Are Observed Classroom Practices Related to Student Language/Literacy Achievement?

Yucheng Cao iD
Young-Suk Grace Kim iD
Minkyung Cho
University of California, Irvine

In this study, we examined the relation of observed classroom practices to language and literacy achievement and the moderation of this relation for students from pre-K to sixth grade. A total of 136 studies (N = 107,882 participants) met the inclusion criteria, of which 108 studies were included for meta-analysis and the other 28 studies were narratively synthesized. The average zero-order (r = .12) and partial correlations (r_p = .04) were statistically significant but weak in magnitude. The relation was slightly weaker in upper than in lower grade levels, and stronger for observations capturing macro quality and instructional dimension than those capturing micro measurement and emotional or structural dimension, respectively. The relation did not vary by observation duration, frequency, adopted statistical approach, or type of covariates. Taken together with the narrative synthesis, the results highlight the complex nature of classroom observation and a need for more classroom research, particularly on higher grade levels.

The classroom is the primary setting for students where formal learning and social interactions take place. A considerable portion of the variance in student learning can be explained at the classroom level (Foorman et al., 2006; Hanushek, 2002). It is well established that effective instruction is contingent on multiple instructional components (e.g., content, organization; Pressley et al., 2001), dynamic teacher-child interactions (Cabell et al., 2013; Crosnoe et al., 2010; Pianta & Hamre, 2009), and transactional child-instruction interactions (Connor, Piasta, et al., 2009; Morrison & Connor, 2009). To capture this complex and multidimensional construct, classroom observation has long been used as a

measurement tool (Pianta & Hamre, 2009). However, early reviews in the 1950s and 1960s consistently revealed inconclusive or confusing associations between teaching acts and student outcomes (Ackerman, 1954; Flanders & Simon, 1969; Morsh & Wilder, 1954), which could be attributed to subjective classroom observations, untenable hypotheses, problematic statistical methods (Frick & Semmel, 1978; Gage, 1963; Medley & Mitzel, 1963), and equivocal predictor and criterion variables (Yamamoto, 1963). More recently, quantitative research has used standardized observations at scale with adequate reliability and validity and has contributed more generalizable and robust empirical evidence on effective teaching (e.g., Connor, Jakobsons, et al., 2009; Early et al., 2006; La Paro et al., 2004).

In the present study, we aimed to extend the prior literature on classroom observations to examine the relation of the quality and quantity of classroom practices to students' language and literacy performance from prekindergarten (pre-K) to sixth grade, using meta-analysis. Previous meta-analyses and qualitative systematic reviews focused more on early childhood education with macro-level observations (e.g., overall classroom quality) and reported weak associations between childcare quality and preschoolers' academic, socioemotional, and behavioral outcomes (Brunsek et al., 2017; Burchinal et al., 2011; Keys et al., 2013; Perlman et al., 2016; Ulferts et al., 2019). Furthermore, few of them examined students in upper elementary grades and their language/literacy development. In this study, we addressed these gaps by investigating the classroom observation–student achievement relations, including (a) both macro- and micro-level observation characteristics and (b) students from pre-K to grade six. Moreover, we explored a series of potential moderators (e.g., student grade level, observation type and dimension, observation duration and frequency, child language and literacy outcomes, adopted statistical approach, and covariates).

Observation Type and Dimension

Numerous observation instruments have been created and developed. Given the complexity of classroom instruction, there is no consensus on how to classify classroom teaching practices. However, one broad approach that has been adopted in previous work is conceptualizing classroom instruction as the macro-level classroom quality and as the microlevel discrete classroom practices (Gosling, 2002; Wragg, 1999; Connor et al., 2014). The macro-level observation typically rates the global quality of the teacher–child interactions (e.g., teacher responsiveness, student engagement) and/or the classroom structural features (e.g., physical environment, class size, teacher qualifications), and these typically have high-inference composite scores or indices. Examples include the Early Childhood Environment Rating scale and its revised edition (ECERS and ECERS-R; Harms et al., 1998), the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008), the Early Childhood Classroom Observation Measure (ECCOM; Stipek, 1996), and the Classroom Practice Inventory (CPI; Hyson et al., 1990). In contrast, the microlevel observation uses a low-inference, time-sampling coding system, commonly measuring the discrete occurrences (e.g., amount, ratio) of certain teacher/student behaviors, pedagogical strategies, and settings, such as the Individualizing Student Instruction classroom observation (ISI; Connor, Morrison, et al., 2009). In addition, there are validated observation

instruments that incorporate both global ratings and counts of discrete instances, such as the Observational Record of the Caregiving Environment (ORCE; National Institute of Child Health and Human Development Early Child Care Research Network [NICHD ECCRN], 1996) and the Classroom Observation System-K-5 (COS-K-5; NICHD ECCRN, 2002).

In addition to the differences in the evaluation approaches such as macro and micro aspects, extant classroom observation instruments also vary in the content of instruction that they evaluate. In general, classroom observation instruments examine three dimensions of content of instruction: (a) the instructional dimension, such as quantity and quality of literacy content delivery, explanation and monitoring, and stimulation and feedback; (b) the emotional dimension, such as classroom climate and organization, praise and discipline, behavior management, sensitivity, responsivity, detachment, and disengagement; and (c) the structural dimension, such as classroom physical environment, book category, and writing materials category. We acknowledge variation within the macro- and microlevel observation systems—different observation instruments were developed with different goals and conceptualizations. For example, sociocultural frameworks such as culturally responsive teaching (Gay, 2002) and critical literacy (Luke, 2012) emphasize that learning occurs through social interactions and encourage learning from experience and discourse (Vygotsky, 1980), whereas social cognitive perspectives highlight individual cognitive skills such as self-regulation (Connor, 2016), socioemotional aspects such as self-efficacy, outcome expectancies, and sociostructural impediments and reinforcements in the learning and performance of actions (Bandura, 2001; Schunk, 2012). Classifying observation systems into observation types and dimensions was not to ignore these differences, but instead to examine whether any global differences in classroom observations such as high-inference rating versus low-inference quantification are differentially related to elementary students' language and literacy skills. The same is true for the content of instruction. Therefore, the results of the present meta-analysis should be interpreted with this in mind.

Relation Between Teaching Practices and Student Achievement

Previous studies have found that the association between global process quality (i.e., emotional and instructional interactions, materials and activities) and student achievement was not consistently significant, and even when significant, the relation was modest in magnitude (Burchinal et al., 2008, 2011; Guo, Connor, et al., 2012; Howes et al., 2008; Mashburn et al., 2008; Taylor et al., 2000; Weiland et al., 2013). The early childhood education and childcare support systems vary geographically, and many European studies have further corroborated this significant but small effect of global process quality on children's development (Abreu-Lima et al., 2013; Cadima et al., 2010; Ulferts et al., 2019). Researchers have suggested that factors such as short pre-post interval and lack of valid, reliable, or suitable measures might help explain the modest or non-significant relations (Burchinal et al., 2011; Keys et al., 2013; Weiland et al., 2013). Furthermore, emerging evidence has suggested a nonlinear relation where the effect is larger in classrooms with a higher quality of instruction (Burchinal et al., 2010, 2016; Cadima et al., 2010; Hatfield et al., 2016).

Despite the limited relation to general academic achievement, previous studies have linked the global quality of teacher-child interactions to students' *language and literacy* development, suggesting that stronger instructional support (Curby et al., 2009; Mashburn et al., 2008), more responsive and sensitive teacher-child interaction (Burchinal et al., 2010; Cornelius-White, 2007; Curby et al., 2009; Howes et al., 2008), and stronger classroom organization (Cadima et al., 2010; Ponitz Rimm-Kaufman, Brock, & Nathanson, 2009) lead to greater literacy gains. Similar patterns have been identified in the studies measuring microlevel classroom practices. For instance, students attain higher achievement in classrooms with more instructional time (time on academic activities; Connor, Morrison, & Katch, 2004; Connor, Morrison, & Petrella, 2004; Taylor et al., 2000; Wharton-McDonald et al., 1998), more positive and scaffolded engagement (Cameron et al., 2008; Pressley et al., 2001), and more teacher control (less off-task activity and disruption; Crocker & Brooker, 1986; Day et al., 2015; Ponitz & Rimm-Kaufman, 2011). Given that observations vary by type and dimension, it is reasonable to speculate differential relations between classroom practices and student achievement as a function of the nature of observation.

Teacher, Student, and Classroom Factors

A variety of teacher, student, and classroom features have been examined in prior research. Teacher credential/education, teaching experience, and teacher knowledge and beliefs have been found to significantly influence student achievement though to a limited extent or indirectly through classroom practices (Cash et al., 2015; Darling-Hammond & Youngs, 2002; Early et al., 2006; Pianta et al., 2005; Wayne & Youngs, 2003). Children's age, gender, and initial skills as well as their home literacy, parent education, and socioeconomic status (SES) are also common predictors/covariates in classroom research (e.g., Burchinal et al., 2000; Connor et al., 2005; Ponitz & Rimm-Kaufman, 2011), though findings are inconsistent. For example, researchers found that student age was negatively related to the classroom quality–student outcome association (Burchinal et al., 2011), students' baseline skills and primary language were significantly associated with instructional effectiveness in reading (Park et al., 2019), and childcare quality varied across and within geographic regions and countries (Vermeer et al., 2016). Furthermore, evidence has suggested that classroom instruction varies by child characteristics (e.g., baseline skill level), such that students differentially benefited from given instruction (Connor, Morrison, & Petrella, 2004). In contrast, Keys et al. (2013) found nonsignificant moderating effects of children's demographic characteristics (race, gender, socioeconomic status), baseline skills, and behaviors. Likewise, previous literature has shown nonsignificant associations between the structural features (e.g., program infrastructure and design such as teacher–child ratio, teacher qualifications, class size) and student academic and social development (Howes et al., 2008; Mashburn et al., 2008). For instance, some found few class-level characteristics were associated with or predictive of classroom quality or children's academic outcomes (Early et al., 2007; Justice et al., 2008; NICHD EECRN, 2002; Walsh & Tracy, 2004). Overall, these studies indicate that the relation between classroom instruction and student achievement

might change as a function of a multitude of student-, family-, teacher-, and class-level features.

Other Factors

Factors such as the outcome domain (e.g., language versus mathematics) and analytic approaches (e.g., multilevel, latent approaches) as well as different types of predictors/covariates are also potential moderators. Specifically, measurement error attenuates the relation of interest. Latent variable approaches account for measurement errors and the dimensionality of constructs with a potential increase in effect size. Multilevel models account for the nested data structure where students are clustered at the classroom level so they yield less biased estimation and significance tests. For example, in a longitudinal meta-analysis, Ulferts and colleagues (2019) detected a lasting impact of child care quality on language and mathematics development throughout the primary school phase, with larger effects found in studies that applied multivariate analyses while controlling for child and family background characteristics.

Current Study

Tremendous heterogeneity exists in how observation is operationalized and measured, what child outcomes are measured, and what statistical approaches are employed (Brunsek et al., 2017; Perlman et al., 2016; Ulferts et al., 2019). However, insufficient research has been conducted to synthesize how both macro- and microlevel classroom practices and different dimensions of instruction, as well as teacher and child factors, are related to early childhood and elementary students' language and literacy achievement (an exception is Park et al., 2019, which is a narrative review). In the present study, we extended prior reviews to examine the relation between both the quality and quantity of classroom practices and students' language and literacy performance from pre-K to sixth grade, using a meta-analysis.

Two research questions guided our investigation: What is the relation of observed classroom practices to students' language and literacy performance from pre-K to sixth grade? Does the relation vary by student grade level, observation type and dimension, observation duration and frequency, child language and literacy outcomes, adopted statistical approach, and covariates? According to the previous research findings, we hypothesized a weak association between the observational results and student language/literacy outcomes, with the association expected to be stronger for younger students than older students (Burchinal et al., 2011). In addition, we expected a varied relation by the features of observation and assessed skills (Brunsek et al., 2017; Keys et al., 2013; Perlman et al., 2016; Ulferts et al., 2019).

Because we focused on only the domain of language and literacy outcomes, we classified child outcomes as reading (phonological awareness, print concept, letter knowledge, word reading, reading fluency, reading comprehension), language (oral language, listening comprehension, vocabulary), and writing (spelling, handwriting, writing quality) in this study. Although the effects of observation duration and frequency have not been widely examined in previous studies, we noticed great ranges of the observation length and interval across instruments and

studies. Hence, it is an open question whether observation duration and frequency might influence results.

Method

Literature Search

To identify relevant studies, six electronic databases (Academic Search Complete, Education Source, ERIC, Primary Search, Teacher Reference Center, and psychINFO) were searched using the following combination of terms: all(class* observ*) AND all(teach* OR instruct* OR organiz* OR act* OR practi* OR control* OR support) AND all(litera* OR lang* OR lingu* OR lexic* OR read* OR letter OR word OR comprehen*). This search strategy was built on the three core constructs in the current review: classroom observation, teacher instruction, and language/literacy outcome. For each construct, we included multiple synonyms and added truncation wildcards to ensure high searching productivity (i.e., replacing word's ending with an asterisk to recruit all possible words with the same root). Moreover, we used the above Boolean operators to ensure any study that simultaneously contains at least one term from each construct to be retrieved. To further restrict the search to include the target population, additional filters and limiters (embedded in the database system by default) were applied (see Table S1, available on the journal website). In addition, 17 relevant journals were digitally searched: *American Educational Research Journal*, *Child Development*, *Contemporary Educational Psychology*, *Early Childhood Research Quarterly*, *Early Childhood Education Journal*, *Early Education and Development*, *Early Child Development and Care*, *Journal of Research in Childhood Education*, *Journal of Educational Psychology*, *Journal of Educational Research*, *Journal of Language Teaching and Research*, *Journal of Literacy Research*, *Reading Research Quarterly*, *Research in the Teaching of English*, *Scientific Studies of Reading*, *The Elementary School Journal*, and *Theory & Practice in Language Studies*. Articles were also collected by contacting scholars and experts on classroom research, which yielded an additional 16 studies. Finally, the reference lists of extant pertinent meta-analyses and synthesis reviews were manually reviewed to recruit possible studies (Brunsek et al., 2017; Burchinal et al., 2011; Keys et al., 2013; Park et al., 2019; Perlman et al., 2016; Ulferts et al., 2019).

