Document 20

Educational Evaluation and Policy Analysis
September 2017, Vol. 39, No. 3, pp. 407–428
DOI: 10.3102/0162373716689489

A Meta-Analysis of Class Sizes and Ratios in Early Childhood Education Programs: Are Thresholds of Quality Associated With Greater Impacts on Cognitive, Achievement, and Socioemotional Outcomes?

Jocelyn Bonnes Bowne
Harvard Graduate School of Education
Katherine A. Magnuson
University of Wisconsin-Madison
Holly S. Schindler
University of Washington
Greg J. Duncan
University of California, Irvine
Hirokazu Yoshikawa
New York University

This study uses data from a comprehensive database of U.S. early childhood education program evaluations published between 1960 and 2007 to evaluate the relationship between class size, child–teacher ratio, and program effect sizes for cognitive, achievement, and socioemotional outcomes. Both class size and child–teacher ratio showed nonlinear relationships with cognitive and achievement effect sizes. For child–teacher ratios 7.5:1 and lower, the reduction of this ratio by one child per teacher predicted an effect size of 0.22 standard deviations greater. For class sizes 15 and smaller, one child fewer predicted an effect size of 0.10 standard deviations larger. No discernible relationship was found for larger class sizes and child–teacher ratios. Results were less clear for socioemotional outcomes due to a small sample.

Keywords: *child-teacher ratio; class size; early childhood education; meta-analysis*

DECADES of program evaluation research have shown that center-based early childhood education (ECE) programs can improve children's cognitive and social development (Camilli, Vargas, Ryan, & Barnett, 2010). As a result, policymakers and educators have pushed for public funding to expand access to ECE. With increased funding, most children now attend ECE before entering elementary school (Burgess, Chien, Morrissey, & Swenson, 2014; Magnuson & Shager, 2010). Yet, with increased enrollment has come greater scrutiny about whether ECE programs are delivering on the promise of improved school readiness that fueled their growth. Increasingly, researchers and policymakers are paying greater attention to how to improve children's early learning by increasing the effectiveness and educational benefits of existing ECE programs.

Although experts agree that the most effective center-based ECE programs provide developmentally appropriate and enriching social and academic environments, most concede that it is hard to define and measure these aspects of settings in cost-effective ways. Therefore, policymakers have often focused on structural dimensions of program quality—aspects of ECE program design that can

be easily defined, measured, and legislated—such as class sizes, teacher education, and child–teacher ratios. These aspects of program design are thought to affect children indirectly, by promoting or impeding safe and positive care environments. As a result, requirements for minimum standards related to structural indicators have long been part of child care business licensing requirements and are now increasingly central features of state and local quality rating and improvement systems (QRIS), as well as Head Start performance standards (U.S. Department of Health and Human Services, 1996). Despite the appeal of using structural indicators to assess program quality, empirical evidence on the association between structural indicators and children's learning is lacking. In this study, we use meta-analytic methods to estimate the association between two of the most-widely used dimensions of ECE structural quality (class size and child–teacher ratio) and children's cognitive and achievement outcomes in early childhood classrooms.

Background

Although ECE pedagogy may differ across programs, many theoretical models of early learning place interactions between children and their teachers, as well as peers, at the center of the learning process (Burchinal et al., 2008; National Scientific Council on the Developing Child, 2004). For this reason, having both smaller classes and lower child–teacher ratios are thought to improve classroom environments and increase ECE's effectiveness in promoting early learning. Smaller classes and lower child–teacher ratios may enable teachers to spend more time interacting with each individual child, which may in turn provide greater opportunities to understand each child's development, tailor activities to children's interests and abilities, and scaffold children's learning.

Evidence in support of this hypothesis comes from analyses of the NICHD Study of Early Child Care data and a sample of child care programs in four states. In both analyses, smaller child–teacher ratios were associated with higher quality child–teacher interactions as measured by warmth, sensitivity, and cognitive stimulation (NICHD, 2002; Phillipsen, Burchinal, Howes, & Cryer, 1997). Smaller class sizes and lower child–teacher ratios have also been related to reduced behavior problems and teacher time spent on classroom management in elementary school classrooms (Blatchford, Bassett, & Brown, 2011; Finn, Pannozzo, & Achilles, 2003; NICHD Early Child Care Research Network, 2004). Though highly related, child–teacher ratios and class sizes may also make distinct contributions to classroom experiences and quality. Although these issues have been studied more extensively in elementary education settings, with conclusions suggesting small positive effects of smaller class sizes, the differences between elementary school classrooms and preschool classrooms, as well as the developmental differences among younger and older children, make it difficult to generalize these findings downward to preschool classroom settings.

In early education classrooms, existing data suggest about a third of children's time is spent in free choice activities and another third is spent in routine activities such as meals (Early et al., 2010). For this reason, lower child–teacher ratios, whether due to more teachers or smaller class sizes, may make it easier for teachers to interact individually with every student and monitor the classroom activity. Even in a large class, an additional teacher may make it easier for the teachers to work together to observe all activities, facilitate teacher intervention or support when necessary, and provide more opportunities for one or two teachers to work individually with children, while another supervises the classroom. At the same time, professional training of early childhood teachers can be quite variable, ranging from little more than a high school education to advanced degrees in early childhood (Early et al., 2006; Mashburn et al., 2008). Although additional teachers may have added importance in early childhood classrooms due to the child-directed nature of activities, potential benefits to child outcomes might not be realized if individual teachers vary greatly in the skill with which they supervise and interact with young children.

Smaller classes in early education settings, even those with higher child–teacher ratios, limit the total number of children with whom teachers and students interact, potentially making it easier for positive relationships to be built and maintained with every student. The smaller number of students also reduces the workload involved in

tracking student progress, may increase the likelihood of effective individualization of instruction, and may make each individual more visible and connected, creating greater social and academic engagement (Finn, et al., 2003). Furthermore, smaller classes may be quieter, with fewer children to contribute to the overall activity levels, potentially making behavior management easier and increasing the likelihood that children have freedom to engage in self-selected, developmentally appropriate activities and cooperative play (Howes, Phillips, & Whitebook, 1992). In sum, although child–teacher ratios and class size are interrelated, they are likely to shape child outcomes in different ways, and it is important to understand the independent impact of each characteristic.

In spite of their theoretical importance, the empirical evidence to date about the benefits for child outcomes of small class sizes and low child–teacher ratios in early childhood classrooms has been mixed and largely inconclusive. One difficulty is that many studies evaluate an existing program model, and within a program there may be little to no variation in dimensions of structural quality. With a few noted exceptions, researchers have not tried to directly manipulate structural aspects of quality. As a result, researchers have turned to large datasets that collect observational measures of program quality across several different program models and consider relationships with child outcomes. A review of the findings from each of these lines of research follows.

In the only experimental study of early childhood settings, Ruopp (1979) randomly assigned 3- and 4-year-old children to preschool classrooms with different child–teacher ratios and different class sizes. Classes with the following combinations of child–teacher ratios and class size were compared: no less than 1:7, no larger than 14; no less than 1:8, no larger than 16; no less than 1:9, no larger than 18; and less than 1:9 but larger than 18. Children assigned to classrooms of smaller size and ratios achieved greater gains on measures of receptive language, general knowledge, cooperative behavior, and verbal initiations, and exhibited less hostility and conflict when compared with groups with larger class sizes and ratios. Children in classes larger than 18 with ratios smaller than 1:9 showed the smallest gains on these outcomes. As a result, Ruopp assigned the greatest significance to the differences in class sizes, acknowledging that ratio is a related construct. Confirming a theoretical model of ECE, this study has served as the primary empirical evidence of the benefits of small classes in the field of ECE for decades.

Perhaps the most well-known study of class size was conducted in Tennessee's early elementary schools in the 1980s, the STAR class size experiment. This experimental evaluation found that small kindergarten classes (of 13–17 children) had small positive impacts on a variety of math and literacy skills relative to larger classes with correspondingly lower child–teacher ratios (Mosteller, 1995).

To date, no other experimental work has been done on this topic, but two recent observational studies (Howes et al., 2008; Mashburn et al., 2008; Pianta et al., 2005) scrutinized a range of structural and process indicators, including class size and child–teacher ratios, as predictors of academic, language, and social learning among 4-year-old children attending state prekindergarten programs. In these studies, neither child–teacher ratios nor class sizes consistently predicted children's growth across cognitive and preacademic measures. However, class size was operationalized as a dichotomous indicator for whether class size was 20 or lower and the ratio was 1:10 or higher, and there was little variation in the sample (more than 80% fell into the small class and low ratio categories).

