Developmental Psychology
2014, Vol. 50, No. 12, 2559–2571

0012-1649/14/$12.00 http://dx.doi.org/10.1037/a0037995

Preschool Center Care Quality Effects on Academic Achievement: An Instrumental Variables Analysis

Anamarie Auger and George Farkas
University of California, Irvine

Margaret R. Burchinal
University of North Carolina–Chapel Hill

Greg J. Duncan and Deborah Lowe Vandell
University of California, Irvine

Much of child care research has focused on the effects of the quality of care in early childhood settings on children's school readiness skills. Although researchers increased the statistical rigor of their approaches over the past 15 years, researchers' ability to draw causal inferences has been limited because the studies are based on nonexperimental designs. The purpose of the present study was to demonstrate how an instrumental variables approach can be used to estimate causal impacts of preschool center care quality on children's academic achievement when applied to a study in which preschool curricula were randomly assigned across multiple sites. We used data from the Preschool Curriculum Evaluation Research Initiative (PCER; $n = 2{,}700$), in which classrooms or preschools were randomly assigned to that grantee's treatment curriculum or "business as usual" conditions in 18 research sites. Using this method, we demonstrate how developmental researchers can exploit the random-assignment designs of multisite studies to investigate characteristics of programs, such as preschool center care quality, that cannot be randomly assigned and their impacts on children's development. We found that the quality of preschool care received by children has significant, albeit modest, effects on children's academic school readiness, with effect sizes of .03 to .14 standard deviation increases in academic achievement associated with a 1 standard deviation increase in quality. Applications and potential policy implications of this method are discussed.

Keywords: child care quality, instrumental variables, PCER data

Causal inferences in many developmental questions are limited because experimental variation is either infeasible or unethical. For example, a large research literature investigates the associations between early childhood care and education and children's academic outcomes (e.g., Mashburn et al., 2008; National Institute of Child Health and Human Development (NICHD) Early Child Care Research Network, 2002; Pianta, Barnett, Burchinal, & Thornburg, 2009). Child care researchers have paid increasing attention to concerns about drawing conclusions from observational studies over the past 15 years and have employed a variety of methods to reduce bias, ranging from using entry skills as covariates to employing regression discontinuity designs (Burchinal, Magnuson, Powell, & Hong, in press; Pianta et al., 2009). None of the previous studies, however, have been able to account for all observed and unobserved potential sources of bias and thus have not been able to confidently estimate causal effects. The purpose of this study was to demonstrate how an econometric technique, instrumental variables (IV), can be used with multisite experimental data to estimate causal effects for topics within child development that cannot be tested in random assignment studies (Crosby, Dowsett, Gennetian, & Huston, 2010; Gennetian, Magnuson, & Morris, 2008). We used the example of understanding the effect of center care quality on preschool children's academic achievement.

Instrumental Variables

Instrumental variables (IV) is a method that allows for estimating causal effects from observational data when certain conditions are met. This approach is predominately used in economics (e.g., Angrist & Krueger, 1991, 2001; Angrist & Pischke, 2008); however, there is increased interest in using this statistical technique to test for causal effects within developmental research (e.g., Crosby et al., 2010; Gennetian et al., 2008). The IV approach requires the identification of variables (*instruments*) that are moderately to strongly related to the predictor of interest and that serve as the

Anamarie Auger and George Farkas, School of Education, University of California, Irvine; Margaret R. Burchinal, Frank Porter Graham Child Development Institute, University of North Carolina–Chapel Hill; Greg J. Duncan and Deborah Lowe Vandell, School of Education, University of California, Irvine.

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award P01HD065704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors wish to thank Tran Dang Keys, Weilin Li, Hirokazu Yoshikawa, and members of the P01 Advisory Group for their thoughtful comments and suggestions.

Correspondence concerning this article should be addressed to Anamarie Auger, School of Education, University of California, Irvine, 3200 Education, Irvine, CA 92697-5500. E-mail: augera@uci.edu

only conduit through which the predictor has an impact on the outcome.

The difficulty of finding variables meeting both criteria has limited the method's use in developmental psychology (e.g., Gennetian et al., 2008; Loeb, Fuller, Kagan, & Carrol, 2004). However, several studies have demonstrated the effectiveness of analyzing multisite experimental data by using treatment and treatment-by-site interactions as instruments in an IV analysis (Bloom, Zhu, & Unlu, 2010; Duncan, Morris, & Rodrigues, 2011; Gennetian et al., 2008; Ludwig & Kling, 2006), and recent studies have applied this methodology to the study of the determinants of children's development (Crosby et al., 2010; Duncan et al., 2011; Gennetian et al., 2008).