Inclusion and Exclusion Criteria

Following the above search procedures, a total of 9165 records from database searches (through February 2020), 352 records from journal searches (through June 2020), 16 records from author requests, and 65 records from a manual search of reference lists were identified and imported to web-based systematic review software, Rayyan (Ouzzani et al., 2016). To ensure that the included studies were most relevant to our research topic and questions, the following inclusion and exclusion criteria were applied in the title and abstract screening phase. First, participants were students from pre-K to sixth-grade level (or 3–12 years old). For example, we excluded studies that solely focused on infants and secondary or higher education (e.g., Lau, 2012; Lucero & Rouse, 2017; Shin & Partyka, 2017).

Second, studies quantitatively measured teacher language and/or literacy instruction through classroom observations. Accordingly, qualitative case studies (e.g., Reyes, 2006), studies that adopted only teacher or student self-reported practices (e.g., Certo et al., 2010; Ollin, 2008), or studies that focused on observation of teaching in subjects other than language/literacy (e.g., science teaching, Arias et al., 2016; math learning, Sun, 2019) were excluded. Third, studies reported the effect sizes (correlation coefficients or linear regression coefficients) between observational variables (e.g., quality rating, the occurrence of certain teaching practices) and student academic outcomes on language and/or literacy achievement. Some studies investigated teacher effectiveness through quantitative classroom observations, but they did not provide the correlations with students' language and literacy skills (e.g., focusing on the development and validation of observation framework, Kington et al., 2011) or provided teacher/expert-rated students' behavior or language use instead (e.g., Hale et al., 2005; patterns of language use based on observations rather than summative assessments, Markova, 2017). Fourth, studies were reported in English (see Table S1 for a detailed description of data sources).

The detailed screening process is shown in Figure 1. After removing the duplicate records, we screened the title and abstract information for 7858 studies. During the initial screening phase, many studies were excluded due to the absence of classroom observation ($n = 1874$) or academic assessments on language/literacy ($n = 1422$), or not meeting other criteria: method (e.g., qualitative study; $n = 1288$), population ($n = 1091$), subject (e.g., medical research; $n = 983$), and language ($n = 1$). In the full-text screening phase, an additional 647 studies were excluded due to the aforementioned reasons and 408 studies were excluded for missing correlation or regression coefficients. For the studies lacking essential data to compute zero-order correlations or standardized regression coefficients, we contacted the authors to retrieve data information (11 out of 46 provided the data). Taken together, a total of 136 studies met our criteria and 108 of them were eligible for meta-analysis while the remaining 28 studies were narratively reviewed to complement our quantitative analysis. The 28 studies included many unique covariates such as interaction terms and lacked necessary statistical information to compute standardized partial correlations, and therefore could not be included in the meta-analysis.

The title/abstract and full-text screenings were conducted by the authors (trained researchers who have expertise in literature search and literacy education) with adequate reliability (91.24%) established prior to the formal screening process.

Coding Procedures

The coding scheme was developed through an iterative process. First, five studies were randomly selected and coded by the first author to generate preliminary coding categories. Then, we modified and consolidated the coding categories based on a random set of 10 studies. Using the finalized coding scheme, two authors independently coded the remaining studies after they reached an agreement of 92%. Finally, 20% of the included studies were randomly selected for double coding, and the overall interrater reliability was 94% agreement.
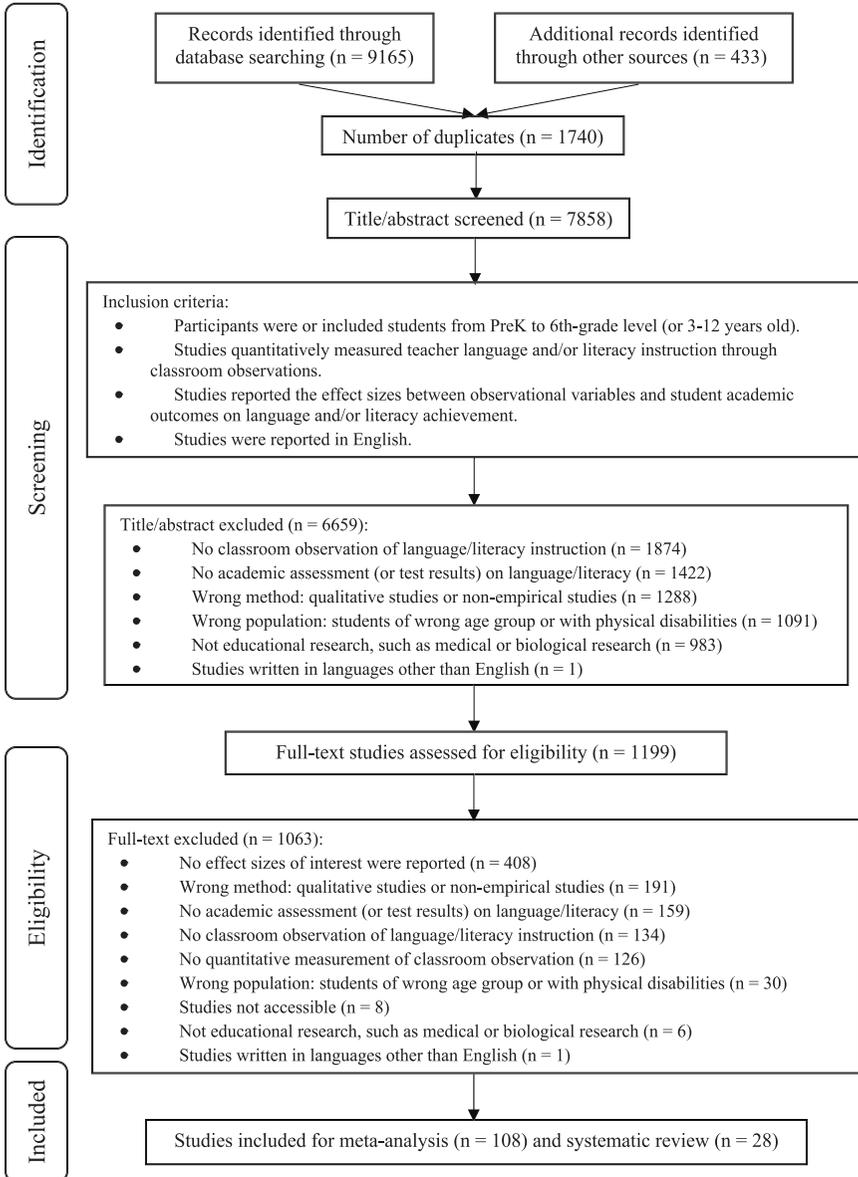
FIGURE 1. *PRISMA chart of screening process.*

The main study information for each study was coded to facilitate statistical analysis (see Appendix A). Specifically, we coded basic study information including authors, publication year and type, study location, and sampling strategy.

Student participants were coded for their race/ethnicity, primary language, maternal education, SES, gender, learning disability status, grade level, and sample size. For teacher participants, we coded their education level, teaching experience, gender, and sample size.

The classroom observation data were coded based on two general categories: macro rating and micro measurement. Under each category, multiple features were coded: measurement point (grade and semester when observation was conducted, duration and frequency of the observation), name of the observation instrument, dimension(s) of the observation (instructional, emotional, structural, mixed; see Table S2 for detailed examples), reliability of the observation instrument, the observer identity (e.g., researcher, teacher), observation mode (e.g., video-based coding, live/real-time rating), interrater/intercoder reliability, the language of instruction, and class size (the number of students per class or teacher–child ratio). With regard to the academic outcome, the measurement point of the test, test name and type (experimental or standardized), assessed skills (reading, language, or writing skills), and test reliability were coded. Lastly, the following were recorded: the correlation coefficients, standardized regression coefficients and corresponding *t*-test statistics, the number and type of associated covariates, and the analytic approach.

Statistical Analysis

We conducted a correlational meta-analysis including both the zero-order correlations and the partial correlations from more complex models. Typically, the correlational meta-analysis extracts only zero-order correlation coefficients. Considering that most classroom research employs multiple and multivariate regression models, eliminating the studies that use complex models would result in a great loss of the studies of interest and thus might misrepresent the population parameters. Therefore, in addition to collecting the zero-order correlations reported in studies, we also collected standardized partial correlations from the linear regression models and analyzed them when the corresponding zero-order correlations were not provided by the authors after contact. For the studies reporting partial correlations, we classified the covariates based on their shared pool of common covariates. Specifically, we identified four types of covariates: student features such as initial achievement level, gender, and race; family characteristics such as parental education, employment, marriage status, and poverty; teacher characteristics such as teacher age, teaching experience, and teacher education level; and classroom features such as class composition and program type. We dummy coded their presence or absence when performing meta-regression (Aloe & Becker, 2012; Aloe & Thompson, 2013).

The effect size indices in this meta-analysis are correlation coefficient, r, and partial correlation coefficient, r_p. Combining zero-order correlations and partial effect sizes in a single data set and analyzing them as a whole would be misleading in terms of the overall effect and level(s) of heterogeneity (Aloe & Thompson, 2013). Therefore, we conducted two separate analyses: one for zero-order correlations and the other for partial effects. The former was prioritized and the latter was performed when the zero-order counterparts could not be retrieved. Furthermore, a narrative synthesis was reported as a supplement to the meta-analysis results.

With regard to moderation analysis, students' grade level, observation frequency, and duration were analyzed in two ways as continuous and then dichotomous variables (grade level below or above fourth grade, observed less or more than three times, observed for less or more than 3 hours per visit), respectively.

All calculations and statistical analyses were conducted in the RStudio open-source software (Version 3.6.3; R Core Team, 2018; Version 1.2.5033; RStudio Team, 2016) using functions available in the metafor package (Version 2.0-0; Viechtbauer, 2010). To ensure that the sampling distributions of *r* values are normal, we applied Fisher's *z* transformation to the data analysis and reconverted *z* values back to *r* values for the presentation and interpretation of the results (Borenstein et al., 2011).

A multilevel random-effects model was used to account for the dependency of the effect sizes with restricted maximum likelihood estimation (Borenstein et al., 2011). Although multiple effect sizes seemed to be nested at the levels of grade, study, and research team, no significant variance was detected at the study or researcher level. Hence, the data set was fitted into a more parsimonious two-level model where individual effect sizes were nested within the grade level. We acknowledge the superiority of robust variance estimation (RVE) for handling dependent effect sizes. However, it has a few important limitations. First, it neither models heterogeneity at multiple levels nor provides corresponding hypothesis tests. Second, the power of the categorical moderator highly depends on the number of studies and features of the covariate (Tanner-Smith, Tipton, & Polanin, 2016). When the number of studies is small, the test statistics and confidence intervals based on RVE can have inflated Type I error (Hedges et al., 2010; Tipton & Pustejovsky, 2015). Relating to our cases, many of our moderators had imbalanced distributions (e.g., see Tables 1 and 3; such as observational dimension and outcome type where some had over 100 cases and some had less than 20). Consequently, tests of particular moderators may be severely underpowered. Given these limitations, we prioritized the multilevel meta-analysis given that many studies contained independent groups of students from different grades and it could meet the goals of operating heterogeneity, moderation, and sensitivity analyses that are not currently available for RVE. In addition, we adopted RVE in sensitivity analysis for robustness check.

The homogeneity statistic Q was computed to assess the variation in correlations between studies (Lipsey & Wilson, 2001), which is usually supplemented by the I^2 statistic for quantifying the percentage of variation across studies due to real heterogeneity (beyond sampling error). By convention, I^2 values range from 0% to 100% and indicate higher levels of heterogeneity with increasing percentages: small = 25%, moderate = 50%, and high = 75% (Higgins et al., 2003). Additionally, moderation analysis was conducted to explain the between-study differences once significant and substantial heterogeneity was confirmed.

Results

Descriptive Information

A total of 136 studies (N = 107,882 participants) were included and reviewed in this study. Of the 108 studies that were eligible for meta-analysis, 70 studies

provided zero-order correlations (k = 736) and 38 studies provided only partial correlations (k = 384). The remaining 28 studies were narratively synthesized.

In all, the majority of the studies (n = 95) were conducted in the United States and used English for instruction, whereas a few of them (n = 13) were conducted in Europe, East Asia, and South America. Most studies (n = 95) focused only on preschoolers and students from lower grades (up to grade 3), and only 13 studies contained students from higher grades (grades 4 to 6). Although the gender distribution was generally balanced, students were of diverse social backgrounds (see detailed maternal education, family income, teacher education level, and class size information in Table S3). Table S4 presents the names of the observation instruments, observer identity, observation mode, and the academic assessments used in the included studies together with their corresponding reliability. Overall, approximately 58% of the included studies reported the observation instruments' reliabilities, ranging from .61 to .98. Most classes were observed live while a small portion relied on field notes or video coding (n = 24). Approximately 81% of the studies reported the interrater reliability between observers, and the range was from .58 to .99. With regard to the assessments, the majority were standardized tests. Approximately 78% of the studies reported reliabilities that ranged from 0.24 to 1.00.