The findings from observational studies of more general child care arrangements among children of varying ages do not provide a clear picture, although this may not be surprising given the diversity of care settings and ages of children being studied. Studies of the NICHD Study of Early Child Care and Youth Development (NICHD & Duncan, 2003; NICHD Early Childhood Research Network, 1999, 2002; Phillipsen et al., 1997) have reported modest relationships between both class size and child–teacher ratio and child cognitive, academic, and behavioral outcomes modeled in a linear and dichotomous manner, but these findings have not been consistent (some analyses found relationships with class size, others with child–teacher ratio). When child–teacher ratio was considered

in a linear form, no significant relationship with the majority of child outcomes was found, with the exception of a positive and significant relationship with children's academic achievement at 54 months (NICHD & Duncan, 2003). Similarly, when child–teacher ratio and class size were modeled by whether they met quality standards or not, the only significant difference found was in behavioral problems and positive social behavior in the expected direction (NICHD Early Childhood Research Network, 1999). Blau's (1999) analysis of nationally representative data found no consistent association between mothers' reports of infant/toddler class sizes or ratios with reading, math, and language assessments or maternal reports of problem behavior at ages 4 or 5. However, smaller group size during the preschool years only was associated with positive effects on the same outcomes in this study. Related nonexperimental research on kindergarten class size also suggests no association between class size and academic or behavioral assessments (Milesi & Gamoran, 2006).

One limitation in the nonexperimental studies is that they have not carefully examined the functional form of the association between class size and child outcomes, and for the most part have used dichotomous measures of "large" versus "small" sizes for child–teacher ratio or assumed a linear relationship with child outcomes. Yet, there are reasons to think that the associations between class size and child–teacher ratio may be nonlinear, with larger associations at the low (higher quality) end of the distributions. Some evidence from an evaluation of 11 state prekindergarten programs suggests that an observational measure of classroom process quality (i.e., CLASS) has this type of nonlinear association, such that increments at the higher end of the quality scale show a significant relationship to child outcomes, but increments at the lower end of the quality scale show no relationship (Burchinal, Vandergrift, Pianta, & Mashburn, 2010). Similarly, when relationships between child care quality and adolescent outcomes in the NICHD sample were evaluated, quality showed a quadratic relationship with cognitive and academic achievement at age 15, such that a positive relationship between quality and cognitive academic outcomes was stronger for quality in the moderately high levels than at low levels of quality (Vandell et al., 2010).

It is possible that similar differences exist in the relationship between child outcomes and class size and ratio. If child–teacher ratios are efficacious because teachers engage more sensitively with more children in classes with small ratios, it is possible that up to a certain point, the demands of classroom management limit their ability to do so, and only with very small ratios do teachers have time to attend to children individually. Similarly, if the effect of class size is due to improved student engagement in the classroom, it is possible that only in particularly small classes do students feel connected enough and visible enough to the teacher to change their behavior. A recent study of primary and secondary school classrooms in Great Britain found a quadratic relationship between class size and the extent to which pupils received the focus of teachers' attention, with greater positive changes at the lower ends of the distribution and greater negative changes at the higher ends of the distribution (Blatchford et al., 2011). Additionally, although not a study of early childhood classrooms, Glass and Smith's (1979) meta-analysis of the relationship between class sizes and achievement in elementary through university classrooms found that the differences in achievement between smaller classes (e.g., 10 vs. 20 students) were much greater than the differences between much larger classes (e.g., 30 vs. 40 students), where there were no meaningful differences in achievement.

Understanding whether class size and child–teacher ratio are associated with child outcomes in center-based, early childhood classrooms, and, if so, whether the association is uniform across the range of class sizes and child–teacher ratios is important, as changes in class size and child–teacher ratio are expensive to implement. Estimates of the size of the relationship, if it exists, are also important, as relationships that have been found in the past have been modest. Some researchers have argued that as smaller class sizes and lower child–teacher ratios have generally been found to have very small effects in studies of elementary schools, the magnitude of effects of other interventions, such as increased investment in teachers, may be greater and easier to implement. Such efforts carry less risk of iatragenic effects, such as might result from the hiring of poorly trained teachers, also associated

with efforts to reduce class sizes (Chingos, 2012, 2013; Cho, Glewwe, & Whitler, 2012).

In this study, we build on prior research in two important ways. First, we conduct a meta-analysis to analyze variation in structural indicators across a broad set of center-based ECE program evaluations to better understand the relationship between class size and child–teacher ratio and children's cognitive, achievement, and socioemotional outcomes. Meta-analysis is an analytic technique that is commonly used to synthesize findings across evaluations to understand how features of the program affect the magnitude of the program's effects. Interestingly, neither overall group size nor child–teacher ratio have been examined in prior meta-analyses of ECE programs (Camilli et al., 2010; Gorey, 2001; Nelson, Westhues, & MacLeod, 2003). Second, we pay particular attention to the functional form of the association, and examine possible nonlinearities in the associations between class size and child–teacher ratio and children's outcomes.

Specifically, we answer the following questions: Are smaller class sizes in center-based ECE programs related to larger effect sizes for children's cognitive, achievement, and socioemotional outcomes and do these relationships vary across the distribution of class size? Are smaller child–teacher ratios related to larger effect sizes for children's cognitive, achievement, and socioemotional outcomes and do these relationships vary across the distribution of child–teacher ratios? Based on prior studies, we expect that both class size and child–teacher ratios will predict children's outcomes. We expect that the associations may be stronger at the lower end of the distribution, such that achieving much smaller classes and ratios is more strongly associated with positive program impacts than reducing large classes or ratios.

Method

The analyses reported in this article are drawn from a comprehensive database of U.S. ECE program evaluations published between 1960 and 2007. For our study, we drew from this sample only evaluations of center-based ECE programs for children aged 3 to 5 years old that provided a dosage of at least 10 hours a week of ECE programs for 4 or more months. Each evaluation study publication in the database is coded so that program impact estimates are transformed into effect sizes that represent differences in outcomes between treatment and control groups as a fraction of the control-group standard deviation. These effect sizes can then be used to estimate average effects across evaluation studies and to explore the differential impact of various program characteristics. We limited our sample to evaluation studies and contrasts that compared a center-based ECE program to a comparison group that was not assigned to an ECE program (a passive control group) and reflected the effect of the program on all participants. Our final sample, drawn from the larger database described below, was comprised of 38 studies, 53 contrasts, and 328 effect sizes.

Our sample of studies is derived from a larger database of published and unpublished studies of early education programs targeting infants through 5-year-olds constructed by the National Forum on Early Childhood Policy and Programs. The Forum's database builds on and updates previous meta-analytic databases (Camilli et al., 2010; Jacob, Creps, & Boulay, 2004; Layzer, Goodson, Bernstein, & Price, 2001). The Forum's research team identified additional studies through exhaustive keyword searches in ERIC, PsychINFO, EconLit, and Dissertation Abstracts databases; manual searches of leading policy institutes (e.g., Abt, Rand, Mathematica Policy Research, NIEER) and state and federal departments (e.g., U.S. Department of Health and Human Services); and a review of references from key ECE summaries and included studies.

Study screening criteria were designed to identify high-quality experimental and quasi-experimental studies with comparable experimental treatment and control groups. To be included in the Forum's meta-analytic database, studies needed to evaluate their programs using a comparison group, have at least 10 participants in each condition at follow-up, and experience less than 50% attrition between the initiation of treatment and time of measurement. The database includes experimental studies, those that employed a high-quality quasi-experimental design, and those that demonstrated baseline equivalency of treatment and control participants. These methodological criteria are more rigorous than those applied by McKey et al. (1985) and

Abt/NIEER. For example, the Forum's criteria exclude all pre-post-only (no comparison group) studies, as well as regression-based studies in which the baseline equivalence of treatment and control groups was not investigated. Using these criteria, the Forum's research team reviewed more than 10,000 reports captured by the search terms (those collected included evaluation reports, as well as many other types of articles, such as program descriptions and recommendations for practitioners) and identified 272 studies for inclusion in the overall database of ECE programs serving children birth to age 5 years old.

A team of nine graduate research assistants used a coding protocol to document relevant information regarding study design, program and sample characteristics, and statistical information needed to compute effect sizes. Prior to coding independently, the Forum implemented a training process that included practice coding, assessing reliability, and holding regular meetings (Wilson, 2009). Specifically, research assistants were trained during a 3- to 6-month process, during which time an overview of the project was provided; each item in the codebook was discussed; a manualized effect size training was completed; and a sample of studies was coded alongside an experienced coder. Trained coders were then required to achieve an interrater reliability agreement rate of 1.00 for effect sizes and .80 for all other study information with a master coder, based on the procedure used in the meta-analytic database the Forum's database built upon (see Camilli et al., 2010). The range of interrater agreements for all study information was .87 to .96. Any discrepancies or questions were resolved through weekly meetings between coders and principal investigators, and decisions were kept in an annotated codebook to ensure that decisions made about any ambiguities during these meetings were followed throughout the coding process (Cooper & Hedges, 2009).

The authors of this article took two further steps to ensure high-quality data. First, at the conclusion of the construction of the database, data entry was thoroughly checked and cleaned. This process included checking outliers, confirming skip patterns, and examining missing values. Then, the program information provided about each study included in this analysis was checked by the authors to ensure accurate coding and to determine if any missing information could be inferred from the information available (for example, state regulations were considered when coding class size and child–teacher ratios for studies of state prekindergarten programs).