These articles illustrate how, if assumptions are met, the method can eliminate several problems that limit causal inference in developmental science, including selection of individuals into treatment conditions as well as bias due to measurement error in predictor variables (Duncan, Magnuson, & Ludwig, 2004). The IV method eliminates selection and measurement error bias because it uses only the portion of the variation in the independent variable of interest (child care quality in the present case) that is caused by the instrumental variables (treatment and treatment by site interactions in the present case). In effect, it tests whether sites with the largest experimental impacts on the key independent variable (in this case, center care quality) are also the sites with the largest experimental impacts on the dependent variable (in this case, children's academic achievement).

Treatment effects on the independent variable of interest (center care quality) can be viewed as causal due to random assignment of treatment. Instrumental variables techniques use variation across sites in the treatment effects on the independent variable of interest to produce a predicted value of the predictor, where the prediction is based only on treatment-by-site variability. This variation is purged of both measurement and selection errors that may bias the estimate of the independent variable of interest. It is purged of bias that could be caused by observed or unobserved factors that could account for quality effects because random assignment determined which settings (schools or classrooms in our case) received the treatment or control condition, and thus, using variability across sites of within-site treatment effects on quality, eliminates the impact of those potential selection biases.

Instrumental variables methods reduce measurement error bias because child outcomes are a function of predicted (not actual) values of the independent variable of interest. These predicted values have been purged of idiosyncratic aspects of, in our case, center care quality. IV analyses that capitalize on random assignment of treatment in multiple sites have been applied to estimate the effect of the type of preschool care on the child's externalizing behavior (Crosby et al., 2010), the effect of maternal education on preschooler's cognitive and school outcomes (Gennetian et al., 2008), and the effect of family income on young children's academic achievement (Duncan et al., 2011).

An IV analysis using multisite experimental data typically consists of a "two-stage" regression procedure. The IV analysis begins with a regression analysis of the predictor of interest as a function of site, treatment, site by treatment, and covariates. The predicted values (of center care quality in our case) from this first-stage regression become the instrumented variable used in the second stage analysis. In the second stage, a regression analysis is performed in which the outcome variable (academic school readiness skills in our study) is analyzed as a function of this predicted variable (predicted center care quality in our case), site, and covariates. The model underlying our IV analysis is shown below:

$$\text{Predictor} \rightarrow \text{Mediator} \rightarrow \text{Outcome}$$

$$\text{Treatment} \times \text{Site} > \text{change in classroom quality} > \text{change in child outcomes.}$$

It is first important to note that the first stage predicts the mediating variable (center care quality) from the differential impact of the treatment curricula across the study sites. We expected differential impact of the treatment across sites because there were multiple curricula being tested compared with one curriculum being implemented across multiple sites. It is presumed that treatment has its impact on child outcomes through changing center care quality as defined by the quality of the child's experiences in her or his preschool care setting. The site by treatment interactions serve particularly well as instruments because treatment was randomly assigned at the site level and is therefore conceptually independent of child performance at the beginning of the study. To be successful, there must be considerable variability in treatment effects on quality across sites, which is tested by the F statistic for adding the instrumental variables (treatment and treatment by site interactions) to the first-stage prediction equation.

Employing the results from the first-stage analysis (e.g., predicted center care quality from the treatment and site by treatment interactions) as the predictor of interest in the second stage leads to an unbiased estimate of the quality variable's impact on the outcome(s) of interest (e.g., academic achievement). This is because the variation induced in the quality measure from the instruments is "pure," meaning free from selection bias and measurement error.

A potential source of bias in the IV estimates arises if other predictors also mediate the effect of the treatment and its cross-site variation on child outcomes. As in more conventional regression models, these predictors should also be included in the analysis (which in the case of IV amounts to predicting them with additional first-stage regressions and also including their predicted values in second-stage regressions). Because in our data the treatment and treatment by site interactions in multisite experimental studies provided multiple instruments, it would be possible for us to estimate multiple-mediator IV models (see Crosby et al. (2010); Duncan et al. (2011), and Gennetian et al. (2008)).