RQ1. What is the relation of observed classroom practices to students' language and literacy performance from pre-K to sixth grade?

As shown in Figure 2, the overall magnitude of the zero-order correlation between observed classroom practices and students' language/literacy outcomes was weak but significant (r = .12; SE = .02; 95% CI [.09, .15]; p < .001). The heterogeneity test result was significant (Q = 6249.38, df = 735, p < .001), and the I^2 value suggested that over 88% of the total variance could be attributed to between-study differences. With regard to the partial correlations (after controlling for the features at the student, family, teacher, and/or classroom level), the correlation between observed classroom practices and students' language/literacy outcomes was minimal though still statistically significant (r_p = .04; SE = .01; 95% CI [.03, .06]; p < .001). The heterogeneity test was also significant (Q = 1246.78, df = 383, p < .001) whereas the I^2 value was reduced to 69%.

RQ2. Does the relation vary by student grade level, observation type and dimension, observation duration and frequency, child language and literacy outcomes, adopted statistical approach, and covariates?

Given the substantial heterogeneity among the effect sizes, we conducted moderation analysis for grade level, observation type and dimension, observation frequency and duration, and child language and literacy outcomes (Tables 1 and 2). Because grade level had already been controlled in many studies reporting partial correlations, we did not carry out a moderation analysis of grade level for this set of studies (see Table 3).
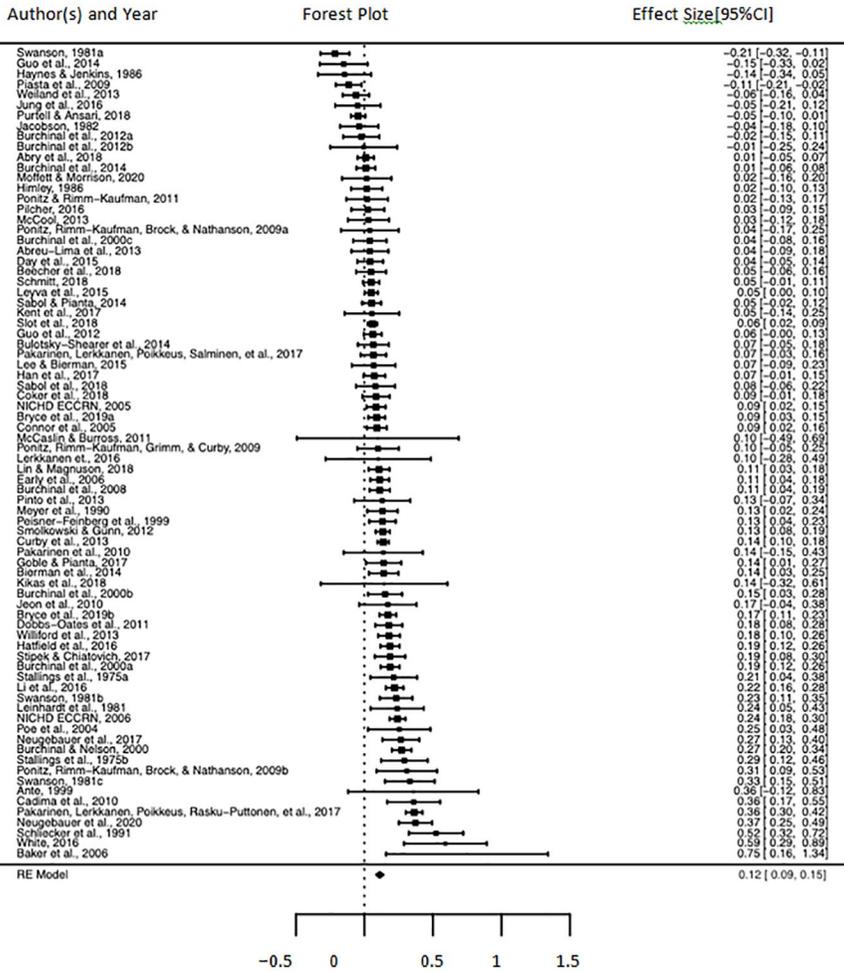
Author(s) and Year | Forest Plot | Effect Size[95%CI]

Author(s) and Year	Effect Size [95%CI]
Swanson, 1981a	−0.21 [−0.32, −0.11]
Guo et al., 2014	−0.15 [−0.33, 0.02]
Haynes & Jenkins, 1986	−0.14 [−0.34, 0.05]
Piasta et al., 2009	−0.11 [−0.21, −0.02]
Weiland et al., 2013	−0.06 [−0.16, 0.04]
Jung et al., 2016	−0.05 [−0.21, 0.12]
Purtell & Ansari, 2018	−0.05 [−0.10, 0.01]
Jacobson, 1982	−0.04 [−0.18, 0.10]
Burchinal et al., 2012a	−0.02 [−0.15, 0.11]
Burchinal et al., 2012b	−0.01 [−0.25, 0.24]
Abry et al., 2018	0.01 [−0.05, 0.07]
Burchinal et al., 2014	0.01 [−0.06, 0.08]
Moffett & Morrison, 2020	0.02 [−0.16, 0.20]
Himley, 1986	0.02 [−0.10, 0.13]
Ponitz & Rimm-Kaufman, 2011	0.02 [−0.13, 0.17]
Pilcher, 2016	0.03 [−0.09, 0.15]
McCool, 2013	0.03 [−0.13, 0.18]
Ponitz, Rimm-Kaufman, Brock, & Nathanson, 2009a	0.04 [−0.17, 0.25]
Burchinal et al., 2000c	0.04 [−0.08, 0.16]
Abreu-Lima et al., 2013	0.04 [−0.09, 0.18]
Day et al., 2015	0.05 [−0.04, 0.14]
Beecher et al., 2018	0.05 [−0.06, 0.16]
Schmitt, 2018	0.05 [−0.01, 0.11]
Leyva et al., 2015	0.05 [0.00, 0.10]
Sabol & Pianta, 2014	0.05 [−0.02, 0.12]
Kent et al., 2017	0.05 [−0.14, 0.25]
Slot et al., 2018	0.06 [0.02, 0.09]
Guo et al., 2012	0.06 [−0.00, 0.13]
Bulotsky-Shearer et al., 2014	0.07 [−0.05, 0.18]
Pakarinen, Lerkkanen, Poikkeus, Salminen, et al., 2017	0.07 [−0.03, 0.16]
Lee & Bierman, 2015	0.07 [−0.09, 0.23]
Han et al., 2017	0.07 [−0.01, 0.15]
Sabol et al., 2018	0.08 [−0.06, 0.22]
Coker et al., 2018	0.09 [−0.01, 0.18]
NICHD ECCRN, 2005	0.09 [0.02, 0.15]
Bryce et al., 2019a	0.09 [0.03, 0.15]
Connor et al., 2005	0.09 [0.02, 0.16]
McCaslin & Burross, 2011	0.10 [−0.49, 0.69]
Ponitz, Rimm-Kaufman, Grimm, & Curby, 2009	0.10 [−0.05, 0.25]
Lerkkanen et., 2016	0.10 [−0.28, 0.49]
Lin & Magnuson, 2018	0.11 [0.03, 0.18]
Early et al., 2006	0.11 [0.04, 0.18]
Burchinal et al., 2008	0.11 [0.04, 0.19]
Pinto et al., 2013	0.13 [−0.07, 0.34]
Meyer et al., 1990	0.13 [0.02, 0.24]
Peisner-Feinberg et al., 1999	0.13 [0.04, 0.23]
Smolkowski & Gunn, 2012	0.13 [0.08, 0.19]
Curby et al., 2013	0.14 [0.10, 0.18]
Pakarinen et al., 2010	0.14 [−0.15, 0.43]
Goble & Pianta, 2017	0.14 [0.01, 0.27]
Bierman et al., 2014	0.14 [0.03, 0.25]
Kikas et al., 2018	0.14 [−0.32, 0.61]
Burchinal et al., 2000b	0.15 [0.03, 0.28]
Jeon et al., 2010	0.17 [−0.04, 0.38]
Bryce et al., 2019b	0.17 [0.11, 0.23]
Dobbs-Oates et al., 2011	0.18 [0.08, 0.28]
Williford et al., 2013	0.18 [0.10, 0.26]
Hatfield et al., 2016	0.19 [0.12, 0.26]
Stipek & Chiatovich, 2017	0.19 [0.08, 0.30]
Burchinal et al., 2000a	0.19 [0.12, 0.25]
Stallings et al., 1975a	0.21 [0.04, 0.38]
Li et al., 2016	0.22 [0.16, 0.28]
Swanson, 1981b	0.23 [0.11, 0.35]
Leinhardt et al., 1981	0.24 [0.05, 0.43]
NICHD ECCRN, 2006	0.24 [0.18, 0.30]
Poe et al., 2004	0.25 [0.03, 0.48]
Neugebauer et al., 2017	0.27 [0.13, 0.40]
Burchinal & Nelson, 2000	0.27 [0.20, 0.34]
Stallings et al., 1975b	0.29 [0.12, 0.46]
Ponitz, Rimm-Kaufman, Brock, & Nathanson, 2009b	0.31 [0.09, 0.53]
Swanson, 1981c	0.33 [0.15, 0.51]
Anto, 1999	0.36 [−0.12, 0.83]
Cadima et al., 2010	0.36 [0.17, 0.55]
Pakarinen, Lerkkanen, Poikkeus, Rasku-Puttonen, et al., 2017	0.36 [0.30, 0.42]
Neugebauer et al., 2020	0.37 [0.25, 0.49]
Schliecker et al., 1991	0.52 [0.32, 0.72]
White, 2016	0.59 [0.29, 0.89]
Baker et al., 2006	0.75 [0.16, 1.34]
RE Model	0.12 [0.09, 0.15]

−0.5   0   0.5   1   1.5

FIGURE 2. *Forest plot for zero-order correlation.*

Grade Level

We first treated grade level as a continuous variable in the moderation analysis and used (weighted) average grade level for the aggregated effect size associated with multiple grades. A significant and negative relation of grade level on the overall zero-order correlation was found between observed classroom practices and students' language/literacy outcomes ($b = −.03$, $p < .001$). Furthermore, this negative relation was consistent after controlling for the observation type and dimension, and language and literacy outcomes (see Table 2), indicating that the relation was getting weaker ($−.03$) with the increase of grade level.

TABLE 1

Multilevel random effects model: Meta-regression of moderators on zero-order correlation.

Moderator	k	Intercept(SE)	b(SE)	Q(df)
Grade level	736	.17***(.02)	−.03***(.01)	37.96***(1)
Grade level dichotomous (<grade 4)[a]	655	.13***(.02)	−.09(.06)	2.83(1)
Observation type (micro measurement)[a]	299	.05**(.02)		43.06***(1)
Macro rating	437		.09***(.01)	
Observation dimension (instructional)[a]	401	.12***(.02)		34.35***(3)
Emotional	186		−.02***(.00)	
Structural	18		−.03*(.01)	
Mixed	131		.03**(.01)	
Observation frequency	698	.11***(.02)	.00(.00)	0.46(1)
Observation frequency dichotomous (<3 visits)[a]	698	.11***(.02)	.01(.04)	.10(1)
Observation duration dichotomous (<3 hours/visit)[a]	639	.09***(.02)	.05(.03)	2.68(1)
Outcome type (reading)[a]	520	.12***(.02)		6.93(3)
Language	157		.00(.00)	
Writing	28		.00(.02)	
Mixed	31		−.03**(.01)	

Note. For grade level, observation frequency, and observation duration, models were fitted treating them as continuous and dichotomous variables, respectively. Those unmarked are continuous variables.

[a]Reference group for moderator in parentheses.

k = number of effect sizes.

*p < .05; **p < .01; ***p < .001.

We then treated grade level as a dichotomous variable (preK3rd versus 4th–6th grade) given the shift in instructional focus from learning to read in lower grades to reading to learn in grade 4. The relation was slightly weaker in upper grades, but it was not statistically significant ($b = -.09$, $p = .06$). This implied that the overall correlation between instruction and student outcome did not vary by primary grades versus upper elementary grades. However, this finding should be taken with caution because our effect sizes were predominantly clustered in lower grades.

Observation Type and Dimension

As shown in Table 1, when using zero-order correlation data, we found a slightly stronger relation for the studies using macro-level observations with quality ratings ($b = .09$, $p < .001$) than those quantitatively measuring discrete

TABLE 2

Multilevel random effects model: Meta-regression of moderators controlling for grade level.

	Model 1	Model 2	Model 3
Intercept(*SE*)	.10***(.02)	.17***(.02)	.17***(.02)
Grade level	−.03***(.01)	−.03***(.01)	−.03***(.01)
Observation type (micro measurement)[a]			
Macro rating	.08***(.01)		
Observation dimension (instructional)[a]			
Emotional		−.02***(.01)	
Structural		−.04*(.01)	
Mixed		.02(.01)	
Outcome type (reading)[a]			
Language			.00(.00)
Writing			.00(.02)
Mixed			−.03*(.01)
k	736	736	736

Note. [a]Reference group for moderator in parentheses.
k = number of effect sizes.
p < 0.05; **p* < 0.01; ***p* < 0.001.

practices at a microlevel. Moreover, the overall zero-order relation between observed classroom practices and students' language/literacy outcomes was stronger for the instructional dimension compared to the emotional dimension ($b = -.02$, $p < .001$) or the structural dimension ($b = -.03$, $p < .05$). In other words, the correlations with student outcomes were as follows: .12 for the instructional dimension (see intercept in Table 1), .10 for the emotional dimension, and .09 for the structural dimension. When using partial correlation data (see Table 3), no significant moderating effect was found for observation type on the overall partial correlation ($b = .004$, $p = .63$). In contrast, a stronger partial correlation was found for the instructional dimension than the emotional dimension ($b = -.01$, $p < .001$) or the structural dimension ($b = -.06$, $p < .001$). The structural dimension only contained 10 effect sizes whereas the instructional dimension contained over 200 cases, so the negative result on the former one might have inflated Type I error and should be taken with caution.