The database consists of three levels of nested data: study, contrast, and effect size. Studies are defined as independent investigations of ECE programs. Contrasts are comparisons of groups that experienced different conditions within a study. For these analyses, we were interested in all comparisons between one group of children provided with center-based ECE programs and another who were not provided with ECE (although in most cases these participants were free to seek other services). Most studies only reported on one contrast of interest, but in some cases one study provided information on more than one contrast (see the Appendix, available in the online version of the journal). For example, in some cases two different groups were provided with ECE using two different curricula and each was compared to a control group, or different cohorts of children were analyzed separately. Effect sizes are standardized comparisons of these treatment and control groups on a set of outcome measures. Subcontrasts, which provide more detailed information on effect sizes for a subgroup of a main contrast, for example, by gender or race, were not used in this study.

The Forum's research team coded each study's outcome measures into standardized mean difference effect sizes, computed using the Comprehensive Meta-Analysis computer software program (Borenstein, Hedges, Higgins, & Rothstein, 2005). Effect sizes were converted to Hedges's g, an effect size statistic that adjusts the standardized mean difference (Cohen's d) to account for bias in the d estimator when sample sizes are small. Because single contrasts frequently provided multiple effect sizes with varying levels of precision, in our analyses, we weight effect sizes by the inverse of the variance of each effect size multiplied by the inverse of the number of effect sizes per contrast (Cooper & Hedges, 2009; Lipsey & Wilson, 2001).

Analytic Sample

Our analytic sample focused on programs that provided center-based early care and education

to children aged 3 to 5 for 6 months or more for at least 10 hours a week. We included all studies of specified program models serving children of these ages, but excluded programs serving younger children and studies of nonspecific heterogeneous early care and education (analyses of national surveys in which parents reported on whether their child attended center-based education or not, but did not include details of the program attended). We limited the sample to evaluation studies that compared a center-based ECE program to a comparison group that was not assigned to an ECE program (a passive control group). We did not include studies that compared two alternative ECE programs, because the comparisons available did not differ on class size or child–teacher ratios and therefore could not inform our question of interest.

Effect sizes capture the overall impact of each program compared to a passive control group, and our analyses estimate differences in the effect size impacts by the reported class sizes and child–teacher ratios. In calculating child–teacher ratios, we only included counts of teachers, not volunteers or parents in the classroom. We included all effect sizes representing measures of children's cognitive, achievement, and socioemotional skills taken between the time the child received two thirds of the intended treatment and up to 1 year after treatment ended.

The resulting sample comprised of 38 studies, 53 contrasts, and 328 effect sizes. Within this sample, 270 effect sizes (within 50 contrasts) were found for cognitive and achievement outcomes and 58 effect sizes (within 20 contrasts) were found for socioemotional and behavioral outcomes. Of the 53 contrasts included in this study, three only included socioemotional effect sizes, 33 only included cognitive and achievement effect sizes, and 17 included outcomes in both domains.

Measures

Dependent Variables: Effect Sizes. We examined measures of program impacts in two domains. The first and largest category consists of measures of children's cognitive and achievement (preacademic) skills. Our analytic sample for this domain includes 270 effect sizes drawn from 50 contrasts. This category includes measures of IQ; vocabulary, visual, spatial, and auditory skills; as well as letter recognition and early math skills. The measures include both direct assessments and teacher and parent reports on children's skills. Two of the cognitive effect sizes reported had values greater than 1.5. As is typically done in meta-analysis to ensure that these outliers did not exert undue influence on our analyses, we top-coded these large effect sizes at 1.5 (Lipsey & Wilson, 2001).

The second, much smaller category includes measures of social and emotional capacities. Our analytic sample for this domain includes 58 effect sizes drawn from 20 contrasts. The vast majority of dependent measures in this category assess children's social competence and cooperation, but it also includes measures of emotional expression and behavior problems. These measures include direct assessments, observations, and parent and teacher reports of behavior.

Sometimes evaluation reports provided insufficient information from which to calculate precise effect sizes. Observations (n = 153 for cognitive outcomes and n = 57 for socioemotional outcomes) for which both effect size and significance information were missing were deleted from the primary analyses (resulting in the sample sizes reported above). In cases in which some information was provided, when possible, we used information on the direction and significance of effect sizes to calculate the missing effect sizes. Estimates of missing effect sizes were calculated assuming a p value of .05, thus representing the minimum plausible effect size, a conservative approach. This was done for 40 observations that were missing effect sizes but for which the authors indicated that the difference between treatment and control was significantly different than 0 and noted the direction of the effect. This assumption, along with other statistical information provided by the reports, enabled us to estimate effect sizes for these outcomes. The robustness of our findings to differing assumptions about missing effect sizes was checked through additional analyses.

Independent Variables: Class Size and Child–Teacher Ratio. Child–teacher ratios and class sizes were continuous variables derived from values reported in the studies. Only paid program staff members were considered teachers and

were included in the calculation of this ratio; parent or other adult volunteers may have been present in some programs but were not counted in the ratio. The class size and child–teacher ratio variables were rescaled to center on the lowest value present in our sample (11 for class size, 5 for child–teacher ratio). The qualifications of teachers varied greatly, from teachers with master's degrees to those without any professional training, and classrooms with more than one teacher often had staff with a mix of qualifications. When this information was not provided in the report, other publicly available sources, such as published program guidelines, were consulted to generate reasonable estimates of these values.

Despite these efforts, information on class sizes and child–teacher ratios could not be obtained for 34% of the program class sizes and 38% of child–teacher ratios. To control for missing information, we used a dummy variable approach (Puma, Olsen, Bell, & Price, 2009). If child–teacher ratio or class size was missing, these variables were set to 0 and a dummy variable that indicated missingness was also included in the model. The coefficients associated with these missing variables can be interpreted to reflect the difference in the predicted mean effect size between the programs missing data for each variable and that for the programs with the lowest class size and ratio.

We also included variables that controlled for other potentially important aspects of structural quality: the use of a standardized curriculum, staff education (whether they had a degree focusing on ECE or whether they had a BA or higher), and the provision of staff training. In each case, these variables were dichotomous and coded as one if there was evidence that the program provided each criteria and zero if there was no evidence the program had done so.

We included in our regressions a set of covariates to control for effect size variation resulting from differences in other program, participant, and study design characteristics. Variables were chosen that have shown significant relationships with effect sizes in prior meta-analyses (Shager et al., 2013). Continuous variables controlled for the length of the program, the average age of children on program entrance, and the percentage of the sample classified as a racial or ethnic minority. The variables for start age and percent minority were mean centered, while the length of program was centered on 10 months. Dichotomous variables captured whether the program was provided after 1980, whether the control group had access to other services in the community (active control group), whether baseline covariates were included in the analysis that generated the effect sizes for the program, and, for cognitive and achievement outcomes, whether a performance test or some other type of measure generated outcomes. We also created indicators for contrasts that evaluated Head Start programs and those that evaluated public prekindergarten programs (the reference category reflected experimental programs that have been evaluated). In cases in which data were missing for any of these variables, we used a dummy variable to indicate missingness for each variable (Puma et al., 2009).

Research Approach

Due to the nested nature of the effect-size data (i.e., effect sizes are clustered within contrasts, which in turn are clustered within studies), we employed multilevel modeling procedures. We estimated a two-level model, with Level 1 modeling effect sizes and Level 2 modeling contrasts. We do not estimate a third level of studies for both theoretical and practical reasons. First, multiple contrasts typically arose from multiple treatment arms with different groups of children or different cohorts of children, and we expected differences in effect sizes between contrasts (even within the same study) to be more consequential and important to capture in our modeling than differences between studies. Second, more than half of the studies consisted of only one contrast (and, in most others, no more than two contrasts), and data would not support consistent estimation of a three-level model.

The Level 1 model (effect size level) is as follows:

$$ES_{ij} = \pi_{0j} + \pi_{1j}x_{1ij} + \ \ldots \ + \pi_{kj}x_{kij} + e_{ij}, \qquad (1)$$

where effect size i in contrast j is modeled as a function of the intercept (π_{0j}), which represents the average (covariate adjusted) effect size for contrast j; k independent variables measured at the effect size level ($\pi_{1j}x_{1ij} + \ldots + \pi_{kj}x_{kij}$); and a within-contrast error term (e_{ij}). The Level 2

equation (contrast level) models the intercept as a function of the grand mean effect size (β_{00}), p independent variables measured at the contrast level ($\beta_{01}x_{1j} + \ldots + \beta_{0p}x_{pj}$) and a between-contrast random error term (u_{0j}):

$$\pi_{0j} = \beta_{00} + \beta_{01}x_{1j} + \ \ldots \ + \beta_{0p}x_{pj} + u_{0j}. \qquad (2)$$

The "mixed effects" model can also be expressed in one equation by substituting (2) into (1). We conduct our analyses in SAS, using the PROC MIXED procedure.

Our first set of analyses estimated the predictive linear associations of class size and teacher–child ratio without holding constant other study or program covariates. We began by estimating regression models with the class size and the child–teacher ratio variables entered individually and then together in a simple linear model. Both class size and child–teacher ratio were centered on the first value in the respective range (5 for class size and 11 for ratio). We then added controls for program, study, measure, and participant characteristics, to evaluate the robustness of our findings.