Other possible factors that influence children's school readiness in preschool include teacher qualifications, including teacher education level and specific courses taken, hours in care, student-teacher ratio, and peer effects (Burchinal et al., in press). However, for the Preschool Curriculum Evaluation Research (PCER) Study data analyzed here, the only part of the classroom that was manipulated was the instruction (i.e., implementation of new curricula), so the classroom quality measures used in the study should capture this change. There was no systematic change in teacher qualifications or in the peer composition of the classroom. Because of this, these additional influences were not instrumented and were not included in our study. Our study focused solely on the impact of center care quality on children's academic school readiness, but other mechanisms of center-based care may influence child development, such as quantity of care, and past studies, using different data sets, have examined this mechanism using an IV approach

(e.g., Li, 2013; Loeb, Bridges, Bassok, Fuller, & Rumberger, 2007).

Center Care Quality and the Need for Econometric Methods

Developmental theory suggests that the quality of child care plays an important role in early development. Ecological theories describe the impact of proximal to distal environments on children's development, positing that, across multiple domains, high-quality environments are necessary for children to reach their full developmental potential (Bronfenbrenner & Morris, 2006). In this view, children's experiences in settings (i.e., microsystems) such as the home and child care have strong effects on development because of their proximity and the large amount of time children spend in these microsystems. Important proximal processes for young children within these settings include caregiver warmth and responsiveness, stimulation, and opportunities for learning (e.g., Mashburn et al., 2008).

To date, few studies have been able to randomly assign center care quality to determine its impact on school readiness. Instead, most studies have utilized observational designs and relied on measured covariates to account for potential differences in child, family, and program characteristics that may be related to center care quality (e.g., Duncan & Gibson-Davis, 2006). Several recent meta-analyses (or studies employing meta-analytic techniques) have extended prior studies on the effects of child care quality on preschool children's academic achievement and socio-emotional development (Burchinal, Kainz, & Cai, 2011; Keys et al., 2013). In these analyses, the "value-added" approach was used by including the child's entry skills as covariates. Unlike experimental studies, these correlational designs cannot control for unmeasured confounds or for covariates that are not appropriately modeled (e.g., interactions or nonlinear effects), which can lead estimates to either over- or understate the true impact of child care quality on school readiness skills.

Increased focus on gains during a year in preschool resulted in the use of a value-added type approach, where prior achievement is taken into consideration and thereby may control for the child and family factors that contributed to the entry skills, reducing bias in the estimates (e.g., Burchinal, Kainz, & Cai, 2011; Howes et al., 2008; Keys et al., 2013; Mashburn et al., 2008; NICHD Early Child Care Research Network & Duncan, 2003). However, this approach reduces but may not eliminate bias, because it assumes that a prior achievement measure can control for all possible confounding factors, whereas this may not be the case. In the current study, we used an IV approach to eliminate bias and determine the causal effect of center care quality on children's achievement. We hypothesized that when IV is used to estimate center care quality effects, a significant positive effect on children's academic achievement will be found, with effect sizes that are larger than previously reported.

Method

Data

Preschool Curriculum Evaluation Research (PCER) Initiative Study. Beginning in 2003, 12 grantees around the United States were funded to study the effect of preschool curricula on children's academic and socio-emotional outcomes into kindergarten (Preschool Curriculum Evaluation Research Consortium, 2008). The PCER study was a multisite evaluation of preschool curricula, where multiple grantees implemented and studied different curricula and their effects on classroom quality and children's development. Each of the 12 grantees chose the curriculum to be examined, with some grantees testing the effects of more than one curriculum. A total of 14 different curricula were tested in 18 different locations. Each grantee was in charge of its own evaluation, thus curriculum effects cannot be compared with each other as noted in the PCER report (Preschool Curriculum Evaluation Research Consortium, 2008). Mathematica Policy Research (Princeton, NJ) and Research Triangle Institute (Rockville, MD) assisted with the evaluation of the curricula, so that data collection and the type of data (i.e., parent and teacher interviews/surveys, classroom observations, and child assessments) collected at each site were consistent, allowing us to pool the data across sites.