Observation Frequency and Duration
   There was no significant moderating effect of either observation frequency or duration (see Tables 1 and 3). We treated frequency as a continuous and then a dichotomous variable (whether or not the class was observed less than three times), and the overall zero-order correlation did not vary by the number of visits (continuous: $b = .003$, $p = .50$; dichotomous: $b = .01$, $p = .75$) or the observation duration (whether or not the class was observed for less than 3 hours) per visit ($b = .05$, $p = .10$). Similar nonsignificant findings were observed in partial

TABLE 3

Multilevel random effects model: Meta-regression of moderators on partial correlation.

Moderator	k	Intercept(SE)	b(SE)	Q(df)
Observation type (micro measurement) [a]	179	.04***(.01)		.23(1)
Macro rating	205		.00 (.01)	
Observation dimension (instructional) [a]	227	.05***(.01)		65.96***(3)
Emotional	101		-.01***(.00)	
Structural	10		-.06***(.01)	
Mixed	46		.01(.01)	
Observation frequency	323	.04**(.01)	.00(.00)	.90(1)
Observation frequency dichotomous (<3 visits) [a]	323	.04***(.01)	.02(.01)	2.77(1)
Observation duration dichotomous (<3 hours/vsit) [a]	322	.05***(.01)	.01(.01)	.16(1)
Outcome type (reading)[a]	264	.04***(.01)		23.08***(3)
Language	106		-.00*(.00)	
Writing	11		.08***(.02)	
Mixed	3		-.01 (.19)	
Covariates	384	.01(.04)		7.24(4)
Student features	330		.06(.04)	
Family features	184		-.03(.01)	
Teacher features	87		.02(.02)	
Class features	264		-.02(.01)	
Analytic approach (multilevel model)[a]	293	.05***(.01)		2.66(2)
Multiple regression	31		-.02(.02)	
Latent growth model	60		-.04(.03)	

Note. For observation frequency, and observation duration, models were fitted treating them as continuous and dichotomous variables, respectively. Those unmarked are continuous variables.

[a]Reference group for moderator in parentheses.

k = number of effect sizes.

*p < 0.05; **p < 0.01; ***p < 0.001.

correlations: frequency (continuous: $b = .004$, $p = .34$; dichotomous: $b = .02$, $p = .10$) and duration ($b = .01$, $p = .69$).

Child Language and Literacy Outcomes

Compared to reading skills (see Table 1), the zero-order correlation was weaker for mixed skills ($b = -.03$, $p < .01$). However, the partial correlation data showed a different pattern: weaker relations between classroom observations and language skills ($b = -.004$, $p < .05$), and stronger relations for writing skills ($b = .08$, $p < .001$) than reading skills (see Table 3).

Covariates and Analytic Approach

Based on the nature of the covariates and the analytic approaches employed in the studies reporting partial correlations, we classified four types of covariates (student, family, teacher, and class features) and three types of analysis (multilevel model, multiple regression, and latent growth model). We found that neither of them significantly moderated the overall partial correlation between observed classroom practices and students' language/literacy outcomes (see Table 3).

Sensitivity Analysis

Prior to the substantive meta-analysis, diagnostic tests for outliers, influential cases, publication bias, and potential threats from studies with lower quality and small sample sizes were performed in an effort to justify the robustness of our analysis.

To identify the potential outliers and influential cases, we plotted the studentized residuals, Cook's distances, and covariance ratios of our main model (i.e., overall zero-order correlation estimation). Four studies were consistently identified as unusual cases (see Figure S1). However, refitting the model without the four studies still led to essentially the same overall correlation: $r = .12$; $SE = .01$; 95% CI [.09, .14]; $p < .001$. Hence, our pooled estimate did not hinge on these unusual cases.

To visualize and statistically determine the existence of publication bias, we performed the funnel plot and Egger's regression test (Sterne & Egger, 2006). The distribution of our data points was symmetric by and large with a few missing to the left, especially near the bottom, suggesting a lack of smaller effect sizes associated with studies of lower precision (see Figure S2). Likewise, the Egger's test indicated that the intercept significantly deviated from zero ($b = .07$, $p < .01$), suggesting the existence of publication bias.

Moreover, we evaluated the quality of the included studies in terms of the research question, study population, classroom observation, outcome report, and statistical analysis. We rated the overall quality of each study as "strong," "moderate," or "weak" using the criteria shown in Appendix B (National Heart, Lung, and Blood Institute, n.d.). The comparison between the models fitted with and without the effect sizes associated with the "weak" studies was conducted and a nonsignificant difference was detected ($b = .002$, $p = .98$). Therefore, including lower-quality studies would not distort our pooled estimate.

Additionally, we accounted for both hierarchical effects cases (i.e., multiple studies nested within a larger cluster) and correlated effects cases (i.e., the same participant group provides multiple effect sizes) with small-sample corrections using robust variance estimation (RVE). We ran RVE with the Robumeta package (version 2.0; Fisher et al., 2017). The RVE meta-analysis confirmed a weak but significant overall zero-order correlation, $r = .11$, $SE = .01$, 95% CI [.08, .14], $p < .001$, and a substantial heterogeneity, $I^2 = 88\%$.

We had 11 studies that reported partial correlations originally but provided zero-order correlations after we contacted the authors. For these studies, we refitted the main model (i.e., overall partial correlation estimation) by adding back these original partial correlations. The overall partial correlation was similar to the model results without these effect sizes, $r_p = .04$, $SE = .01$, 95% CI [.03, .06], $p < .001$, with a substantial heterogeneity, $I^2 = 68\%$. Thus, we retained the 11 studies as part of the zero-order correlation data.

In sum, our main analysis of the overall zero-order correlation between observed classroom practices and student language/literacy outcomes was not subject to any potential outliers/influential cases, study quality, or small sample issues except publication bias. Also, removing the studies that provided zero-order correlations upon author request from the partial correlation analysis did not statistically affect the main estimation for partial correlation either.

Narrative Synthesis

A total of 28 studies (see online supplementary materials) were included in the narrative analysis. Overall, the majority of the studies reported weak positive partial correlations between classroom practices and students' language/literacy outcomes (e.g., Gersten et al., 2010; Howes et al., 2008; Kwan et al., 1998) whereas some studies reported moderate to strong positive partial correlations (e.g., McCartney, 1984; McIntosh et al., 2007) and a few reported weak negative partial correlations (e.g., Crocker & Brooker, 1986; Howes et al., 2008). Aligning with our meta-analysis results mentioned previously, Crocker and Brooker (1986) reported that the relation between classroom practices and students' achievement was weaker for higher grades. Furthermore, there was a general tendency that the studies using macro-level observations with quality ratings yielded more stable positive relations (Connor et al., 2014; Kwan et al., 1998; McCartney, 1984; McIntosh et al., 2007; Pianta, Belsky, et al., 2008) whereas those using quantitative measurements of microlevel practices showed either negative or relatively weaker positive relations (Crocker & Brooker, 1986; Gersten et al., 2010). In line with the meta-analysis results, studies also showed that the association between classroom practices and student outcomes was consistently stronger when the observation was focused on the instructional dimension rather than the emotional or structural ones (Crocker & Brooker, 1986; Guo, Justice, et al., 2012; Howes et al., 2008).

In addition, these studies suggested several notable moderation effects. First, the direct and indirect relations between classroom practices and student achievement might differ by the language/literacy outcome of interest (Baroody & Diamond, 2016). For example, Kwan and colleagues (1998) found that child center quality was related to students' verbal fluency but not word reading. McCartney

(1984) reported that the strength of the relation was stronger for mixed measures incorporating both reading and oral language components than for measures examining only oral language.

Second, there was an interaction between observation types or dimensions in predicting students' language/literacy outcomes (Connor et al., 2014; Guo, Justice, et al., 2012; Pianta, Belsky, et al., 2008). Students' gains in vocabulary and comprehension were greater not only when teachers provided a high-quality classroom learning environment but also when they spent greater amounts of time on meaning-focused instruction in small groups (Connor et al., 2014). In addition, the classroom physical literacy environment (i.e., presence of writing materials) was positively related to children's growth in alphabet knowledge and name-writing ability only in the context of high-quality, instructional supportive classrooms (Guo, Justice, et al., 2012). It was also found that the negative relation between the quantity of time spent on reading instruction and improvement in reading was mitigated when there was a higher level of emotional support in the classroom (Pianta, Belsky, et al., 2008).

Third, there was a moderation between classroom quality and student characteristics in predicting student achievement (Gosse et al., 2014; Vitiello et al., 2012). The relation between instructional support and language development was stronger for those children who had higher initial language skills (Gosse et al., 2014). It was also found that high emotional support was more positively associated with children's gains in language/literacy for children who were resilient than those who were overcontrolled (Vitiello et al., 2012).

Lastly, teachers and peers also influenced the association between classroom practices and student achievement (Guo et al., 2011; Mashburn et al., 2009). For example, Guo and colleagues (2011) demonstrated that teachers' sense of collegiality (i.e., collaboration among teachers within schools, implying shared responsibility and commitment to common educational goals) in combination with higher language and literacy instructional quality, predicted greater gains in children's vocabulary scores. Another example is Mashburn and colleagues' (2009) study, where better-managed classrooms had stronger relations between peers' expressive skills and children's growth in receptive language.

Discussion

The purpose of this study was to summarize and characterize the relation between observed classroom practices and children's language/literacy achievement. Because of the variety of the models and the analytic approaches employed in the reviewed studies, we analyzed both zero-order and partial correlations and narratively synthesized studies containing interaction terms or indirect effects. This meta-analysis did not focus on the relation to student academic *growth* over the school year, but rather on status outcomes, assuming a level playing field at the onset.

First and foremost, aligning with our hypotheses—a weak association between the classroom observation and student language/literacy outcomes, and the association expected to be stronger for younger students than older students (Burchinal et al., 2011)—we found that the observed classroom practices were significantly but weakly associated with the language/literacy outcomes of students from pre-K

to sixth grade ($r = .12$ & $r_p = .04$). The overall zero-order relation was stronger for younger children regardless of the observation type, observation dimension, or outcome type. These findings overall indicate that although classroom observations are widely used, they do not explain a large amount of variation in students' language/literacy achievement, and this is particularly pronounced for students in upper elementary grades (Brunsek et al., 2017; Burchinal et al., 2011; Keys et al., 2013; Perlman et al., 2016; Ulferts et al., 2019). In other words, few observational measures target specifically language/literacy learning, and extant observation instruments are limited in predicting student language/literacy outcomes.

There are several possibilities to explain the overall weak association. As suggested in Perlman et al. (2016), the weak association could reflect the unneglectable impact of family and other factors. In fact, compared to the zero-order correlation estimation, our smaller pooled partial correlation and its corresponding moderate heterogeneity estimate implied that on top of the teacher/child classroom behaviors and interactions, a multitude of student, family, teacher as well as class features could predict the teaching and learning outcomes. According to our narrative review, there were multiple moderation effects among different observation types, dimensions, and individual characteristics, which showed that the relation between classroom quality and students' gains in language/literacy is a function of a complex set of factors. In addition, the measurement issues regarding the difference in the measurement unit for classroom observation and student performance should be noted. Specifically, the majority of studies reported observation results at the classroom level, not at the child level. Given that students differ in the extent to which they engage in and learn from the same instruction, observation results at the classroom level are not precise estimates of student learning.

There is another mismatch between the scope and specificity of the classroom observation and testing construct. Researchers have pointed out the validity issues inherent in present observation instruments as the majority were developed by child development experts based on conceptual rather than psychometric considerations. Consequently, although they capture a broad context (i.e., general interpersonal interaction and environmental/structural provisions), there is a lack of focus on cognitive and academic skills (Burchinal et al., 2011). Also echoing Connor's statement (2013, p. 4) "observation tools are most useful when developed to serve a particular purpose and are put to that purpose," it is possible that more focused observation on a particular dimension would yield a stronger relation to its targeting academic outcome (e.g., the relation between instructional quality or time spent on spelling and its corresponding spelling outcome). Ulferts and colleagues (2019) found instruments that measured teacher–child interactions outperformed those measuring material–spatial surroundings (more of a precondition for quality teaching) in capturing what matters for student learning. However, there are findings showing that more positive associations were found between ECERS/ECERS-R global scores and student outcomes as compared to its subscale scores (e.g., teaching and interactions, provisions for learning; Brunsek et al., 2017). In sum, efforts on improving observation validity are

warranted and the tradeoff between broadly and narrowly focused observation scopes is well worth considering.

In line with the above findings and our hypothesis of a varied relation by the features of observation and assessed skills (Brunsek et al., 2017; Keys et al., 2013; Perlman et al., 2016; Ulferts et al., 2019), our results from zero-order and partial correlation data showed a stronger relation between observed classroom practices and student language/literacy outcomes for studies capturing the instructional dimension (quantity and quality of literacy content delivery, cognitive explanation and monitoring, stimulation and feedback) than the emotional dimension (classroom climate and organization, praise and discipline, behavior management, sensitivity/responsivity/detachment/disengagement) or the structural dimension (classroom physical environment, book category, writing materials category). These findings suggest that for students' academic outcomes, variation in the instructional dimension matters more than that in the emotional and structural dimensions. Despite statistically significant moderation effects, these findings should be taken with caution because the difference in magnitude was small and the relevant measures varied in reliability (see instrument reliability and observer reliability coefficients in Table S4).