Our second set of analyses was designed to test alternative functional forms of the association between the key independent and dependent variables. In these analyses, we estimated separate linear spline regressions at predetermined inflection points in the distribution corresponding to small and large class sizes and child–teacher ratios. A dummy variable was set at 0 if the class size or ratio was lower than the cutoff score (or knot) and 1 if higher than or equal to the cutoff, and interacted with the difference between the ratio or class size and the cutoff score. The Level 2 equation then is as follows if knot = t(using class size as an example):

$$\begin{aligned}\pi_{0j} = {} & \beta_{00} + \beta_{01}\text{Class size}_i + \\ & \beta_{02}\text{Large class size}_j \times \\ & \left(\text{Class size}_j - t\right) \ + \ \ldots \ + \\ & \beta_{0p}x_{pj} + u_{0j},\end{aligned} \qquad (3)$$

β_{00} represents the estimated mean effect size for the smallest observed class size. β_{01} then represents the slope for class sizes smaller than and equal to the knot (t) and β_{02} the difference in slope for those with larger class sizes. The knots used to define small and large class sizes and child–teacher ratios were identified based on hypotheses applied to the distribution of data. Knots at the median, the lower third, and the lower quartile of the distribution were tested. We first tested these relationships with class size and the child–teacher ratio variables entered individually and then together into the model. We then added controls for program, study, measure, and participant characteristics, to evaluate the robustness of our findings.

Results

Table 1 displays the characteristics of our full sample of studies, including both studies that assessed cognitive and achievement outcomes, and those that assessed socioemotional and behavioral outcomes. The samples of studies used in the analysis of each of these types of dependent measures varied slightly: 82% of our analytic sample included measures of cognitive and achievement outcomes, whereas only 18% included measures of socioemotional outcomes. The only significant difference between the sample of ECE programs with cognitive and achievement outcomes and those with socioemotional outcomes was that cognitive and achievement dependent measures were more likely to be performance tests whereas socioemotional dependent measures were more likely to be observations or ratings of the child.

In our sample of studies, child–teacher ratios ranged from 5:1 to 15:1 with a mean close to 9:1. Class sizes ranged from 11 to 25, with a mean of about 17. Children were, on average, 49.41 months old when they entered the programs studied, and the average program length was 11.02 months. Many of these programs served a significant population of racial or ethnic minority children (average of 61% African American and/or Latino/Hispanic children). Thirty to forty percent of programs reported the use of a standardized curriculum or hiring teachers with bachelors' degrees or ECE certification, whereas 19% provided staff training. Forty percent of the studies evaluated a Head Start program, and 43% evaluated a public prekindergarten program. Programs with lower class sizes and ratios tended to serve younger children ($r = .27$, $r = .24$, respectively). The sample of programs with low ratios and small

TABLE 1

Sample Sizes, Means, and Standard Deviations or Percentages of Variables

Variable	*n*	*M* (*SD*)/ Percentage
Contrast level	Contrasts	
Child–teacher ratio	33	8.90 (2.43)
Class size	35	16.60 (3.05)
Start age	49	49.41 (5.88)
Length (months)	53	11.02 (6.09)
Percent minority	39	61.00 (32.19)
Class size missing	53	34%
Child–teacher ratio missing	53	38%
Standardized curriculum	53	36%
Many staff have ECE degree	53	30%
Many staff have BA or higher	53	42%
Staff training provided	53	19%
Study done after 1980	53	49%
Active control group	53	26%
Baseline covariates included	53	25%
Start age missing	53	8%
Percent minority missing	53	26%
Head Start program	53	40%
Public prekindergarten	53	43%
Effect size level	Effect sizes	
Performance measure	328	72%
Cognitive/ achievement outcomes	328	82%
Socioemotional outcomes	328	18%

classes were slightly less likely to report large percentages of teachers with BA degrees or ECE certifications and to use a standardized curriculum. There were no notable differences across the program types in staff training. Programs missing data on class size or child–teacher ratios were more likely to lack information for other program variables as well. Examination of the cross tabulations for the variables detailing characteristics of the teaching staff, training, curriculum, ratios, and class size did not raise any concerns about collinearity.

The average inverse variance weighted effect size of all programs that provided center-based ECE services to children ages 3 to 5 years old was 0.31 (see Model 1, Table 2). This effect size was significantly different than zero ($p < .001$). A common concern in meta-analyses is publication bias, the potential of which can be explored by looking at a funnel plot. Our review of a funnel plot did not suggest any such bias. Moreover, a fail-safe *N* test indicated that 4,910 studies would need to be found to reduce the mean effect size to nonsignificance. Analyses revealed significant heterogeneity in these effect sizes ($Q = 44.07$, $p < .001$) and the I^2 statistic indicated that 65% of the variance across studies cannot be explained by chance.

The average inverse variance weighted effect size for socioemotional outcomes was 0.17 ($p < .01$). Again, a funnel plot did not raise concerns of publication bias; a fail-safe *N* test indicated that 147 additional studies with effect sizes of zero would be necessary to reduce the results to a nonsignificant difference from zero. There was significant heterogeneity in these effect sizes ($Q = 65.42$, $p < .001$) and the I^2 statistic indicated that 57% of the variance across studies cannot be explained by chance.

Cognitive and Achievement Outcomes

Do continuous measures of class size and child–teacher ratio predict ECE program impacts on cognitive and achievement outcomes? Analyses of the linear relationship between child–teacher ratio and class size, and cognitive and achievement outcomes showed that there were no significant linear relationships when these variables were entered either individually or in combination (see Table 2).

Is there support for nonlinear associations between class size and child–teacher ratio and children's cognitive and achievement outcomes? When different inflection points (or knots) were tested that model the slope of the linear association differently across the distribution, some

TABLE 2

Summary of Results From Hierarchical Linear Modeling (HLM) Models With Cognitive and Achievement Effect Sizes Predicted by Linear Specifications of Child–Teacher Ratio and Class Size, Standard Errors in Parentheses (N = 50 Contrasts, 270 Effect Sizes)

Variable	Model 1	Model 2	Model 3	Model 4
Child–teacher ratio		−0.02 (0.01)		−0.02 (0.01)
Class size			−0.01 (0.01)	0.00 (0.01)
Missing class size			−0.11 (0.10)	−0.26* (0.13)
Missing child–teacher ratio		−0.07 (0.08)		0.11 (0.13)
Intercept	0.31*** (0.03)	0.39*** (0.07)	0.39** (0.08)	0.44*** (0.10)

*$p < .05$. **$p < .01$. ***$p < .001$.

significant associations were detected between child–teacher ratio, as well as group size, and cognitive and achievement outcomes. Specifically, knots set at 30% of distribution or lower (5:1–7.5:1 or 12–15) resulted in significant coefficients for both the slope at the lower third of the distribution and the change in the slope above the knot. The slope for the top two thirds of the distribution (7.75:1–15:1 and 16–25) was estimated to be close to zero, whereas the slope in the lower third of the distribution was negative and significant (see Model 3, Table 3). These results were largely robust to the addition of controls to the model, although the slope of class size in the lowest third of the distribution became nonsignificant (see Model 4, Table 3). Figure 1 presents the distribution of effect sizes and the predicted slope.

For child–teacher ratios 7.5:1 and lower, the reduction of this ratio by one child per teacher predicts an effect size of 0.22 standard deviations greater. For class sizes 15 and smaller, one child fewer in the class predicts an effect size of 0.10 standard deviations larger. For larger child–teacher ratios and class sizes, there was no discernible relationship with cognitive and achievement outcomes. Figure 2 shows the distribution of effect sizes and predicted relationship between child–teacher ratio and cognitive and achievement outcomes.

Finally, we tested the robustness of these models to the addition of variables measuring other potentially important program characteristics that might be confounded with class size or child–teacher ratio. When controls for the use of a standardized curriculum, the education level of teachers, and the training of teachers were added, none of these additional variables showed a significant relationship with effect sizes. Table 4 presents the results of these analyses. The relationship between class size and child–teacher ratio and cognitive and achievement effect sizes remained stable. When controls identifying Head Start and public prekindergarten programs were added, the estimates changed slightly and class size was no longer significant; neither the Head Start nor the public prekindergarten program variables significantly predicted effect sizes.

Socioemotional Outcomes

Is there an association between child–teacher ratio and class sizes and children's socioemotional outcomes? And if so, what is the functional form of these associations? Due to the smaller sample of effect sizes derived from measures of socioemotional outcomes, statistical power was limited in models including the full set of controls that were used in the analyses of cognitive and achievement outcomes. For this reason, we focus on findings from models without controls (see Table 5 and Table 6), but also report information about models with controls.

