (TCRS; Hightower et al., 1986) is a 38-item teacher-report measure with subscales for problem behaviors, and social competence. The problem behaviors subscale measures students' acting-out, levels of shyness/anxiousness, and learning problems; it requires the teacher to rate the degree to which each item is a problem for the student (1 = not a problem, 5 = very serious problem). The social competence subscale measures students' reactions to limits/frustration tolerance, assertive social skills, and task orientation; teachers respond to each item by rating how well it describes the student (1 = not at all, 5 = very well). Excellent psychometric properties have been indicated for this measures when used with preschoolers (internal consistency and test–retest reliabilities range from .85 to .95; Hightower et al., 1986); concurrent validity has been established between the TCRS and other behavioral checklists (e.g., Trickett, McBride-Chang, & Putnam, 1994), and we found high internal consistency for the problem behaviors and social competence subscales (α = .90 and .94, respectively).

Teachers' self-efficacy. The Teachers' Sense of Efficacy Scale–Short Form (Tschannen-Moran & Woolfolk Hoy, 2001) is a 12-item Likert-type questionnaire in which teachers report their perceived levels of effectiveness on student engagement, instructional strategies, and classroom management. Internal reliability of this scale has been reported to be high, with alphas ranging from .81 to .86 on the subscales and .90 overall (Tschannen-Moran & Woolfolk Hoy, 2001); our analyses also showed high internal reliability in our sample (α = .90). Construct validity has also been confirmed with .16 and .64 correlations to subscales of the Teacher Efficacy Scale (Hoy & Woolfolk, 1993).

Analyses

The overall goal of our analysis was to examine (a) the association between teacher ratings of mathematical skills and students' directly assessed mathematics abilities, and (b) which factors influence teachers' ratings of the students' mathematical skills, after controlling for the variation explained by students' directly assessed mathematics abilities. We used ordinary least squares regression to predict teacher ratings of student mathematical skills first from the direct assessment of students' skills. The resulting coefficient represents the concordance between direct assessments and teacher ratings of students' mathematical skills.

Next, we added student demographic characteristics (sex, age, race/ethnicity, and SES), students' social-emotional competence (social competence and problem behaviors), and teacher education, experience, and self-efficacy, while controlling for students' directly assessed mathematical skills. The resulting coefficients represent the degree to which each factor is associated with systematic differences in teacher ratings, after accounting for the variance explained by students' directly assessed abilities and the other variables in the model. We also examined partial correlations between the predictor variables and teacher ratings, while controlling for direct assessments. These correlations represent the relationship between each predictor and teacher ratings, after accounting for the variance explained by students' directly assessed abilities. We use a cluster design to account for the nesting of students within classrooms. Analyses were run in Mplus Version 7.0 using full information maximum likelihood estimation so that data analyses used all available data when estimating parameters, increasing the precision and accuracy of the estimated parameters (Enders & Bandalos, 2001). Models were run using Mplus's type equals complex mode, which adjusts standard errors and p values, thereby accounting for the nesting of students within teachers. We include the intervention condition as a covariate in both models, and the partial correlation analyses, using two dummy coded variables representing each of the intervention groups with the control group as the reference category. We graphically examined the residuals of our models to examine the assumptions of normality and equal variances, and did not find evidence of any substantial violations.

To further answer Research Question 2, and better understand the relationship between the predictor variables and direct assessments of student skills, we examined a regression model predicting students' direct assessment scores from all of the independent variables in the original predictive model. The resulting coefficients represent the degree to which each independent variable relates to students' directly assessed mathematical skills.

Results

When examining the relationship between direct assessments and ratings alone, the concordance is 0.50 ($SE = 0.06, p < .001$). We observed that 25% of the variability in teacher ratings could be attributed to students' directly assessed abilities, indicating that a substantial amount of variance remained to be potentially explained by student and/or teacher characteristics. After adding the other variables in our regression model, we observed that the predictive equation could explain 55% of the variability in teacher ratings. This 30% difference suggests that a larger proportion of the variability in teacher ratings is associated with construct-irrelevant factors than is associated with the direct assessment of student skill. The coefficients from the estimated model are presented in Table 2. In addition, we examined partial correlations between each of the predictor variables and teacher ratings, while controlling for direct assessments. The resulting correlations are presented in Table 2.