In addition, the types of student outcomes—language, reading, or writing—showed consistently significant moderating effects in both sets of analyses, but the patterns of results were opposite for zero-order data versus partial correlations. We found a larger effect size associated with comprehensive tests (assessing mixed skills) for zero-order correlations whereas a larger effect size was associated with writing tests among partial correlations. Keeping in mind that the two sets of analyses contained different samples and used diverse observation measurements as well as child language and literacy outcomes, the two sets of analyses are not comparable, but the overall findings underscore that the relation is contingent on both the observation dimension and child outcomes.

Our meta-analysis findings did not show a varied relation by observation frequency and duration, type of covariates, or analytic approach, although the latter two had significant moderating effects in Ulferts and colleagues' (2019) study. There might be a couple of explanations. For example, observation frequency and duration were congruent and limited among the reviewed studies (most researchers observed twice or three times in total with 2 to 3 hours per visit despite a few extreme values that extended the overall ranges) such that their moderating effects may have been underestimated. Another reason is the variation in covariates included in studies. There was a great deal of variability among the covariates, and we broadly classified them into four types, which might have disguised some significant factors such as baseline skills and teacher education level. More consistent approaches across studies would help illuminate consistency in findings.

Limitations and Future Directions

There are several limitations of this study. First, although the aim of this review was to investigate the relation between observed classroom practices and students' language/literacy outcomes from pre-K to sixth grade, more than half of the studies meeting our criteria focused on a lower grade level (pre-K to third grade). Therefore, the generalization of our results to upper-grade levels is more limited. These results

indicate a need for more classroom research beyond early primary grade levels. Second, as mentioned earlier, the complexity and variability of the employed observation instruments and child outcomes rendered a relatively broad classification that might also obscure many findings. For example, we classified observational instruments according to their evaluation approaches and instructional content, which only allowed us to test the global differences among classroom observations. Another caveat was that too few studies assessed writing and mixed skills in comparison to the number of studies that assessed reading skills though we detected differential relations among studies assessing reading versus those assessing mixed or writing skills. Third, many studies did not report information on potential moderators such as teachers' years of teaching, education level, class size, students' learning disability status, and SES (see Table S3). Therefore, we could not explore whether relations differ by these features. Lastly, it should be noted that the pooled zero-order and partial correlations were generated from independent samples so they are not directly comparable. Hence, our comparative findings from zero-order versus partial correlations should be taken with caution. In addition, for the studies that provided both zero-order correlations and partial correlations, we prioritized zero-order correlations in the data analysis phase. We recognize that zero-order correlation does not account for the teaching impact on student academic growth over the school year because it does not account for students' previous skill levels. Finally, this study cannot draw any conclusion about whether the instrument captures or fails to capture the latent construct of teaching or teaching practices. This is an important validity question that should be addressed in the original studies. Meta-analysis is a powerful tool but it relies on the quality of the original studies that are included. In this review, many included studies did not report validity information. Future reviews may need to further account for the validity and reliability of observation instruments.

Conclusion

This study adds to our understanding of the relation between observed classroom practices and language/literacy development for elementary students. Our study found a significant but weak association on average and brought to light several challenges in synthesizing various classroom research. Extant observation instruments differ in the observational scope, dimensions, and purposes. The present findings indicate that classroom observations provide only a limited picture of students' language and literacy skills. Of course, this does not deny the importance of attending to teaching practices in the classrooms. Instead, what the results suggest is a need for a comprehensive picture of the factors that influence students' language and literacy skills, including student factors (e.g., traits and behaviors) and their language and literacy environments and resources in the home and community as well as classroom instruction.

Classroom instruction is a complex construct that requires a highly reliable and valid observation system that can capture the complexity and its relation to students' language and literacy development. In addition, we highlighted the need for more consistent measurement approaches across studies in order for the field to develop a clearer picture of the relation between classroom practices and student achievement.

Appendix A

Coding Scheme

- Study ID
- Authors
- APA citation
- Publication type: dissertation, journal article, unpublished manuscript
- Study location
- Sampling strategy
- Student participants
  - Race/ethnicity: %Caucasian, African American, Hispanic, Asian or Pacific Islander, Native American, multiracial, other
  - Primary language: %English, Spanish, other, English language learner
  - Maternal/caregiver education: %Under high school, high school, associate, bachelor, master or above, missing
  - SES: %low, mid, high
  - Gender: %boy, girl
  - Learning disability status: %average, learning disability, etc.
  - Grade (age)
  - Sample size
- Teacher participants
  - Teacher education: %high school, associate, bachelor, above bachelor
  - Teaching experience: mean years
  - Gender: %male, female
  - Under PD program: Yes or no
  - Sample size
- Classroom observation
  - Macro rating
    - Timing (grade-semester, duration, frequency)
    - Observation instrument
    - Rating dimension(s)
    - Reliability of observation instrument
    - Observer
    - Observation mode (video recording, live, etc.)
    - Inter-rater/coder reliability
  - Micro measurement
    - Timing (grade-semester, duration, frequency)
    - Observation instrument
    - Amount/ratio (the amount of time, the proportion or frequency of certain practice)
    - Observation dimension(s)
    - Reliability of observation instrument
    - Observer
    - Observation mode (video recording, live, etc.)
    - Inter-rater/coder reliability
  - Language of instruction: English, Spanish, etc.
  - Class size

- Academic outcome
  - Timing (grade-semester)
  - Test name
  - Test type (experimental or standardized)
  - Test skills
  - Test reliability
- Effect sizes
  - Correlation or regression coefficients with t-test results
  - Number of covariates and their type
  - Analytic approach

Appendix B

Study Quality Evaluation Scale.

1. Was the research question or objective clearly stated?	Yes	Somewhat	No
2. Was the study population clearly defined?	Yes	Somewhat	No
3. Was the sample selected from the same or similar populations and clearly described (i.e., age, gender, race, disabilities)?	Yes	Somewhat	No
4. Were inclusion and exclusion criteria for being in the study prespecified and applied uniformly to all participants?	Yes	Somewhat	No
5. Was sample size justification, power description, or variance and effect estimates provided?	Yes	Somewhat	No
6. For longitudinal studies, was attrition (withdrawals and drop-outs) reported in terms of numbers and/or reasons per group?	Yes	Somewhat	No
7. Were the classroom observation procedures clearly defined across all study participants?	Yes	Somewhat	No
8. Were the classroom observation procedures reliable across all study participants?	Yes	Somewhat	No
9. Were the classroom observations implemented consistently across all study participants?	Yes	Somewhat	No
10. Were the academic assessments clearly defined across all study participants?	Yes	Somewhat	No
11. Were the academic assessments reliable across all study participants?	Yes	Somewhat	No
12. Were the academic assessments implemented consistently across all study participants?	Yes	Somewhat	No
13. Did they report basic information (i.e., mean, standard deviations, range, correlations for all assessments given)?	Yes	Somewhat	No
14. Did they report reliability estimates on all measures? And, if so, were they above .7?	Yes	Somewhat	No
15. Were the statistical measures appropriate for the study design?	Yes	Somewhat	No
Rating criteria			
Strong		>13	
Moderate		13	
Weak		<13	

ORCID iDs

Yucheng Cao [iD] https://orcid.org/0000-0002-4842-0667
Young-Suk Grace Kim [iD] https://orcid.org/0000-0002-4328-3843

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Abreu-Lima, I. M. P., Leal, T. B., Cadima, J., & Gamelas, A. M. (2013). Predicting child outcomes from preschool quality in Portugal. *European Journal of Psychology of Education*, *28*(2), 399–420. https://doi.org/10.1007/s10212-012-0120-y

*Abry, T., Granger, K. L., Bryce, C. I., Taylor, M., Swanson, J., & Bradley, R. H. (2018). First grade classroom-level adversity: Associations with teaching practices, academic skills, and executive functioning. *School Psychology Quarterly*, *33*(4), 547. http://dx.doi.org/10.1037/spq0000235

Ackerman, W. I. (1954). Teacher competence and pupil change. *Harvard Educational Review*, *24*, 273–289.

Aloe, A. M., & Becker, B. J. (2012). An effect size for regression predictors in meta-analysis. *Journal of Educational and Behavioral Statistics*, *37*(2), 278–297. https://doi.org/10.3102/1076998610396901

Aloe, A. M., & Thompson, C. G. (2013). The synthesis of partial effect szes. *Journal of the Society for Social Work and Research*, *4*(4), 390–405. https://doi.org/10.5243/jsswr.2013.24

*Ante, B. C. (1999). *An exploratory study on the implementation of collaborative learning environments: Challenges and constraints* [Unpublished doctoral dissertation]. The Catholic University of America.

Arias, A. M., Bismack, A. S., Davis, E. A., & Palincsar, A. S. (2016). Interacting with a suite of educative features: Elementary science teachers' use of educative curriculum materials. *Journal of Research in Science Teaching*, *53*(3), 422–449. https://doi.org/10.1002/tea.21250

*Auger, A., Farkas, G., Burchinal, M. R., Duncan, G. J., & Vandell, D. L. (2014). Preschool center care quality effects on academic achievement: An instrumental variables analysis. *Developmental Psychology*, *50*(12), 2559–2571. https://doi.org/10.1037/a0037995

*Baker, S. K., Gersten, R., Haager, D., & Dingle, M. (2006). Teaching practice and the reading growth of first-grade English learners: Validation of an observation instrument. *The Elementary School Journal*, *107*(2), 199–220. https://doi.org/10.1086/510655

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, *52*(1), 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

*Barnett, W. S., Yarosz, D. J., Thomas, J., Jung, K., & Blanco, D. (2007). Two-way and monolingual English immersion in preschool education: An experimental comparison. *Early Childhood Research Quarterly*, *22*(3), 277–293. https://doi.org/10.1016/j.ecresq.2007.03.003

Baroody, A. E., & Diamond, K. E. (2016). Associations among preschool children's classroom literacy environment, interest and engagement in literacy activities, and early reading skills. *Journal of Early Childhood Research*, *14*(2), 146–162. https://doi.org/10.1177/1476718X14529280

*Beecher, C. C., Strand, P., & French, B. F. (2018). Investigation of the development of pre-academic skills for preschoolers in Head Start. *Journal of Education for Students Placed at Risk (JESPAR)*, *23*(3), 230–249. https://doi.org/10.1080/10824669.2018.1477601

*Bierman, K. L., Nix, R. L., Heinrichs, B. S., Domitrovich, C. E., Gest, S. D., Welsh, J. A., & Gill, S. (2014). Effects of Head Start REDI on children's outcomes 1 year later in different kindergarten contexts. *Child Development*, *85*(1), 140–159. https://doi.org/10.1111/cdev.12117

*Bitter, C., O'Day, J., Gubbins, P., & Socias, M. (2009). What works to improve student literacy achievement? An examination of instructional practices in a balanced literacy approach. *Journal of Education for Students Placed at Risk*, *14*(1), 17–44. https://doi.org/10.1080/10824660802715403

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Brunsek, A., Perlman, M., Falenchuk, O., McMullen, E., Fletcher, B., & Shah, P. S. (2017). The relationship between the Early Childhood Environment Rating Scale and its revised form and child outcomes: A systematic review and meta-analysis. *PLoS One*, *12*(6), e0178512. https://doi.org/10.1371/journal.pone.0178512

*Bryce, C. I., Bradley, R. H., Abry, T., Swanson, J., & Thompson, M. S. (2019). Parents' and teachers' academic influences, behavioral engagement, and first- and fifth-grade achievement. *School Psychology*, *34*(5), 492. http://dx.doi.org/10.1037/spq0000297

*Bulotsky-Shearer, R. J., Bell, E. R., Carter, T. M., & Dietrich, S. L. (2014). Peer play interactions and learning for low-income preschool children: The moderating role of classroom quality. *Early Education and Development*, *25*(6), 815–840. http://dx.doi.org/10.1080/10409289.2014.864214

*Burchinal, M., Field, S., López, M. L., Howes, C., & Pianta, R. (2012). Instruction in Spanish in pre-kindergarten classrooms and child outcomes for English language learners. *Early Childhood Research Quarterly*, *27*(2), 188–197. https://doi.org/10.1016/j.ecresq.2011.11.003

*Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-Kindergarten teacher–child interactions and instruction. *Applied Developmental Science*, *12*(3), 140–153. https://doi.org/10.1080/10888690802199418

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 11–31). Paul H. Brookes Publishing Co.