Child–teacher ratio and class size were not significantly associated with effect sizes in the linear analyses (see Table 5). Child–teacher ratio remained a nonsignificant predictor of effect size in the spline analysis, although class size did show a modest and significant association at the lowest end of the distribution (Table 6). Tests of inflection points for class size at the mean of class size and below consistently resulted in

TABLE 3

Summary of Results From Hierarchical Linear Modeling (HLM) Models With Cognitive and Achievement Effect Sizes Predicted by Spline Specifications of Child–Teacher Ratios and Class Sizes, Standard Errors in Parentheses (N = 50 Contrasts, 270 Effect Sizes)

Variable	Model 1	Model 2	Model 3	Model 4
Child–teacher ratio, slope for ratios of 7.5 and less	−0.33*** (0.07)		−0.22* (0.09)	−0.26** (0.09)
Child–teacher ratio, difference in slope for ratios greater than 7.5	0.33*** (0.08)		0.22* (0.09)	0.28** (0.10)
Class size, slope for class sizes 15 and less		−0.17*** (0.03)	−0.10* (0.04)	−0.07~ (0.04)
Class size, difference in slope for class sizes greater than 15		0.20*** (0.03)	0.13** (0.05)	0.11* (0.05)
Missing class size		−0.62*** (0.12)	−0.45** (0.15)	−0.45** (0.22)
Missing child–teacher ratio	−0.78*** (0.18)		−0.42* (0.21)	−0.41~ (0.22)
After 1980				0.00 (0.06)
Percent minority				0.00 (0.00)
Domain (achievement effect sizes)				0.12*** (0.03)
Active control				−0.13* (0.06)
Baseline covariates				−0.11* (0.05)
Performance measure				−0.05 (0.05)
Start age				−0.01~ (0.00)
Length				−0.01** (0.00)
Missing start age				−0.17~ (0.09)
Missing percent minority				−0.08 (0.05)
Intercept	1.08*** (0.17)	0.89*** (0.11)	1.15*** (0.17)	1.10*** (0.18)

Note. These spline specifications allow for a discontinuity in the linear slope of the estimated associations, with the inflection point modeled at the 30% of the class size (15) and of child–teacher ratio (7.5) distribution. "Difference in Slope" estimates indicate how the linear association changes at the inflection point specified (7.5 or 15). As a result, for points higher than the inflection, the overall slope is the sum of the "Slope" and "Difference in Slope" estimates.
~$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

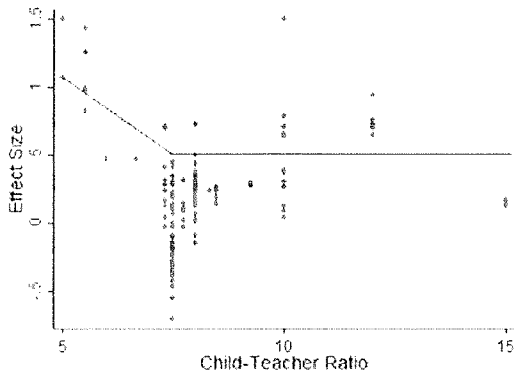

FIGURE 1. *Scatterplot showing the relationship between child–teacher ratio and cognitive and achievement effect sizes, with the regression results of the spline analysis.*

significant associations at the lower end of the range and a slope close to zero at the higher end of the range. To be consistent with other models reported, the results of the spline analysis using an inflection point at the lower third of the distribution are presented here. Although there is no way to precisely identify the exact point of inflection that is the best fit for the relationship between class size and socioemotional outcomes, our findings indicate that the association between class size and socioemotional outcomes is stronger at the lower end of the distribution than the higher.

When all the controls used in our analyses of cognitive and achievement outcomes were added, the slope of class size and difference in slope at

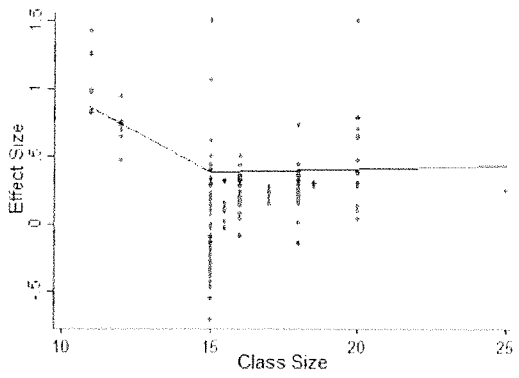

FIGURE 2. *Scatterplot showing the relationship between class size and cognitive and achievement effect sizes, with the regression results of the spline analysis.*

the high end of the distribution both become smaller and nonsignificant (Model 4, Table 6). When the small sample of studies contributing to these findings was examined closely, it appeared that one evaluation of a program with a small class size (12) and larger child–teacher ratio (12:1) was largely responsible for producing this pattern of findings. The rest of the effect sizes were smaller on average and came from programs with classes greater than 15 (see Figure 3).

Robustness Checks

To confirm that the pattern of findings we presented was robust, we tested a variety of alternative specifications. First, we were concerned that programs with very small class sizes and ratios might be unduly influencing our findings. Therefore, we eliminated the effect sizes from the contrasts with smallest class size and the lowest child–teacher ratio (The Perry Preschool and Karnes Ameliorative Curriculum Pre-K) and reestimated our spline regression models. With controls, the magnitudes of the resulting associations were similar to those presented in Table 3 and all were significant. Without controls, child–teacher ratio did not show an association with effect sizes, but the association with class size remained a similar size and was significant.

Our study inclusion criteria was carefully considered, but we wanted to be sure that the inclusion of studies in which coders noted the potential for bias in estimated program impacts (both experimental and nonexperimental designs) did not bias our findings. We dropped all studies that coders rated with more than one potential bias. The studies dropped included both quasi-experimental and random assignment studies, suggesting that in our sample the implementation quality of the research method was important for understanding bias in estimates of ECE program effectiveness. Without controls, the parameter estimate for child–teacher ratio was smaller and nonsignificant, whereas the estimate for class size was smaller but remained significant. When all controls were included in the spline regression analysis, the findings were consistent with Table 3.

The decision to utilize a two-level model with contrasts at Level 2 might raise concerns about nonindependence in results between two contrasts from the same study. To check if this was influencing our results, we also reran the analyses with the study as the Level 2 unit (effect sizes nested within studies instead of nested within contrasts). The size and direction of coefficients was consistent with our original findings.

Meta-analysis combines results from measures that vary in numerous and possibly consequential ways. To ensure that measures without strong psychometric properties, and specifically evaluation researcher designed measures, were not biasing our findings, we estimated our analyses including only measures for which we could find reliability data, either from the study population or a normed sample. The relationship between child–teacher ratio and effect sizes was much larger and significant in this sample, but the relationship with class size was smaller and nonsignificant.

Our analysis focused on short-term outcomes and used effect sizes measured up to 12 months after program completion, but not later follow-ups. To test whether the results would differ if we included later follow-up time points, we ran the same analyses on the sample of programs reporting outcomes between 12 months and 72 months and the sample of programs with outcomes after 12 months (which was characterized by a large number of outcomes from a small number of studies that followed children into middle childhood and adulthood). The association between class size and child–teacher ratio is smaller and nonsignificant in these models, suggesting that

TABLE 4

Summary of Hierarchical Linear Modeling (HLM) Models of Cognitive Outcomes Predicted by Spline Specifications of Child–Teacher Ratios and Class Sizes and Other Program Characteristics (With Controls for Study Design and Participant Characteristics Not Shown), Standard Errors in Parentheses (N = 20 Contrasts, 58 Effect Sizes)

Variable	Model 1	Model 2
Child–teacher ratio, slope for ratios of 7.5 and less	−0.24** (0.09)	−0.30** (0.10)
Child–teacher ratio, difference in slope for ratios greater than 7.5	0.27** (0.09)	0.32** (0.11)
Class size, slope for class sizes 15 and less	−0.09* (0.04)	−0.06 (0.04)
Class size, difference in slope for class sizes greater than 15	0.11* (0.05)	0.09~ (0.05)
Missing class size	−0.48** (0.14)	−0.43~ (0.16)
Missing child–teacher ratio	−0.37~ (0.21)	−0.49* (0.24)
Standardized curriculum	−0.07 (0.04)	
Majority teachers with ECE degree	0.18~ (0.09)	
Majority teachers with BA degree	0.04 (0.11)	
Teacher training provided	−0.01 (0.07)	
Head Start program		0.06 (0.13)
Public prekindergarten program		−0.05 (0.14)
Other study design and participant controls		
Intercept	1.14*** (0.17)	1.11*** (0.18)

Note. These spline specifications allow for a discontinuity in the linear slope of the estimated associations, with the inflection point modeled at the 30% of the class size (15) and of child–teacher ratio (7.5) distribution. "Difference in Slope" estimates indicate how the linear association changes at the inflection point specified (7.5 or 15). As a result, for points higher than the inflection, the overall slope is the sum of the "Slope" and "Difference in Slope" estimates. ECE = early childhood education. ~$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 5

Summary of Results From Hierarchical Linear Modeling (HLM) Models of Socioemotional and Behavioral Outcomes Predicted by Linear Specification of Class Size and Child–Teacher Ratio (N = 20 Contrasts, 58 Effect Sizes)

Variable	Model 1	Model 2	Model 3	Model 4
Child–teacher ratio		0.11~ (0.05)		0.09 (0.05)
Class size			−0.05~ (0.03)	−0.04 (0.03)
Missing class size and child–teacher ratio		0.45* (0.20)	−0.23 (0.19)	0.14 (0.30)
Intercept	0.17** (0.05)	−0.23 (0.18)	0.45* (0.18)	0.08 (0.29)

~$p < .10$. *$p < .05$. **$p < .01$.

any relationship with child outcomes is a short-term one.