After controlling for students' directly assessed skills, the results indicate that students' race/ethnicity and social competence, as well as teacher self-efficacy were significantly related to Pre-K teacher ratings of students' mathematical skills, independent of students' directly assessed abilities. Specifically, students in the Other race/ethnicity category (e.g., Hispanics and Asians) were rated as having significantly lower levels of mathematical competence than were Caucasians, and students with higher social competence were rated more highly in their mathematical skills. Teachers with greater self-efficacy also provided higher ratings of their students. Partial correlations for student social competence, as well as teacher self-efficacy, showed significant relations with teacher ratings, in directions consistent with results of the regression model. However, partial correlations for student race showed that without controlling for other variables, race was not significantly related to teacher ratings.

Results examining direct assessments as an outcome indicated that students' social competence was positively associated with their directly assessed mathematical skills ($b = 0.321, p < .001$). However, despite being related to teacher ratings in the previous model, students' race and teachers' self-efficacy were not significantly associated with their directly assessed mathematical skills.

Discussion

Our results suggest that Pre-K teachers' ratings of students' mathematical skills at the end of the year are moderately aligned with concurrent direct assessments of those skills ($b = 0.50$). This is consistent with our previous finding that the concordance was also .50 in the beginning of the year (Kilday et al., 2012), and is considerably lower than the average concordance for K-12 general academic direct and indirect assessments ($b = 0.63$) (Südkamp et al., 2012). In addition, more of the variability in teacher ratings could be accounted for by construct-irrelevant student and teacher characteristics than by students' directly assessed mathematical skills (30% and 25%, respectively). The finding that only 25% of the variance in teachers' ratings could be explained by students' performance on direct assessments of the those skills suggests that Pre-K teachers are even less accurate at assessing students' skills in mathematics as compared with literacy and school readiness, where student performance on direct assessments was been found to explain 30% of the variance in teacher ratings (Mashburn & Henry, 2004).

Student Characteristics

Teachers in our study tended to rate students in the Other racial category (e.g., Hispanics and Asians) as having lower levels of mathematical knowledge and skills, as compared with Caucasian students. Similarly, Ready and Wright (2011) found that Pre-K and kindergarten

Table 2. Standardized Regression Coefficients for Predicting Pre-K Teachers' Ratings of Math Skills.

Predictor	Estimate (*SE*)	*p* value	Partial *r*	*p* value
Model 1: Direct assessment only (R^2 = .25)				
Plus dummy code	−0.07 (0.123)	.57	—	—
Basic dummy code	0.159 (0.108)	.14	—	—
DA-Math[a]	0.497 (0.058)	<.001	—	—
Model 2: Direct assessment plus student and teacher characteristics (R^2 = .55, ΔR^2 = .30)				
Plus dummy code	−0.059 (0.089)	.51	—	—
Basic dummy code	0.063 (0.074)	.39	—	—
DA-Math[a]	0.306 (0.059)	<.001	—	—
Student age	0.028 (0.041)	.50	.069	.234
Student sex: female	0.051 (0.033)	.12	.086	.118
Student race: African American	−0.037 (0.059)	.53	.002	.964
Student race Other	−0.083 (0.04)	.04	−.046	.404
Student income-needs ratio	0.023 (0.045)	.62	−.008	.899
Student social competence	0.41 (0.082)	<.001	.469	<.001
Student problem behaviors	−0.017 (0.065)	.79	−.322	<.001
Teacher education	−0.046 (0.089)	.61	−.021	.728
Teacher experience	0.077 (0.078)	.32	.140	.013
Teacher self-efficacy	0.234 (0.084)	.005	.253	<.001

Note. The partial-*r* estimates and corresponding *p* values represent the relationship between the given predictor and teacher ratings, while controlling for DA-Math, and the intervention dummy variables.
[a] DA-Math indicates the composite mathematics direct assessment score.

teachers rated Asian and Hispanic students as having lower literacy skills. The relatively small proportion of both students and teachers in the Other racial category (8% and 2.5%, respectively) may have made it more difficult to detect true patterns in teachers' ratings as dependent on students' race, as demonstrated by our finding that the partial correlation between students' race and teacher ratings was not significant. Future research should examine whether teacher ratings are influenced by students' race in populations that have a more even distribution of students' and teachers' races.