*Burchinal, M. R., & Nelson, L. (2000). Family selection and child care experiences: Implications for studies of child outcomes. *Early Childhood Research Quarterly*, *15*(3), 385–411. https://doi.org/10.1016/S0885-2006(00)00072-7

*Burchinal, M., Peisner-Feinberg, E., Bryant, D. M., & Clifford, R. (2000). Children's social and cognitive development and child-care quality: Testing for differential associations related to poverty, gender, or ethnicity. *Applied Developmental Science*, *4*(3), 149–165. https://doi.org/10.1207/S1532480XADS0403_4

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*, *25*(2), 166–176. https://doi.org/10.1016/j.ecresq.2009.10.004

*Burchinal, M., Vernon-Feagans, L., Vitiello, V., Greenberg, M., & Family Life Project Key Investigators. (2014). Thresholds in the association between child care quality and child outcomes in rural preschool children. *Early Childhood Research Quarterly*, *29*(1), 41–51. http://dx.doi.org/10.1016/j.ecresq.2013.09.004

Burchinal, M., Xue, Y., Auger, A., Tien, H. C., Mashburn, A., Peisner-Feinberg, E., Cavadel, E., Zaslow, M., & Tarullo, L. (2016). Testing for quality thresholds and features in early care and education. *Monographs of the Society for Research in Child Development*, *81*(2), 46–63. https://doi.org/10.1111/mono.12238

Cabell, S. Q., DeCoster, J., LoCasale-Crouch, J., Hamre, B. K., & Pianta, R. C. (2013). Variation in the effectiveness of instructional interactions across preschool classroom settings and learning activities. *Early Childhood Research Quarterly*, *28*(4), 820–830. https://doi.org/10.1016/j.ecresq.2013.07.007

*Cadima, J., Leal, T., & Burchinal, M. (2010). The quality of teacher-student interactions: Associations with first graders' academic and behavioral outcomes. *Journal of School Psychology*, *48*(6), 457–482. https://doi.org/10.1016/j.jsp.2010.09.001

*Cameron, C. E., Connor, C. M. D., Morrison, F. J., & Jewkes, A. M. (2008). Effects of classroom organization on letter-word reading in first grade. *Journal of School Psychology*, *46*(2), 173–192. https://doi.org/10.1016/j.jsp.2007.03.002

*Carlisle, J., Kelcey, B., Berebitsky, D., & Phelps, G. (2011). Embracing the complexity of instruction: A study of the effects of teachers' instruction on students' reading comprehension. *Scientific Studies of Reading*, *15*(5), 409–439. https://doi.org/10.1080/10888438.2010.497521

Cash, A. H., Cabell, S. Q., Hamre, B. K., DeCoster, J., & Pianta, R. C. (2015). Relating pre-Kindergarten teacher beliefs and knowledge to children's language and literacy development. *Teaching and Teacher Education*, *48*(1), 97–105. https://doi.org/10.1016/j.tate.2015.02.003

Certo, J., Moxley, K., Reffitt, K., & Miller, J. A. (2010). I learned how to talk about a book: Children's perceptions of literature circles across grade and ability levels. *Literacy Research and Instruction*, *49*(3), 243–263. https://doi.org/10.1080/19388070902947352

*Coker, D. L., Jr., Jennings, A. S., Farley-Ripple, E., & MacArthur, C. A. (2018). When the type of practice matters: The relationship between typical writing instruction, student practice, and writing achievement in first grade. *Contemporary Educational Psychology*, *54*, 235–246. https://doi.org/10.1016/j.cedpsych.2018.06.013

Connor, C. M. (Ed.). (2016). *The cognitive development of reading and reading comprehension*. Routledge.

*Connor, C. M., Jakobsons, L. J., Crowe, E. C., & Meadows, J. G. (2009). Instruction, student engagement, and reading skill growth in Reading First classrooms. *The Elementary School Journal*, *109*(3), 221–250. https://doi.org/10.1086/592305

Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., Piasta, S. B., Crowe, E. C., & Schatschneider, C. (2009). The ISI classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher*, *38*(2), 85–99. https://doi.org/10.3102/0013189X09332373

*Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, *8*(4), 305–336. https://doi.org/10.1207/s1532799xssr0804_1

*Connor, C. M., Morrison, F. J., & Petrella, J. N. (2004). Effective reading comprehension instruction: Examining child × instruction interactions. *Journal of Educational Psychology*, *96*(4), 682–698. https://doi.org/10.1037/0022-0663.96.4.682

*Connor, C. M., Morrison, F. J., & Underwood, P. S. (2007). A second chance in second grade: The independent and cumulative impact of first- and second-grade reading instruction and students' letter-word reading skill growth. *Scientific Studies of Reading*, *11*(3), 199–233. https://doi.org/10.1080/10888430701344314

*Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., Underwood, P., & Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child × instruction interactions on first graders' literacy development. *Child Development*, *80*(1), 77–100. https://doi.org/10.1111/j.1467-8624.2008.01247.x

*Connor, C. M., Son, S. H., Hindman, A. H., & Morrison, F. J. (2005). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology*, *43*(4), 343–375. https://doi.org/10.1016/j.jsp.2005.06.001

Connor, C. M., Spencer, M., Day, S. L., Giuliani, S., Ingebrand, S. W., McLean, L., & Morrison, F. J. (2014). Capturing the complexity: Content, type, and amount of instruction and quality of the classroom learning environment synergistically predict third graders' vocabulary and reading comprehension outcomes. *Journal of Educational Psychology*, *106*(3), 762–778. https://doi.org/10.1037/a0035921

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research*, *77*(1), 113–143. https://doi.org/10.3102/003465430298563

*Crawford, A. D., Zucker, T. A., Williams, J. M., Bhavsar, V., & Landry, S. H. (2013). Initial validation of the Pre-Kindergarten Classroom Observation Tool and goal setting system for data-based coaching. *School Psychology Quarterly*, *28*(4), 277. https://doi.org/10.1037/spq0000033

Crocker, R. K., & Brooker, G. M. (1986). Classroom control and student outcomes in Grades 2 and 5. *American Educational Research Journal*, *23*(1), 1–11. https://doi.org/10.3102/00028312023001001

Crosnoe, R., Morrison, F., Burchinal, M., Pianta, R., Keating, D., Friedman, S. L., & Clarke-Stewart, K. A. (2010). Instruction, teacher-student relations, and math achievement trajectories in elementary school. *Journal of Educational Psychology*, *102*(2), 407–417. https://doi.org/10.1037/a0017762

*Curby, T. W., Brock, L. L., & Hamre, B. K. (2013). Teachers' emotional support consistency predicts children's achievement gains and social skills. *Early Education & Development*, *24*(3), 292–309. https://doi.org/10.1080/10409289.2012.665760

Curby, T. W., Rimm-Kaufman, S. E., & Ponitz, C. C. (2009). Teacher-child interactions and children's achievement trajectories across kindergarten and first grade. *Journal of Educational Psychology*, *101*(4), 912–925. https://doi.org/10.1037/a0016647

Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teachers": What does "scientifically-based research" actually tell us? *Educational Researcher*, *31*(9), 13–25. https://doi.org/10.3102/0013189X031009013

*Day, S. L., Connor, C. M. D., & McClelland, M. M. (2015). Children's behavioral regulation and literacy: The impact of the first grade classroom environment. *Journal of School Psychology*, *53*(5), 409–428. https://doi.org/10.1016/j.jsp.2015.07.004

*de Bilde, J., Boonen, T., Speybroeck, S., De Fraine, B., Van Damme, J., & Goos, M. (2015). Experiential education in kindergarten: Associations with school adjustment and the moderating role of initial achievement. *The Elementary School Journal*, *116*(2), 173–197. https://doi.org/10.1086/683802

*Dobbs-Oates, J., Kaderavek, J. N., Guo, Y., & Justice, L. M. (2011). Effective behavior management in preschool classrooms and children's task orientation: Enhancing emergent literacy and language development. *Early Childhood Research Quarterly*, *26*(4), 420–429. https://doi.org/10.1016/j.ecresq.2011.02.003

*Dotterer, A. M., Burchinal, M., Bryant, D., Early, D., & Pianta, R. C. (2013). Universal and targeted pre-kindergarten programmes: A comparison of classroom characteristics and child outcomes. *Early Child Development and Care*, *183*(7), 931–950. https://doi.org/10.1080/03004430.2012.698388

*Dynia, J. M. (2012). *The literacy environment of early childhood special education classrooms: Predictors of print knowledge* [Unpublished doctoral dissertation]. The Ohio State University.

*Early, D. M., Bryant, D. M., Pianta, R. C., Clifford, R. M., Burchinal, M. R., Ritchie, S., Howes, C., & Barbarin, O. (2006). Are teachers' education, major, and credentials related to classroom quality and children's academic gains in pre-kindergarten? *Early Childhood Research Quarterly*, *21*(2), 174–195. https://doi.org/10.1016/j.ecresq.2006.04.004

Early, D. M., Maxwell, K. L., Burchinal, M., Bender, R. H., Ebanks, C., Henry, G. T., Iriondo-Perez, J., Mashburn, A. J., Pianta, R. C., Alva, S., Bryant, D., Cai, K., Clifford, R. M., Griffin, J. A., Howes, C., Jeon, H.-J., Peisner-Feinberg, E., Vandergrift, N., & Zill, N. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development*, *78*(2), 558–580. https://doi.org/10.1111/j.1467-8624.2007.01014.x

*Early, D. M., Sideris, J., Neitzel, J., LaForett, D. R., & Nehler, C. G. (2018). Factor structure and validity of the Early Childhood Environment Rating Scale–Third Edition (ECERS-3). *Early Childhood Research Quarterly*, *44*, 242–256. https://doi.org/10.1016/j.ecresq.2018.04.009

Fisher, Z., Tipton, E., & Zhipeng, H. (2017, May 29). *Package 'robumeta'*. https://cran.r-project.org/web/packages/robumeta/robumeta.pdf

Flanders, N. A., & Simon, A. (1969). Teacher effectiveness. *Classroom Interaction Newsletter*, *5*(1), 18–37. http://www.jstor.org/stable/23869471

Foorman, B. R., Schatschneider, C., Eakin, M. N., Fletcher, J. M., Moats, L. C., & Francis, D. J. (2006). The impact of instructional practices in Grades 1 and 2 on reading and spelling achievement in high poverty schools. *Contemporary Educational Psychology*, *31*(1), 1–29. https://doi.org/10.1016/j.cedpsych.2004.11.003

Frick, T., & Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research*, *48*(1), 157–184. https://doi.org/10.3102/00346543048001157

Gage, N. L. (1963). *Handbook of research on teaching*. Rand McNally.

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of Teacher Education*, *53*(2), 106–116. https://doi.org/10.1177/0022487102053002003

*Gerde, H. K., Bingham, G. E., & Pendergast, M. L. (2015). Reliability and validity of the Writing Resources and Interactions in Teaching Environments (WRITE) for preschool classrooms. *Early Childhood Research Quarterly*, *31*, 34–46. http://dx.doi.org/10.1016/j.ecresq.2014.12.008

Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S., & Santoro, L. E. (2010). Teacher study group: Impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *American Educational Research Journal*, *47*(3), 694–739. https://doi.org/10.3102/0002831209361208

*Goble, P., & Pianta, R. C. (2017). Teacher–child interactions in free choice and teacher-directed activity settings: Prediction to school readiness. *Early Education and Development*, *28*(8), 1035–1051. https://doi.org/10.1080/10409289.2017.1322449

Gosling, D. (2002). *Models of peer observation of teaching*. LTSN Generic Centre.

Gosse, C. S., McGinty, A. S., Mashburn, A. J., Hoffman, L. M., & Pianta, R. C. (2014). The role of relational and instructional classroom supports in the language development of at-risk preschoolers. *Early Education & Development*, *25*(1), 110–133. https://doi.org/10.1080/10409289.2013.778567

*Guo, Y., Connor, C. M. D., Yang, Y., Roehrig, A. D., & Morrison, F. J. (2012). The effects of teacher qualification, teacher self-efficacy, and classroom practices on fifth graders' literacy outcomes. *Elementary School Journal*, *113*(1), 3–24. https://doi.org/10.1086/665816

Guo, Y., Justice, L. M., Kaderavek, J. N., & McGinty, A. (2012). The literacy environment of preschool classrooms: Contributions to children's emergent literacy growth. *Journal of Research in Reading*, *35*(3), 308–327. https://doi.org/10.1111/j.1467-9817.2010.01467.x

*Guo, Y., Kaderavek, J. N., Piasta, S. B., Justice, L. M., & McGinty, A. (2011). Preschool teachers' sense of community, instructional quality, and children's language and literacy gains. *Early Education and Development*, *22*(2), 206–233. https://doi.org/10.1080/10409281003641257

*Guo, Y., Tompkins, V., Justice, L., & Petscher, Y. (2014). Classroom age composition and vocabulary development among at-risk preschoolers. *Early Education and Development*, *25*(7), 1016–1034. https://doi.org/10.1080/10409289.2014.893759

Hale, J. B., Fiorello, C. A., & Brown, L. (2005). Determining medication treatment effects using teacher ratings and classroom observations of children with ADHD: Does neuropsychological impairment matter. *Educational and Child Psychology*, *22*(2), 39–61. https://search.proquest.com/docview/620862987?accountid=14509

*Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher–child interactions: Associations with preschool children's development. *Child Development*, *85*(3), 1257–1274. https://doi.org/10.1111/cdev.12184

*Hamre, B. K., Justice, L. M., Pianta, R. C., Kilday, C., Sweeney, B., Downer, J. T., & Leach, A. (2010). Implementation fidelity of MyTeachingPartner literacy and language activities: Association with preschoolers' language and literacy growth. *Early Childhood Research Quarterly*, *25*(3), 329–347. https://doi.org/10.1016/j.ecresq.2009.07.002

*Han, J., Schlieber, M., & Gregory, B. (2017). Associations of home and classroom environments with Head Start children's code-related and oral language skills. *Journal of Education for Students Placed at Risk (JESPAR)*, *22*(4), 200–219. https://doi.org/10.1080/10824669.2017.1347044

Hanushek, E. A. (2002). *The long-run importance of school quality*. NBER Working Paper 9071, National Bureau of Economic Research.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale* (rev. ed.). Teachers College Press.