Our analyses of socioemotional outcomes combined measures of behavior problems and social skills, which, given their differing emphasis on negative and positive behavioral repertoires, might show different relationships with class size and ratio. To explore this issue, we estimated the same models dropping all measures of behavior problems and focusing instead on the set of social skills (the sample of behavioral outcomes was too small to analyze on its own). The direction and pattern of the results was the same, although some of the coefficients were larger. This confirmed that our approach of combining measures was not biasing our findings, but given the instability of models once we added covariates, we choose to present results from analyses of the larger sample with greater statistical power.

Finally, we estimated models using samples in which missing effect sizes were estimated three

TABLE 6

Summary of Hierarchical Linear Modeling (HLM) Models of Socioemotional and Behavioral Outcomes Predicted by Spline Specifications of Child–Teacher Ratios and Class Sizes, Standard Errors in Parentheses (N = 20 Contrasts, 58 Effect Sizes)

Variable	Model 1	Model 2	Model 3	Model 4
Child–teacher ratio, slope for ratios of 7.5 and less	−0.23 (0.30)			
Child–teacher ratio, difference in slope for ratios greater than 7.5	0.36 (0.31)			
Child–teacher ratio			−0.08 (0.06)	0.16 (0.09)
Class size, slope for class sizes 15 and less		−0.28** (0.07)	−0.39** (0.11)	−0.19 (0.14)
Class size, difference in slope for class sizes greater than 15		0.28** (0.09)	0.40** (0.13)	0.05 (0.18)
Missing class size and child–teacher ratio	−0.34 (0.72)	−0.98** (0.28)	−1.64* (0.56)	−0.34 (0.71)
After 1980				0.15 (0.09)
Percent minority				−0.00 (0.00)
Active control				−0.28~ (0.13)
Baseline covariates				0.15~ (0.07)
Start age				−0.02~ (0.01)
Length				0.10 (0.05)
Missing start age				0.71* (0.21)
Missing percent minority				0.36* (0.12)
Intercept	0.56 (0.72)	1.19** (0.28)	1.84** (0.55)	0.34 (0.74)

Note. Domain and performance measures were not included in the controls as there was not enough variance in the sample to support use in the analysis. These spline specifications allow for a discontinuity in the linear slope of the estimated associations, with the inflection point modeled at the 30% of the class size (15) and of child–teacher ratio (7.5) distribution. "Difference in Slope" estimates indicate how the linear association changes at the inflection point specified (7.5 or 15). As a result, for points higher than the inflection, the overall slope is the sum of the "Slope" and "Difference in Slope" estimates.
~$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

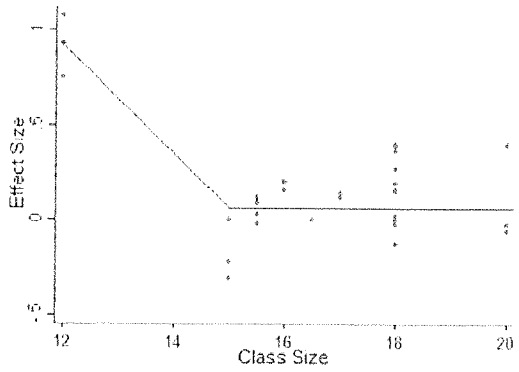

FIGURE 3. *Scatterplot showing the relationship between class size and socioemotional effect sizes, with the regression results of the spline analysis.*

different ways: (a) When the difference between the treatment and comparison groups was noted to be nonsignificant, the missing effect size was estimated as zero; (b) when a direction for the effect was given, the largest possible nonsignificant difference in the indicated direction was estimated (with p values of .11 if favoring the treatment group and .99 if favoring the comparison group); and (c) the smallest possible effect size was estimated (with p values of .99 if favoring the treatment group and .11 if favoring the comparison group). Our primary analyses were run with these three samples and we found that, although the size of the coefficients varied slightly, the pattern of significant results remained the same.

In summary, we did not find that the pattern of negative relationships between class size and child–teacher ratio at the low end of the distribution was unique to the inclusion of specific effect sizes, types of measures, quality of evaluation methods, specific programs, or our treatment of missing data. This pattern provides us with confidence that, given the existing data available for

the analyses, the pattern of findings is robust and is not dependent on specific analytic decisions.

Discussion

This study used almost 60 years of early childhood program evaluation research to identify the associations between child–teacher ratio and class size and children's cognitive, achievement, and socioemotional outcomes. The relationship between these early childhood program characteristics and children's outcomes has been difficult to discern in prior research, and our findings offer some possible insight into prior inconsistencies in results. The findings for cognitive and achievement outcomes were more robust than those for socioemotional outcomes, as the sample of programs assessing these outcomes was larger. For cognitive and achievement outcomes, child–teacher ratio and class size were associated with more positive outcomes for children at the lower end of the distribution of class size and child–teacher ratio; that is, only very low child–teacher ratios (7.5:1 and lower) or very small class sizes (15 or less) were associated with significant, although not large, differences for children's cognitive and achievement outcomes. Small changes in class size or ratio (the reduction by one child) in very small, well-staffed classrooms (i.e., 15 children and two teachers) were only associated with small effect sizes (0.22 and 0.10, respectively). For socioemotional outcomes, there was the suggestion that very small classes, but not child–teacher ratios, might be important, but we caution that our sample was too small to warrant confidence in our conclusions for these outcomes and the effect sizes were also small.

These findings do provide some tentative support to the hypothesis that both class size and child–teacher ratios shape children's classroom experiences and their learning processes in different ways. Child–teacher ratio was positively related to cognitive and achievement outcomes in classrooms with very low child–teacher ratios, which corresponds with the work suggesting that these ratios may play a role in facilitating high-quality interactions between teachers, both those that are emotionally supportive and cognitively stimulating (NICHD Early Childhood Research Network, 2002; Phillipsen et al., 1997). Class size was also significantly related to these outcomes in very small classes, even when controlling for the child–teacher ratio, providing support for the hypothesis that teachers in smaller classes may be more effective, whether through the provision of more developmentally appropriate activities, individualized instruction, or increased effectiveness keeping the class on task (Blatchford et al., 2011; Finn et al., 2003; Howes et al., 1992). However, the effect sizes for smaller classes and those with lower child–teacher ratios were modest and relationships with class size were not consistently significant, suggesting that incremental changes in class sizes and ratios would have limited use as a mechanism to improve center-based ECE effectiveness.

In this study, among programs with child–teacher ratios above 7.5:1 and class sizes above 15, there were no associations between these structural characteristics and child outcomes. These findings align with prior studies of older children (Chingos, 2012; Cho et al., 2012) and studies that find that only variation in the higher end of process quality predict improved child outcomes (Burchinal et al., 2010; Vandell et al., 2010). Class size and child–teacher ratio may both need to be very small to facilitate higher quality interactions at the level necessary to yield modest increases in child cognitive and achievement outcomes.

An important area of future research is to examine why lower class sizes and teacher–child ratios matter at the low end of these distributions, but not at the higher end. Indeed, most theoretical models of classroom quality do not a priori suggest that these should only matter at the low end. These findings also suggest that changes in class sizes and ratios may play a limited role in shaping the classroom ecology, and more work is also needed to understand how teachers adapt their interactions in smaller classes or those with lower child–teacher ratios. Qualitative research and classroom observations may be especially useful in this endeavor.

Given the expense of making substantial reductions to class sizes and child–teacher ratios, it is important to note that all of the programs in our sample provided child–teacher ratios and class sizes within a range that is considered current reasonable practice for 3- and 4-year-olds. Ratios were not larger than 15:1 (in fact in all but

one study, ratios were 10:1 or smaller) and class sizes were not larger than 25 (in all but one study these were smaller than 20). As a result, it may be that these features of programs are not especially effective targets of policies to increase the impacts of preschool programs on early learning. The costs associated with reducing class sizes and ratios to the point that would likely improve children's outcomes would be quite large and would require significant and large changes to typical standards found in QRIS systems, Head Start, or state prekindergarten regulations (Sabol, Soliday Hong, Pianta, & Burchinal, 2013). Perhaps more important, even if implemented, the expected resulting benefits from small changes in class sizes and ratios would also be relatively modest (0.22 if considering the ratio and 0.10 if considering class size). Thus, the high cost of the changes in staffing combined with the potentially small return to children's learning suggest that this would not be a cost-effective approach to improve children's early learning in preschool programs.

Class size and ratio are not the only structural program features that are often the target of ECE policies, but more theoretical and empirical work is needed to identify the cost-effective and policy-relevant aspects of classrooms that are likely to increase children's learning. Requiring higher educational degrees or credentials for ECE classroom teachers and a comprehensive curriculum may also be the target of improvement initiatives (Early et al., 2008; Howes et al., 2008; Mashburn et al., 2008). Confirming prior work in this area, our analyses found these characteristics did not predict effect sizes, with nonsignificant associations of −0.07, 0.04, 0.18, and −0.01 for the use of a standardized curriculum, the majority of teachers holding a BA, majority of teachers holding a degree with ECE specialization, and the provision of teacher training, respectively. Although teacher training is a potentially a cost-effective approach to improving child outcomes, more work is needed to better design effective training and professional development education that can be scaled up to produce meaningful effects on children's learning.