Teachers rated mathematical skills more highly for students whom they perceived to be more socially competent. Our findings are supported by those of Hinnant and colleagues (2009) who found that elementary school teachers rated students with better social skills as being more competent in both reading and mathematics. Further explaining this finding, our post hoc regression analyses suggest that more socially competent students are, in fact, more proficient in mathematics, but that teachers may be oversensitive to this pattern (teacher perceptions of students' social competence were associated with students' directly assessed mathematical skills).

Our hypotheses that girls would be rated higher than boys, older students would be rated higher than younger students, and that students from higher socioeconomic backgrounds would be rated higher, were not supported by our results. This is inconsistent with previous research suggesting that students' sex, age, and socioeconomic backgrounds are associated with Pre-K and kindergarten teacher ratings of general academic performance, literacy skills, and communication skills (Mashburn & Henry, 2004; Ready & Wright, 2011). It may be that teacher ratings of students' mathematical skills are influenced by different factors than teacher ratings of students' general academics, literacy, and communication skills. For instance, teacher ratings of preschool mathematics abilities may be influenced by cultural biases (Lubienski, 2008) that favor males over females in mathematics; this could lead to a neutralizing effect whereby girls and boys are actually rated the same at an aggregate level.

Teacher Characteristics

Teachers reporting higher levels of self-efficacy also rated students' higher on their mathematical skills, independent of students' directly assessed mathematical skills. There is evidence that academic achievement is higher among students whose teachers report higher levels of self-efficacy (Anderson et al., 1988; Ashton & Webb, 1986; Ross, 1992); however, teachers may overestimate their effect on students' development. Our post hoc analyses revealed that teacher self-efficacy in our study was not tied to students' performance on direct assessments. This suggests that Pre-K teachers may be inaccurately attributing their students' end-of-year mathematical skills to their beliefs about their own abilities to promote desired outcomes in students. Further research in this area is needed to disentangle the ways in which self-efficacy influences teachers' ratings of their students' skills.

Teachers' education and experience were not related to systematic patterns in their ratings. This is contrary to previous research suggesting that teachers with lower levels of education and less experience rate students higher (Mashburn & Henry, 2004; Ready & Wright, 2011). Our results may be different because all of the teachers in our sample had at least a bachelor's degree, and the majority had a master's, whereas the significant findings from Mashburn and Henry's (2004) study included a group of teachers who had less than a bachelor's degree. In addition, our findings with regard to both teacher education and experience may have been affected by our relatively small sample (42 teachers) as compared with that of Ready and Wright (2011), who found significant associations around teacher experience with a much larger sample of teachers.

Implications

Knowing the factors that influence teacher ratings can help ensure appropriate interpretations by researchers who rely on teacher ratings as assessments of student performance. For instance, if these trends were uniformly found, researchers could potentially control for the sources of systematic variation in teacher ratings of student skills, to produce a result that more closely replicates a direct assessment of student skills. It may also be possible to train teachers to become more objective assessors; scholars have found that training teachers in giving direct assessments of student skills improves their accuracy when estimating their students' performance (Begeny & Buchanan, 2010), and their ability to administer a battery of direct and indirect assessments as intended by measure developers (Williford, Downer, Hamre, & Pianta, 2014).

Limitations

One limitation to this study is that our teacher rating scale and direct assessments quantify students' skills on different scales, and therefore cannot be directly compared without concealing a great deal of variance in each of the independent measures. To avoid directly comparing the measures, we use a predictive model, thereby allowing us to keep the scales of the separate measures intact. However, the limitation of this analysis is that the coefficients are only able to tell us whether teacher ratings tend to be higher or lower than direct assessments, rather than whether teachers overrate or underrate relative to the direct assessments. To determine whether teacher ratings are over or under estimating students' abilities, a discrepancy score is required and thus measures using the same scales (same items and same scoring/rating) are necessary. Another limitation is the problem of measurement error, which is a challenge to both direct and indirect assessments. Although this study focuses on the potential biases in teacher ratings, it is important to remain cognizant of the error that is also present in direct assessments, which are subject to child-related error (e.g., fatigue, distractibility, discomfort; Vacc & Ritter, 1995). Finally, there were a higher proportion of African American students in the sample with only fall or only spring data; thus, results around the influence of students' race on teachers' ratings should be interpreted with caution.