*Hatfield, B. E., Burchinal, M., Pianta, R. C., & Sideris, J. (2016). Thresholds in the association between quality of teacher-child interactions and preschool children's school readiness skills. *Early Childhood Research Quarterly*, *36*, 561–571. https://doi.org/10.1016/j.ecresq.2015.09.005

*Haynes, M. C., & Jenkins, J. R. (1986). Reading instruction in special education resource rooms. *American Educational Research Journal*, *23*(2), 161–190. http://www.jstor.com/stable/1162952

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

*Himley, O. T. (1986). *Iowa Chapter 1, ECIA Reading Study*. Iowa State Department of Public Instruction, Des Moines.

*Hoffman, A. J. (2011). *Indicators of school readiness: An investigation of classroom quality and language, literacy, and social-emotional competence in a sample of Head Start children* [Unpublished doctoral dissertation]. University of Wisconsin-Madison.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-Kindergarten programs. *Early Childhood Research Quarterly*, *23*(1), 27–50. https://doi.org/10.1016/j.ecresq.2007.05.002

Hyson, M. C., Hirsh-Pasek, K., & Rescorla, L. (1990). The classroom practices inventory: An observation instrument based on NAEYC's guidelines for developmentally appropriate practices for 4- and 5-year-old children. *Early Childhood Research Quarterly*, *5*(4), 475–494. https://doi.org/10.1016/0885-2006(90)90015-S

*Jacobson, K. R. (1982). *The relationship of individual student time allocation to reading and mathematics achievement* [Unpublished doctoral dissertation]. The University of Wisconsin-Madison.

*Jeon, H. J., Langill, C. C., Peterson, C. A., Luze, G. J., Carta, J. J., & Atwater, J. B. (2010). Children's individual experiences in early care and education: Relations with overall classroom quality and children's school readiness. *Early Education and Development*, *21*(6), 912–939. https://doi.org/10.1080/10409280903292500

*Jung, Y., Zuniga, S., Howes, C., Jeon, H. J., Parrish, D., Quick, H., Manship, K., & Hauser, A. (2016). Improving Latino children's early language and literacy development: Key features of early childhood education within family literacy programmes. *Early Child Development and Care*, *186*(6), 845–862. https://doi.org/10.1080/03004430.2015.1062374

Justice, L. M., Mashburn, A. J., Hamre, B. K., & Pianta, R. C. (2008). Quality of language and literacy instruction in preschool classrooms serving at-risk pupils. *Early Childhood Research Quarterly*, *23*(1), 51–68. https://doi.org/10.1016/j.ecresq.2007.09.004

*Kelcey, B., & Carlisle, J. F. (2013). Learning about teachers' literacy instruction from classroom observations. *Reading Research Quarterly*, *48*(3), 301–317. https://doi.org/10.1002/rrq.51

*Kent, S. C., Wanzek, J., & Al Otaiba, S. (2017). Reading instruction for fourth-grade struggling readers and the relation to student outcomes. *Reading & Writing Quarterly*, *33*(5), 395–411. https://doi.org/10.1080/10573569.2016.1216342

Keys, T. D., Farkas, G., Burchinal, M. R., Duncan, G. J., Vandell, D. L., Li, W., Ruzek, E. A., & Howes, C. (2013). Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. *Child Development*, *84*(4), 1171–1190. https://doi.org/10.1111/cdev.12048

*Kikas, E., Pakarinen, E., Soodla, P., Peets, K., & Lerkkanen, M. K. (2018). Associations between reading skills, interest in reading, and teaching practices in first grade. *Scandinavian Journal of Educational Research*, *62*(6), 832–849. https://doi.org/10.1080/00313831.2017.1307272

Kington, A., Sammons, P., Day, C., & Regan, E. (2011). Stories and statistics: Describing a mixed methods study of effective classroom practice. *Journal of Mixed Methods Research*, *5*(2), 103–125. https://doi.org/10.1177/1558689810396092

Kwan, C., Sylva, K., & Reeves, B. (1998). Day care quality and child development in Singapore. *Early Child Development and Care*, *144*(1), 69–77. https://doi.org/10.1080/0300443981440108

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the pre-Kindergarten year. *The Elementary School Journal*, *104*(5), 409–426. https://doi.org/10.1086/499760

Lau, K. L. (2012). Instructional practices and self-regulated learning in Chinese language classes. *Educational Psychology*, *32*(4), 427–450. https://doi.org/10.1080/01443410.2012.674634

*Le, V. N., Schaack, D. D., & Setodji, C. M. (2015). Identifying baseline and ceiling thresholds within the qualistar early learning quality rating and improvement system. *Early Childhood Research Quarterly*, *30*, 215–226. https://doi.org/10.1016/j.ecresq.2014.03.003

*Lee, P., & Bierman, K. L. (2015). Classroom and teacher support in kindergarten: Associations with the behavioral and academic adjustment of low-income students. *Merrill-Palmer Quarterly (Wayne State University Press)*, *61*(3), 383. https://doi.org/10.13110/merrpalmquar1982.61.3.0383

*Leinhardt, G., Zigmond, N., & Cooley, W. W. (1981). Reading instruction and its effects. *American Educational Research Journal*, *18*(3), 343–361. http://www.jstor.com/stable/1162667

*Lerkkanen, M. K., Kiuru, N., Pakarinen, E., Poikkeus, A. M., Rasku-Puttonen, H., Siekkinen, M., & Nurmi, J. E. (2016). Child-centered versus teacher-directed teaching practices: Associations with the development of academic skills in the first grade at school. *Early Childhood Research Quarterly*, *36*, 145–156. https://doi.org/10.1016/j.ecresq.2015.12.023

*Leyva, D., Weiland, C., Barata, M., Yoshikawa, H., Snow, C., Treviño, E., & Rolla, A. (2015). Teacher–child interactions in Chile and their associations with pre-Kindergarten outcomes. *Child Development*, *86*(3), 781–799. https://doi.org/10.1111/cdev.12342

*Li, K., Pan, Y., Hu, B., Burchinal, M., De Marco, A., Fan, X., & Qin, J. (2016). Early childhood education quality and child outcomes in China: Evidence from Zhejiang

Province. *Early Childhood Research Quarterly*, *36*, 427–438. http://dx.doi.org/10.1016/j.ecresq.2016.01.009

*Lin, Y. C., & Magnuson, K. A. (2018). Classroom quality and children's academic skills in child care centers: Understanding the role of teacher qualifications. *Early Childhood Research Quarterly*, *42*, 215–227. http://dx.doi.org/10.1016/j.ecresq.2017.10.003

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE Publications.

*López, F. (2011). The nongeneralizability of classroom dynamics as predictors of achievement for Hispanic students in upper elementary grades. *Hispanic Journal of Behavioral Sciences*, *33*(3), 350–376. https://doi.org/10.1177/0739986311415222

*López, F. (2012). Moderators of language acquisition models and reading achievement for English language learners: The role of emotional warmth and instructional support. *Teachers College Record*, *114*(8), 1–30.

Lucero, E., & Rouse, M. (2017). Classroom interaction in ELTE undergraduate programs: Characteristics and pedagogical implications. *Colombian Applied Linguistics Journal*, *19*(2), 193–208. http://dx.doi.org/10.14483/22487085.10801

Luke, A. (2012). Critical literacy: Foundational notes. *Theory into Practice*, *51*(1), 4–11. https://doi.org/10.1080/00405841.2012.636324

*Maier, M. F., Vitiello, V. E., & Greenfield, D. B. (2012). A multilevel model of child- and classroom-level psychosocial factors that support language and literacy resilience of children in Head Start. *Early Childhood Research Quarterly*, *27*(1), 104–114. http://dx.doi.org/10.1016/j.ecresq.2011.06.002

Markova, I. (2017). Effects of academic and non-academic instructional approaches on preschool English language learners' classroom engagement and English language development. *Journal of Early Childhood Research*, *15*(4), 339–358. https://doi.org/10.1177%2F1476718X15609390

*Mashburn, A. J. (2008). Quality of social and physical environments in preschools and children's development of academic, language, and literacy skills. *Applied Developmental Science*, *12*(3), 113–127. https://doi.org/10.1080/10888690802199392

Mashburn, A. J., Justice, L. M., Downer, J. T., & Pianta, R. C. (2009). Peer effects on children's language achievement during pre-Kindergarten. *Child Development*, *80*(3), 686–702. https://doi.org/10.1111/j.1467-8624.2009.01291.x

*Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, D. M., & Howes, C. (2008). Measures of classroom quality in pre-Kindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749. https://doi.org/10.1111/j.1467-8624.2008.01154.x

McCartney, K. (1984). Effect of quality of day care environment on children's language development. *Developmental Psychology*, *20*(2), 244–260. https://doi.org/10.1037/0012-1649.20.2.244

*McCaslin, M., & Burross, H. (2011). Research on individual differences within a sociocultural perspective: Co-regulation and adaptive learning. *Teachers College Record*, *113*(2), 325–349. https://www.researchgate.net/publication/289169398

*McCool, Y. D. (2013). *Scientifically based reading research strategies in the preschool classroom: An investigation into quality early childhood reading practices and literacy acquisition in one northern Michigan early reading first program* [Unpublished doctoral dissertation]. Eastern Michigan University.

*McGinty, A. S., Justice, L. M., Piasta, S. B., Kaderavek, J., & Fan, X. (2012). Does context matter? Explicit print instruction during reading varies in its influence by child and classroom factors. *Early Childhood Research Quarterly*, *27*(1), 77–89. https://doi.org/10.1016/j.ecresq.2011.05.002

McIntosh, A. S., Graves, A., & Gersten, R. (2007). The effects of response to intervention on literacy development in multiple-language settings. *Learning Disability Quarterly*, *30*(3), 197–212. https://doi.org/10.2307/30035564

Medley, D. M., & Mitzel, H. E. (1963). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 247–328). Rand McNally.

*Meyer, L. A., Wardrop, J. L., & Hastings, C. N. (1990). *The development of reading ability in first and second grade*. Center for the Study of Reading Technical Report (no. 516).

*Moffett, L., & Morrison, F. J. (2020). Off-task behavior in kindergarten: Relations to executive function and academic achievement. *Journal of Educational Psychology*, *112*(5), 938. http://dx.doi.org/10.1037/edu0000397

Morrison, F. J., & Connor, C. M. (2009). The transition to school: Child-instruction transactions in learning to read. In A. Sameroff (Ed.), *The transactional model of development: How children and contexts shape each other* (pp. 183–201). American Psychological Association. https://doi.org/10.1037/11877-010

Morsh, J. E., & Wilder, E. W. (1954). *Identifying the effective instructor: A review of the quantitative studies, 1900-1952*. USAF Personnel Training Research Center. Research Bulletin.

National Heart, Lung, & Blood Institute. (n.d.). *Study quality assessment tools*. https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools

National Institute of Child Health and Human Development Early Child Care Research Network. (1996). Characteristics of infant child care: Factors contributing to positive caregiving. *Early Childhood Research Quarterly*, *11*(3), 269–306. https://doi.org/10.1016/S0885-2006(96)90009-5

National Institute of Child Health and Human Development Early Child Care Research Network. (2002). The relation of global first-grade classroom environment to structural classroom features and teacher and student behaviors. *Elementary School Journal*, *102*(5), 367–387. https://doi.org/10.1086/499709

*National Institute of Child Health and Human Development Early Child Care Research Network. (2005). Early child care and children's development in the primary grades: Follow-up results from the NICHD Study of Early Child Care. *American Educational Research Journal*, *42*(3), 537–570. https://www.jstor.org/stable/3700462

*National Institute of Child Health and Human Development Early Child Care Research Network. (2006). Child-care effect sizes for the NICHD Study of Early Child Care and Youth Development. *American Psychologist*, *61*(2), 99–116. https://doi.org/10.1037/0003-066X.61.2.99

*National Institute of Child Health and Human Development Early Child Care Research Network, & Duncan, G. J. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development*, *74*(5), 1454–1475. https://doi.org/10.1111/1467-8624.00617

*Neugebauer, S., Sandilos, L., Coyne, M., McCoach, D. B., & Ware, S. (2020). Highly potent and vastly conditional instructional practices: Variations in use and utility of

language interactions for kindergarten. *Early Education and Development*, *31*(4), 541–560. https://doi.org/10.1080/10409289.2019.1686928

*Neugebauer, S. R., Gámez, P. B., Coyne, M. D., McCoach, D. B., Cólon, I. T., & Ware, S. (2017). Promoting word consciousness to close the vocabulary gap in young word learners. *The Elementary School Journal*, *118*(1), 28–54. https://doi.org/10.1086/692986

Ollin, R. (2008). Silent pedagogy and rethinking classroom practice: Structuring teaching through silence rather than talk. *Cambridge Journal of Education*, *38*(2), 265–280. https://doi.org/10.1080/03057640802063528

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

*Pakarinen, E., Kiuru, N., Lerkkanen, M.-K., Poikkeus, A.-M., Siekkinen, M., & Nurmi, J.-E. (2010). Classroom organization and teacher stress predict learning motivation in kindergarten children. *European Journal of Psychology of Education*, *25*(3), 281–300. https://doi.org/10.1007/s10212-010-0025-6

*Pakarinen, E., Lerkkanen, M.-K., Poikkeus, A.-M., Rasku-Puttonen, H., Eskelä-Haapanen, S., Siekkinen, M., & Nurmi, J.-E. (2017). Associations among teacher–child interactions, teacher curriculum emphases, and reading skills in Grade 1. *Early Education and Development*, *28*(7), 858–879. https://doi.org/10.1080/10409289.2017.1289768

*Pakarinen, E., Lerkkanen, M.-K., Poikkeus, A.-M., Salminen, J., Silinskas, G., Siekkinen, M., & Nurmi, J.-E. (2017). Longitudinal associations between teacher-child interactions and academic skills in elementary school. *Journal of Applied Developmental Psychology*, *52*, 191–202. https://doi.org/10.1016/j.appdev.2017.08.002

Park, Y., Brownell, M. T., Bettini, E. F., & Benedict, A. E. (2019). Multiple dimensions of instructional effectiveness in reading: A review of classroom observation studies and implications for special education classrooms. *Exceptionality*, *27*(1), 1–17. https://doi.org/10.1080/09362835.2017.1283628

*Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., Yazejian, N., Byler, P., Rustici, J., & Zelazo, J. (1999). *The children of the cost, quality, and outcomes study go to school: Technical report*. University of North Carolina at Chapel Hill, Frank Porter Graham Child Development Center.