Limitations

Although this study synthesizes the results of nearly six decades of evaluations of center-based ECE programs, there are some important limitations to our findings to highlight. First, although the mean effect sizes reported here are drawn from impact evaluations, our analyses are correlational and cannot provide causal estimates of the impacts of different class sizes and ratios on child outcomes. We have attempted to rule out alternative plausible explanations for the estimated associations by controlling for relevant characteristics of the program evaluations, but it is possible that some unobserved characteristics remain confounded with class size or ratio, and are biasing our findings.

Second, the meta-analytic database we created does not include studies conducted after 2007. Recent center-based ECE evaluations are thus not represented in our findings, including studies of programs that have shown variable impacts on children, such as the recent regression discontinuity study of Boston's Prekindergarten program (Weiland & Yoshikawa, 2013) and the Tennessee Voluntary Prekindergarten (Lipsey, Hofer, Dong, Farran, & Bilbrey, 2013). However, it is worth noting that most of the more recent studies do not provide much additional variation in class sizes or ratios. For example, Boston and Tennessee had maximum class sizes of 22 (1:11 ratio) and 20 (1:10), respectively. Nor does our study include more recent longitudinal follow-ups to some of the evaluations included here, such as the Head Start Impact evaluation fifth-grade follow-up (Puma, Bell, Cook, & Heid, 2010). Future research should focus on what can be learned about class size and ratios from the most recent studies.

Finally, missing data was a challenge. Although we included studies with missing data and used a dummy variable approach to control for their influence on the results to ensure that we estimated all other covariates as accurately as possible, we cannot know for certain whether our findings would differ if this information was not missing. It is surprising that many studies do not provide this type of basic information about their classrooms in an evaluation study. We urge authors and editors to recognize the importance of this information for subsequent research efforts and thus include in published reports.

Conclusions

Based on our findings, we conclude that current regulations that hold class sizes at or below 20 and child–teacher ratios at or below 10:1 are

largely adequate for most children. There is no clear advantage to slight reductions in these numbers. We did find that very small and/or well-staffed classrooms might confer some small benefits for children's cognitive and academic learning. However, the programs in our sample that fell on the lowest end of the distribution tended to come from demonstration programs, like Perry Preschool, or programs that were designed to serve higher risk populations, such as Head Start. There are reasons to be skeptical that if implemented at scale such benefits would be seen. Studies in K–12 have found that large-scale efforts to reduce class sizes by hiring more teachers may yield even smaller benefits than found in demonstration studies because of the difficulty in finding a sufficiently skilled pool of new teachers (Jepsen & Rivkin, 2009; Milesi & Gamoran, 2006). Moreover, the costs of sustaining such small classes and ratios would be high and could well lead to a reduction in resources in other areas, like teacher training, that may affect classroom quality and the long-term impacts of ECE.

Authors' Note

Jocelyn Bonnes Bowne is now at the MA Department of Early Education and Care.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We are grateful to the following funders of the National Forum on Early Childhood Policy and Programs: the Birth to Five Policy Alliance, the Buffett Early Childhood Fund, Casey Family Programs, the McCormick Tribune Foundation, the Norlien Foundation, Harvard University, and an Anonymous Donor. We are also grateful to the Institute of Education Sciences, U.S. Department of Education for supporting this research (#R305A110035), to Abt Associates, Inc. and the National Institute for Early Education Research for making their data available to us. Yoshikawa's work on the article was partially supported by a grant from the NYU Abu Dhabi Research Institute to the Global TIES for Children Center at New York University.

References

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

*Abbott-Shim, M., Lambert, R., & McCarty, F. (2003). A comparison of school readiness outcomes for children randomly assigned to a Head Start program and the program's wait list. *Journal of Education for Students Placed at Risk, 8*, 191–214.

*Ametjian, A. (1965). *The effects of a preschool program upon the intellectual development and social competency of lower class children* (Unpublished doctoral dissertation). Stanford University, Palo Alto, CA.

*Barnett, W. S., Jung, K., Lamy, C., Wong, V., & Cook, T. (2007, March). *Effects of five state prekindergarten programs on early learning. Paper presented at the SRCD annual meeting*, Boston, MA.

Blatchford, P., Bassett, P., & Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher-pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. *Learning and Instruction, 21*, 715–730.

Blau, D. M. (1999). The effect of child care characteristics on child development. *The Journal of Human Resources, 34*, 786–822.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis* (Version 2). Englewood, NJ: Biostat.

*Bridge, C. A., Townley, K. F., Hemmeter, M. L., & de Mesquita, P. B. (1994). *Third party evaluation of the Kentucky Education Reform Act preschool programs*. Lexington: University of Kentucky, College of Education and College of Human Environmental Sciences.

*Bryant, D. M., Peisnar-Feinberg, E. S., & Clifford, R. M. (1993). *Evaluation of public preschool programs in North Carolina*. Chapel Hill: University of North Carolina.

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*, 140–153. doi:10.1080/10888690802199418

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*, 166–176.

Burgess, K., Chien, N., Morrissey, T., & Swenson, K. (2014). *Trends in the use of early care and education, 1995–2011: Descriptive analysis of child care arrangements from national survey data*

(Report from the Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services). Washington, DC: U.S. Department of Health and Human Services.

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record, 112*, 579–620.

*Chesterfield, R., Chavez, R., Chesterfield, K. B., Hayes-Latimer, K., LaBelle, T., Levine, H., . . . Watson, P. (1982). *An evaluation of the Head Start Bilingual Bicultural Curriculum Development Project. Final report* (Prepared for the Administration on Aging [DHHS]). Los Angeles, CA: Juarez & Associates.

Chingos, M. M. (2013). Class size and student outcomes: Research and policy implications. *Journal of Policy Analysis and Management, 32*, 411–438.

Chingos, M. M. (2012). The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. *Economics of Education Review, 31*, 543–562.

Cho, H., Glewwe, P., & Whitler, M. (2012). Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools. *Economics of Education Review, 31*, 77–95.

*Cicarelli, V. G., Cooper, W. H., & Granger, R. L. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development* (Vol. 2, Prepared for the Office of Economic Opportunity). Athens: Westinghouse Learning Corporation and Ohio University.

*Coffman, A. O., & Dunlap, J. M. (1968). *The effects of assessment and personalized programming on subsequent intellectual development of prekindergarten and kindergarten children* (Office of Education, U.S. Department of Health, Education and Welfare, Cooperative Research Project #6-1328). University City, MO: School District of University City.

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 3–17). New York, NY: Russell Sage Foundation.

*Di Lorenzo, L. T., Salter, R., & Brady, J. J. (1969). *Prekindergarten programs for educationally disadvantaged children* (Final report). New York: New York State Department of Education.

Early, D. M., Iruka, I. U., Ritchie, S., Barbarin, O. A., Winn, D.-M. C., Crawford, G. M., Frome, P. M., Clifford, R. M., Burchinal, M., Howes, C., Bryant, D. M., & Pianta, R. C. (2010). How do pre-kindergartners spend their time? Gender, ethnicity, and income as predictors of experiences in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 25*, 177–193.

Early, D. M., Bryant, D. M., Pianta, R. C., Clifford, R. M., Burchinal, M. R., Ritchie, S., & 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*, 174–195. doi:10.1016/j.ecresq.2006.04.004

*Erickson, E. L., McMillan, J., Bonnell, J., Hofman, L., & Callahan, O. D. (1969). *Experiments in Head Start and early education: The effects of teacher attitude and curriculum structure on preschool disadvantaged children* (Final report). Washington, DC: The Office of Economic Opportunity.

*Esteban, M. (1987). *A comparison of Head Start and non-Head Start reading readiness scores of low-income kindergarten children of Guam* (Doctoral dissertation). Retrieved from Dissertation Abstracts International. (UMI No. 8808677)

Finn, J. D., Pannozzo, G. M., & Achilles, C. M. (2003). The "why's" of class size: Student behavior in small classes. *Review of Educational Research, 73*, 321–368.

*Frede, E., Jung, K., Barnett, W. S., Lamy, C. E., & Figueras, A. (2007). *The Abbott preschool program longitudinal effects study (APPLES)*. New Brunswick, NJ: National Institute for Early Education Research.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of class size and achievement. *Educational Evaluation and Policy Analysis, 1*, 2–16.

Gorey, K. M. (2001). Early childhood education: A meta-analytic affirmation of the short- and long-term benefits of educational opportunity. *School Psychology Quarterly, 16*, 9–30.

*Gormley, W. T., & Gayer, T. (2005). Promoting school readiness in Oklahoma: An evaluation of Tulsa's pre-k program. *Journal of Human Resources, 40*, 533–558.

*Herzog, E., Newcomb, C. H., & Cisin, I. H. (1973). *Preschool and postscript: An evaluation of the inner-city program*. Washington, DC: Social Research Group, Washington University.

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*, 27–50.

Howes, C., Phillips, D. A., & Whitebook, M. (1992). Thresholds of quality: Implications for the social

development of children in center-based child care. *Child Development, 63*, 449–460.

Jacob, R. T., Creps, C. L., & Boulay, B. (2004). *Meta-analysis of research and evaluation studies in early childhood education.* Cambridge, MA: Abt Associates.