Conclusion

Early mathematics assessment is required at the beginning of kindergarten in 22 states (NCES, 2013). Several of these states (e.g., Connecticut, Louisiana, Wyoming; Daily, Burkhauser, & Halle, 2010) and a nationwide assessment program (Kim et al., 2013) rely on teacher ratings of student skills for either all or part of their assessments. The findings of the current study suggest that teachers are able to draw on their rich experiences interacting with students to rate student proficiencies; however, teacher biases pose potential complications for use of their ratings as measures of students' academic skills. Research triangulating students' actual abilities with various modes of assessment will inform researchers and policy makers so that they can better measure students' school readiness, guide instruction, and make decisions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A120631 to the University of Virginia.

References

Anderson, R. N., Greene, M. L., & Loewen, P. S. (1988). Relationships among teachers' and students' thinking skills, sense of efficacy, and student achievement. *Alberta Journal of Educational Research*, *34*, 148-165.

Ashton, P. T., & Webb, R. B. (1986). Making a difference: Teachers' sense of efficacy and student achievement. New York: Longman.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191-215. doi:10.1037/0033-295X.84.2.191

Begeny, J. C., & Buchanan, H. (2010). Teachers' judgments of students' early literacy skills measured by the Early Literacy Skills Assessment: Comparisons of teachers with and without assessment administration experience. *Psychology in the Schools*, *47*, 859-868. doi:10.1002/pits.20509

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

Braun, C. (1976). Teacher expectation: Sociopsychological dynamics. *Review of Educational Research*, *46*, 185-213. doi:10.3102/00346543046002185

Cabell, S. Q., Justice, L. M., Zucker, T. A., & Kilday, C. R. (2009). Validity of teacher report for assessing the emergent literacy skills of at-risk preschoolers. *Language, Speech, and Hearing Services in Schools*, *40*, 161-173.

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-Based Early Maths Assessment. *Educational Psychology*, *28*, 457-482. doi:10.1080/01443410701777272

Clements, D. H., Sarama, J., & Wolfe, C. (2011). *TEAM—Tools for Early Assessment in Mathematics*. Columbus, OH: McGraw-Hill Education.

Daily, S., Burkhauser, M., & Halle, T. (2010). A review of school readiness practices in the states: Early learning guidelines and assessments. *Early Childhood Highlights*, *1*(3). Retrieved from http://www.childtrends.org/wp-content/uploads/2013/05/2010-14-SchoolReadinessStates.pdf

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

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*, 430-457.

Ferguson, R. F. (2003). Teachers' perceptions and expectations and the Black-White test score gap. *Urban Education*, *38*, 460-507. doi:10.1177/0042085903038004006

Fives, H., & Buehl, M. M. (2012). Spring cleaning for the "messy" construct of teachers' beliefs: What are they? Which have been examined? What can they tell us? In K. R. Harris, S. Graham, T. Urdan, S. Graham, J. M. Royer, & M. Zeidner (Eds.), *APA educational psychology handbook, Vol. 2: Individual differences and cultural and contextual factors* (pp. 471-499). Washington, DC: American Psychological Association. Retrieved from http://content.apa.org/books/13274-019

Ginsburg, H. P., & Baroody, A. J. (2003). *TEMA-3 examiners manual* (3rd ed.). Austin, TX: Pro-Ed.

Hightower, A. D., Work, W. C., Cowen, E. L., Lotyczewski, B., Spinell, A., Guare, J., & Rohrbeck, C. (1986). The Teacher-Child Rating Scale: A brief objective measure of elementary children's school problem behaviors and competencies. *School Psychology Review*, *15*, 393-409.

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

Hoy, W. K., & Woolfolk, A. E. (1993). Teachers' sense of efficacy and the organizational health of schools. *The Elementary School Journal*, *93*, 355-372.