*Peisner-Feinberg, E. S., LaForett, D. R., Schaaf, J. M., Hildebrandt, L. M., Sideris, J., & Pan, Y. (2014). *Children's outcomes and program quality in the North Carolina Pre Kindergarten Program: 2012–2013 Statewide evaluation*. The University of North Carolina, FPG Child Development Institute.

Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the Classroom Assessment Scoring System) in early childhood education and care settings and child outcomes. *PLoS One*, *11*(12), e0167660. https://doi.org/10.1371/journal.pone.0167660

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*(2), 365–397. https://doi.org/10.3102/0002831207308230

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity.

Educational Researcher, *38*(2), 109–119. https://doi.org/10.3102/001318
9X09332374

Pianta, R. C., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre-Kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science*, *9*(3), 144–159. https://doi.org/10.1207/s1532480xads0903_2

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System*™*: Manual K-3*. Paul H Brookes Publishing.

*Pianta, R. C., Whittaker, J. E., Vitiello, V., Ruzek, E., Ansari, A., Hofkens, T., & DeCoster, J. (2020). Children's school readiness skills across the pre-K year: Associations with teacher-student interactions, teacher practices, and exposure to academic content. *Journal of Applied Developmental Psychology*, *66*, 101084. https://doi.org/10.1016/j.appdev.2019.101084

*Piasta, S. B., Connor, C. M., Fishman, B. J., & Morrison, F. J. (2009). Teachers' knowledge of literacy concepts, classroom practices, and student reading growth. *Scientific Studies of Reading*, *13*(3), 224–248. https://doi.org/10.1080/108884309 02851364

*Pilcher, H. (2016). *Classroom organization by prior performance interactions as predictors of literacy and language achievement* [Unpublished doctoral dissertation]. The Florida State University.

*Pinto, A. I., Pessanha, M., & Aguiar, C. (2013). Effects of home environment and center-based child care quality on children's language, communication, and literacy outcomes. *Early Childhood Research Quarterly*, *28*(1), 94–101. http://dx.doi.org/10.1016/j.ecresq.2012.07.001

*Poe, M. D., Burchinal, M. R., & Roberts, J. E. (2004). Early language and the development of children's reading skills. *Journal of School Psychology*, *42*(4), 315–332. https://doi.org/10.1016/j.jsp.2004.06.001

*Ponitz, C. C., & Rimm-Kaufman, S. E. (2011). Contexts of reading instruction: Implications for literacy skills and kindergarteners' behavioral engagement. *Early Childhood Research Quarterly*, *26*(2), 157–168. https://doi.org/10.1016/j.ecresq.2010.10.002

*Ponitz, C. C., Rimm-Kaufman, S. E., Brock, L. L., & Nathanson, L. (2009). Early adjustment, gender differences, and classroom organizational climate in first grade. *The Elementary School Journal*, *110*(2), 142–162. https://doi.org/10.1086/605470

*Ponitz, C. C., Rimm-Kaufman, S. E., Grimm, K. J., & Curby, T. W. (2009). Kindergarten classroom quality, behavioral engagement, and reading achievement. *School Psychology Review*, *38*(1), 102–120. https://doi.org/10.1080/02796015.200 9.12087852

Pressley, M., Wharton-McDonald, R., Allington, R., Block, C. C., Morrow, L., Tracey, D., Baker, K., Brooks, G., Cronin, J., Nelson, E., & Woo, D. (2001). A study of effective first-grade literacy instruction. *Scientific Studies of Reading*, *5*(1), 35–58. https://doi.org/10.1207/S1532799XSSR0501_2

*Purtell, K. M., & Ansari, A. (2018). Classroom age composition and preschoolers' school readiness: The implications of classroom quality and teacher qualifications. *AERA Open*, *4*(1), 1–13. https://doi.org/10.1177/2332858418758300

R Core Team. (2018). *R: A language and environment for statistical computing*. https://www.r-project.org/

Reyes, I. (2006). Exploring connections between emergent biliteracy and bilingualism. *Journal of Early Childhood Literacy*, *6*(3), 267–292. https://doi.org/10.1177/1468798406069801

RStudio Team. (2016). *RStudio: Integrated development for R*. http://www.rstudio.com/.

*Sabol, T. J., Bohlmann, N. L., & Downer, J. T. (2018). Low-income ethnically diverse children's engagement as a predictor of school readiness above preschool classroom quality. *Child Development*, *89*(2), 556–576. https://doi.org/10.1111/cdev.12832

*Sabol, T. J., & Pianta, R. C. (2014). Do standard measures of preschool quality used in statewide policy predict school readiness? *Education Finance and Policy*, *9*(2), 116–164. https://doi.org/10.1162/EDFP_a_00127

*Schliecker, E., White, D. R., & Jacobs, E. (1991). The role of day care quality in the prediction of children's vocabulary. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *23*(1), 12–24. https://doi.org/10.1037/h0078960

Schunk, D. H. (2012). Social cognitive theory. In K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra, & J. Sweller (Eds.), *APA educational psychology handbook, Vol. 1. Theories, constructs, and critical issues* (pp. 101–123). American Psychological Association. https://doi.org/10.1037/13273-005

*Schmitt, H. A. (2018). *Upper elementary social studies instruction: Exploring the relationship between text use, comprehension instruction, and student engagement* [Unpublished doctoral dissertation]. Michigan State University.

Shin, M., & Partyka, T. (2017). Empowering infants through responsive and intentional play activities. *International Journal of Early Years Education*, *25*(2), 127–142. https://doi.org/10.1080/09669760.2017.1291331

*Slot, P. L., Bleses, D., Justice, L. M., Markussen-Brown, J., & Højen, A. (2018). Structural and process quality of Danish preschools: Direct and indirect associations with children's growth in language and preliteracy skills. *Early Education and Development*, *29*(4), 581–602. https://doi.org/10.1080/10409289.2018.1452494

*Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the Classroom Observations of Student–Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly*, *27*(2), 316–328. https://doi.org/10.1016/j.ecresq.2011.09.004

*Stallings, J., Almy, M., Resnick, L. B., & Leinhardt, G. (1975). Implementation and child effects of teaching practices in Follow Through classrooms. *Monographs of the Society for Research in Child Development*, 1–133. https://www.jstor.org/stable/1165828

Sterne, J. A. C., & Egger, M. (2006). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). John Wiley & Sons. https://doi.org/10.1002/0470870168.ch6

Stipek, D. (1996). Motivation and instruction. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 85–113). Macmillan.

*Stipek, D., & Chiatovich, T. (2017). The effect of instructional quality on low- and high-performing students. *Psychology in the Schools*, *54*(8), 773–791. https://doi.org/10.1002/pits.22034

Sun, K. L. (2019). The mindset disconnect in mathematics teaching: A qualitative analysis of classroom instruction. *The Journal of Mathematical Behavior*, *56*, 100706. https://doi.org/10.1016/j.jmathb.2019.04.005

*Swanson, B. B. (1981, November 6–7). *The beginner's concepts about reading, attitudes, and reading achievement in relationship to language environment* [Paper presentation]. *Annual Meeting of the University of South Carolina Conference of Educational Research*, Columbia, SC.

*Sylva, K., Siraj-Blatchford, I., Taggart, B., Sammons, P., Melhuish, E., Elliot, K., & Totsika, V. (2006). Capturing quality in early childhood through environmental rating scales. *Early Childhood Research Quarterly*, *21*(1), 76–92. https://doi.org/10.1016/j.ecresq.2006.01.003

Taylor, B. M., Pearson, P. D., Clark, K., & Walpole, S. (2000). Effective schools and accomplished teachers: Lessons about primary-grade reading instruction in low-income schools. *Elementary School Journal*, *101*(2), 121–165. https://doi.org/10.1086/499662

*Taylor, B. M., Pearson, P. D., Peterson, D., & Rodriguez, M. C. (2002). *The CIERA school change project: Supporting schools as they implement home-grown reading reform. CIERA Report*. Center for the Improvement of Early Reading Achievement, Ann Arbor, MI.

*Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal*, *104*(1), 3–28. http://www.jstor.com/stable/3203047

*Taylor, B. M., Peterson, D. S., Pearson, P. D., & Rodriguez, M. C. (2002). Looking inside classrooms: Reflecting on the "how" as well as the "what" in effective reading instruction. *The Reading Teacher*, *56*(3), 270–279.

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology*, *2*(1), 85–112. http://doi.org/10.1007/s40865-016-0026-5

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Ulferts, H., Wolf, K. M., & Anders, Y. (2019). Impact of process quality in early childhood education and care on academic outcomes: Longitudinal meta-analysis. *Child Development*, *90*(5), 1474–1489. https://doi.org/10.1111/cdev.13296

Vermeer, H. J., van IJzendoorn, M. H., Cárcamo, R. A., & Harrison, L. J. (2016). Quality of child care using the environment rating scales: A meta-analysis of international studies. *International Journal of Early Childhood*, *48*(1), 33–60. https://doi.org/10.1007/s13158-015-0154-9

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

*Vitiello, V. E., Bassok, D., Hamre, B. K., Player, D., & Williford, A. P. (2018). Measuring the quality of teacher–child interactions at scale: Comparing research-based and state observation approaches. *Early Childhood Research Quarterly*, *44*, 161–169. https://doi.org/10.1016/j.ecresq.2018.03.003

Vitiello, V. E., Moas, O., Henderson, H. A., Greenfield, D. B., & Munis, P. M. (2012). Goodness of fit between children and classrooms: Effects of child temperament and preschool classroom quality on achievement trajectories. *Early*

Education & Development, *23*(3), 302–322. https://doi.org/10.1080/10409289.2011.526415

*Votruba-Drzal, E., Levine Coley, R., & Lindsay Chase-Lansdale, P. (2004). Child care and low-income children's development: Direct and moderated effects. *Child Development*, *75*(1), 296–312. https://www.jstor.org/stable/3696582

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Walsh, K., & Tracy, C. O. (2004). *Increasing the odds: How good policies can yield better teachers*. National Council on Teacher Quality.

Wayne, A., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, *73*(1), 89–122. http://www.jstor.org/stable/3516044

*Weiland, C., Ulvestad, K., Sachs, J., & Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public pre-Kindergarten program. *Early Childhood Research Quarterly*, *28*(2), 199–209. https://doi.org/10.1016/j.ecresq.2012.12.002

Wharton-McDonald, R., Pressley, M., & Hampston, J. M. (1998). Literacy instruction in nine first-grade classrooms: Teacher characteristics and student achievement. *The Elementary School Journal*, *99*(2), 101–128. https://doi.org/10.1086/461918

*White, K. M. (2016). "My teacher helps me": Assessing teacher-child relationships from the child's perspective. *Journal of Research in Childhood Education*, *30*(1), 29–41. https://doi.org/10.1080/02568543.2015.1105333

*White, L. J., Fernandez, V. A., & Greenfield, D. B. (2020). Assessing classroom quality for Latino dual language learners in Head Start: DLL-specific and general teacher-child interaction perspectives. *Early Education and Development*, *31*(4), 599–627. https://doi.org/10.1080/10409289.2019.1680785

*Williford, A. P., Maier, M. F., Downer, J. T., Pianta, R. C., & Howes, C. (2013). Understanding how children's engagement and teachers' interactions combine to predict school readiness. *Journal of Applied Developmental Psychology*, *34*(6), 299–309. https://doi.org/10.1016/j.appdev.2013.05.002

Wragg, E. C. (1999). *An introduction to classroom observation* (2nd ed.). Routledge.

*Xu, Y., Chin, C., Reed, E., & Hutchinson, C. (2014). The effects of a comprehensive early literacy project on preschoolers' language and literacy skills. *Early Childhood Education Journal*, *42*(5), 295–304. https://doi.org/10.1007/s10643-013-0613-6

Yamamoto, K. (1963). Evaluating teacher effectiveness: A review of research. *Journal of School Psychology*, *2*(1), 60–71. https://doi.org/10.1016/0022-4405(63)90017-7

Authors

YUCHENG CAO is a postdoctoral researcher in the School of Education at University of California, Irvine; 3200 Education Bldg, Irvine, CA 92697; *e-mail: yuchc10@uci.edu*. Her research interests include reading and writing development, teacher professional development, and classroom effectiveness.

YOUNG-SUK GRACE KIM, Ed.D. is a professor in the School of Education at University of California, Irvine; 3200 Education Bldg, Irvine, CA 92697; *e-mail: youngsk7@uci.edu*. Her work includes the development and effective instruction of language, cognition, reading, and writing skills for children from diverse linguistic, cultural, and socioeconomic backgrounds.

MINKYUNG CHO is a Ph.D. candidate in the School of Education at University of California, Irvine; 3200 Education Bldg, Irvine, CA 92697; *e-mail: minkyc1@uci.edu*. Her research focuses on the role of higher-order thinking skills and discourse knowledge in oral and written language production for elementary and secondary students.