Jepsen, C., & Rivkin, S. (2009). Class size reduction and student achievement the potential tradeoff between teacher quality and class size. *Journal of Human Resources, 44*, 223–250.

*Karnes, M. B., Hodgins, A. S., Stoneburner, R. L., Studley, W. M., & Teska, J. A. (1968). Effects of a highly structured program of language development on intellectual functioning and psycholinguistic development of culturally disadvantaged three-year-olds. *The Journal of Special Education, 2*, 405–412.

*Kelly, E. J. (1970). *The New Nursery School Research Project: Evaluating the effectiveness of an open, responsive environment in achieving selected objectives of early childhood education* (Final report). Greeley: University of Colorado.

*Larsen, J. M. (1985). *Family influences on competence in low-risk preschool children.* Toronto, Ontario, Canada: Society for Research in Child Development.

*Larson, D. F. (1969). *The effects of a preschool experience upon intellectual functioning among four-year-old, white children in rural Minnesota.* Mankato: Minnesota State University, College of Education.

*Larsen, J. M., & Hite, C. H. (1983). The effects of preschool on educationally-advantaged children: First phases of a longitudinal study. *Intelligence, 7*, 345–352.

Layzer, J. I., Goodson, B. D., Bernstein, L., & Price, C. (2001). *National evaluation of family support programs, Volume A: The meta-analysis, final report.* Cambridge, MA: Abt Associates.

*Lee, V. E., Schnur, E., & Brooks-Gunn, J. (1988). Does Head Start work? A 1-year follow-up comparison of disadvantaged children attending Head Start, no preschool, and other preschool programs. *Developmental Psychology, 24*, 210–222.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: SAGE.

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013). *Evaluation of the Tennessee voluntary prekindergarten program: End of pre-k results from the randomized control design* (Research Report). Nashville, TN: Peabody Research Institute, Vanderbilt University.

Magnuson, K., & Shager, H. (2010). Early education: Progress and promise for low-income children. *Children and Youth Services Review, 32*, 1186–1198.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., . . . Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*, 732–749.

McKey, R. H., Condelli, L., Ganson, H., Barrett, B. J., McConkey, C., & Plantz, M. C. (1985). *The impact of Head Start on children, families and communities: Final report of the Head Start evaluation, synthesis and utilization project.* Washington, DC: CSR.

*McNamara, J. R. (1968). *Evaluation of the effects of Head Start experience in the area of self-concept, social skills, and language skills* (ED 028 832, Pre-Publication Draft). Miami, FL: Dade County Board of Public Instruction.

Milesi, C., & Gamoran, A. (2006). Effects of class size and instruction on Kindergarten achievement. *Educational Evaluation and Policy Analysis, 28*, 287–313.

*Miller, L. B., & Dyer, J. L. (1972). *Four preschool programs: Their dimensions and effects.* Washington, DC: Public Health Service.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5*, 113–127.

National Scientific Council on the Developing Child. (2004). *Young children develop in an environment of relationships* (Working Paper No. 1). Available from www.developingchild.harvard.edu

Nelson, G., Westhues, A., & MacLeod, J. (2003). A meta-analysis of longitudinal research on preschool prevention programs for children. *Prevention & Treatment, 6*, 31a.

NICHD Early Child Care Research Network. (1999). Child outcomes when child care center classes meet recommended standards for quality. *American Journal of Public Health, 89*, 1072–1077.

NICHD Early Child Care Research Network. (2002). Child-care structure → process → outcome: Direct and indirection effects of child-care quality on young children's development. *Psychological Science, 13*, 199–206.

NICHD & Duncan, G. J. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development, 74*, 1454–1475.

NICHD Early Child Care Research Network. (2004). Does class size in first grade relate to children's academic and social performance or observed classroom practices? *Developmental Psychology, 40*, 651–664.

*Nummedal, S. G., & Stern, C. (1971, February). *Head Start graduates: One year later. Paper presented*

at the annual meeting of the American Educational Research Association, New York, NY.

Phillipsen, L. C., Burchinal, M. R., Howes, C., & Cryer, D. (1997). The prediction of process quality from structural features of child care. *Early Childhood Research Quarterly, 12*, 281–303.

Pianta, R., 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? *Developmental Science, 9*, 144–159.

*Pilcher, L. C. (1994). *Georgia Prekindergarten Program evaluation [with] executive summary.* Atlanta: Georgia State University Early Childhood Education.

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start impact study: Final report.* Washington, DC: U.S. Department of Health and Human Services.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized trials.* Cambridge, MA: Abt Associates.

*Reynolds, A. J. (1995). One year of preschool intervention or two: Does it matter? *Early Childhood Research Quarterly, 10*, 1–31.

Ruopp, R. (1979). *Children at the center: Summary findings and their implications.* Cambridge, MA: Abt Books.

Sabol, T. J., Soliday Hong, S. L., Pianta, R. C., & Burchinal, M. R. (2013). Can rating Pre-K programs predict children's learning. *Science, 341*, 845–846.

Shager, H., Schindler, H., Magnuson, K., Duncan, G., Yoshikawa, H., & Hart, C. (2013). Can research design explain variation in head start research results? A meta-analysis. *Education Evaluation and Policy Analysis, 35*, 76–95.

*Smith, E. J., Pellin, B. J., & Agruso, S. A. (2003). *Bright beginnings: An effective literacy-focused PreK program for educationally disadvantaged four-year-old children.* Arlington, VA: Educational Research Service.

*Sontag, M., Sella, A. P., & Thorndike, R. L. (1969). The effect of Head Start training on the cognitive growth of disadvantaged children. *The Journal of Educational Research, 62*, 387–389.

*U.S. Department of Health and Human Services, Administration for Children and Families. (2005). *Head Start impact study: First year findings.* Washington, DC: Author.

U.S. Department of Health and Human Services. (1996). *Head Start program performance standards.* Washington, DC: Head Start Bureau.

*Vance, B. J. (1967). *The effect of preschool group experience on various language and social skills in disadvantaged children. Final report.* Retrieved from ERIC database. (ED019989)

Vandell, D. L., Belsky, J., Burchinal, M., Steinberg, L., Vandergrift, N., & NICHD Early Child Care Research Network. (2010). Do effects of early child care extend to age 15 years? Results from the NICHD study of early child care and youth development. *Child Development, 81*, 737–756.

*Warden, B. A. (1998). *A study to determine the effectiveness of preschool on kindergarten readiness and achievement* (Unpublished master's thesis). Salem-Teikyo University, NC.

*Weikart, D. P., Bond, J. T., & McNeil, J. T. (1978). *The Ypsilanti Perry Pre-School Project: Preschool years and longitudinal results through fourth grade. Monographs of the High/Scope Educational Research Foundation.* Ypsilanti, MI: High/Scope Press.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function and emotional skills. *Child Development, 84*, 2112–2130.

*Weisberg, H. I. (1974). *Short term cognitive effects of head start programs: A report on the third year of planned variation–1971-72.* Cambridge, MA: Huron Institute.

Wilson, D. (2009). Systematic coding. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 159–176). New York, NY: Russell Sage Foundation.

*Xiang, Z., & Schweinhart, L. J. (2002). *Effects five years later: The Michigan School Readiness Program evaluation through age 10* (Prepared for the Michigan State Board of Education). Ypsilanti, MI: High/Scope Educational Research Foundation.

*Zigler, E. F., Abelson, W. D., Trickett, P. K., & Seitz, V. (1982). Is an intervention program necessary in order to improve economically disadvantaged children's IQ scores? *Child Development, 53*, 340–348.

Authors

JOCELYN BONNES BOWNE is the director of Research and Preschool Expansion Grant Administration at the Massachusetts Department of Early Education and Care, 51 Sleeper Street, Boston, MA 02210; jbb961@mail.harvard.edu. Her research focuses on early childhood education policy.

KATHERINE A. MAGNUSON is a professor of social work at the University of Wisconsin-Madison, and the associate director of the Institute for Poverty Research; 1350 University Ave., Madison, WI 53796; kmagnuson@wisc.edu. Her research focuses on early childhood and social welfare policy.

HOLLY S. SCHINDLER is an assistant professor in the College of Education at the University of Washington, Miller Hall, Box 353600, Seattle, WA 98195-3600, USA, hschindl@uw.edu. Her research focuses on early childhood and family studies.

GREG J. DUNCAN is a Distinguished Professor in the School of Education at University of California, Irvine, 2056 Education, Mail code: 5500, Irvine CA 92697; gduncan@uci.edu. He has published extensively on child poverty and the importance of early academic skills, cognitive and emotional self-regulation as well as health in promoting children's eventual success in school and the labor market.

HIROKAZU YOSHIKAWA is the Courtney Sale Ross Professor of Globalization and Education at NYU Steinhardt and a university professor at NYU, and co-director of the Global TIES for Children center at NYU, hy2042@nyu.edu. He is a community and developmental psychologist who studies the effects of public policies and programs related to immigration, early childhood, and poverty reduction on children's development.

Manuscript received July 13, 2015
First revision received June 1, 2016
Second revision received October 30, 2016
Accepted December 20, 2016