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., & Whittaker, J. V. (2012). Accuracy of teacher judgments of preschoolers' math skills. *Journal of Psychoeducational Assessment*, *30*, 148-159. doi:10.1177/0734282911412722

Kim, D.-H., Lambert, R. G., & Burts, D. C. (2013). Evidence of the validity of teaching strategies GOLD® assessment tool for English language learners and children with disabilities. *Early Education and Development*, *24*, 574-595. doi:10.1080/10409289.2012.701500

Kinzie, M. B., Whittaker, J. V., Williford, A. P., DeCoster, J., McGuire, P., Lee, Y., & Kilday, C. R. (2014). MyTeachingPartner-Math/Science Pre-Kindergarten curricula and teacher supports: Associations with children's mathematics and science learning. *Early Childhood Research Quarterly*, *29*, 586-599. doi:10.1016/j.ecresq.2014.06.007

La Paro, K. M., & Pianta, R. C. (2000). Predicting children's competence in the early school years: A meta-analytic review. *Review of Educational Research*, *70*, 443-484.

Le Floch, K. C., Martinez, F., O'Day, J., Stecher, B., Taylor, J., & Cook, A. (2007). *State and local implementation of the "No Child Left Behind Act." Volume III—Accountability under NCLB: Interim report*. Washington, DC: U.S. Department of Education.

Lubienski, S. T. (2008). On "gap gazing" in mathematics education: The need for gaps analyses. *Journal for Research in Mathematics Education*, *39*, 350-356.

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

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment*, *14*, 78-102. doi:10.1080/10627190903039429

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.

National Center for Education Statistics (NCES). (2013). *The Condition of Education 2013* (NCES 2013-037). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch

NCLB ESEA Flexibility. (2015, February 25). [Letters (Correspondence)]. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

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. doi:10.3102/0002831210374874

Rock, D. A., & Pollack, J. M. (2002). *Early Childhood Longitudinal Study-Kindergarten class of 1998-99 (ECLS-K): Psychometric report for kindergarten through first grade* (Working Paper Series). Retrieved from http://nces.ed.gov/pubs2002/200205.pdf

Ross, J. A. (1992). Teacher efficacy and the effects of coaching on student achievement. *Canadian Journal of Education/Revue Canadienne de L'éducation*, *17*, 51-65.

Sabol, T. J., & Pianta, R. C. (2012). Recent trends in research on teacher–child relationships. *Attachment & Human Development*, *14*, 213-231. doi:10.1080/14616734.2012.672262

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology, 15*(2), 201-292. doi:10.2307/1412107

Spies, R. A., Plake, B. S., & Murphy, L. L. (2005). *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*, 743-762. doi:10.1037/a0027627

Trickett, P. K., McBride-Chang, C., & Putnam, F. W. (1994). The classroom performance and behavior of sexually abused females. *Development and Psychopathology*, *6*, 183-194. doi:10.1017/S0954579400005940

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*, 783-805.

Vacc, N. A., & Ritter, S. H. (1995). *Assessment of preschool children*. ERIC Digest. Retrieved from http://files.eric.ed.gov/fulltext/ED389964.pdf

Williford, A. P., Downer, J. T., Hamre, B., & Pianta, R. C. (n.d.). *The Virginia Kindergarten Readiness Project: Executive summary & legislative report: Fall 2014, Phase II.* Richmond, VA: Elevate Early Education. Retrieved from http://curry.virginia.edu/uploads/resourceLibrary/VKRP_Executive_Summary_and_Legislative_Report_2015_01_21_updated_%281%29.pdf (accessed 20 March 2015).

Whittaker, J. V., Kinzie, M. B., Williford, A., & DeCoster, J. (2016). Effects of MyTeachingPartner–Math/Science on teacher–child interactions in prekindergarten classrooms. *Early Education and Development, 27*(1), 110-127. doi:10.1080/10409289.2015.1047711

Xiang, Z., & Schweinhart, L. J. (2002). *Effects five years later: The Michigan School Readiness Program evaluation through age 10*. HighScope Educational Research Foundation. Retrieved from https://test.coradvantage.org/file/Research/Effects%205%20Years%20Later.pdf