Individual grantees were responsible for recruiting preschool centers to participate in the study. These included Head Start (90 centers), private child care (40 centers), and public preschool (180 centers), with the majority of centers serving low-income children, and the majority of the preschool or child care centers being full-day centers (90%). For more information on the study, see Preschool Curriculum Evaluation Research Consortium (2008). All sample sizes are rounded to the nearest 10 in accordance with U.S. Department of Education guidelines for restricted-use data.

Participants

At each grantee site, either classrooms within preschool centers or entire centers themselves were randomly assigned to a treatment (experimental curriculum) or control condition, depending on the feasibility of being able to randomly assign classrooms within centers. A total of approximately 3,000 children in 320 preschool classrooms participated in the study. Because of feasibility issues (including the fear of cross contamination across classrooms), most research sites assigned only one curriculum to each center; although a few sites randomly assigned both treatment and control conditions within a center. Table 1 displays information about the number of preschools, classrooms, and children who participated at each site in the PCER study.

For our purposes, the samples from each grantee-level evaluation were pooled together into one data set to form our analysis sample. Pooling the data was reasonable because we were not interested in estimating the effects of individual curricula on children's achievement but instead in using the *variation* in curricula impacts across grantee sites to estimate the effect of program quality on child outcomes. Examining characteristics of the overall sample, we found families included in the PCER study were racially diverse and predominately low-income.

The children were on average 4½ years old at the time of random assignment (treatment = 54.66 months, control = 54.74 months) and were from families with a mean income of approximately $30,000 (treatment = $31,020, control = $29,310). The majority of mothers in the sample were employed (treatment = 67%, control = 64%), and approximately half were married (treatment = 48%, control = = 46%). Characteristics of children and families were evenly distributed across control and treatment conditions, with no significant differences ($p < .05$) found. Table 2

Table 1
Descriptive Information on Each Study Site Included in the Preschool Curriculum Evaluation Research Consortium Study

Site	Curriculum evaluated	Number of		
		Preschools	Classrooms	Students
Tennessee	Bright Beginnings and Creative Curriculum	20	20	310
North Carolina	Creative Curriculum	10	10	100
Georgia		10	10	90
New Hampshire	Creative Curriculum with Ladders to Literacy	10	10	120
Florida	Curiosity Corner: Success for All	10	10	50
Kansas		10	20	110
New Jersey		10	10	60
Texas	Doors to Discovery and Let's Begin With the Letter People	20	40	300
Florida (UNF)	Early Literacy and Learning Model	30	30	240
Virginia	Language-Focused Curriculum	10	10	200
Florida (FSU)	Literacy Express and DLM Early Childhood Express supplemented with Open Court Reading–Pre-K	20	30	300
California	Pre-K Mathematics supplemented with DLM Early Childhood	20	20	160
New York	Express Math Software	20	20	160
Wisconsin	Project Approach	10	10	200
Missouri	Project Construct	20	20	230
New Jersey	Ready, Set, Leap!	20	40	290

Note. Number of preschools, classrooms, and students is rounded to the nearest 10 in accordance with National Center for Education Statistics data policies. UNF = University of North Florida; FSU = Florida State University.

displays the demographic characteristics of the children and their families.

Measures

Center care quality. In this article, we focused on estimating the effects of several measures of process quality on child outcomes. The PCER study included quality measures that focused on different aspects of the center care environment. Previous studies have suggested that the quality of instruction and of teacher–child interactions are the most important components for improving academic outcomes (e.g., Mashburn et al., 2008; Phillips & Lowenstein, 2011). The interactions between caregivers and children and the overall quality of the center care environment were measured by the Early Childhood Environment Rating Scale–Revised (ECERS–R; Harms, Clifford, & Cryer, 1998). The ECERS–R is a widely used child care quality measure in research and policy (e.g., Côté et al., 2013; Gordon et al., 2013; La Paro, Thomason, Lower, Kintner-Duffy, & Cassidy, 2012), and previous work has shown that it is a much stronger predictor of academic achievement than structural or regulatable measures of child care quality such as child-to-adult ratio, and staff qualifications (Sabol, Hong, Pianta, & Burchinal, 2013). The ECERS–R consists of 43 items rated by trained observers on a 7-point scale (1 = *inadequate quality*, 7 = *excellent quality*) that focus on the quality of space and furnishings, personal care routines, language reasoning, activities, interactions, program structure, and parents and staff. Most recent studies have used two summary scores, Teacher Interactions and Provisions for Learning (e.g., Côté et al., 2012; Keys et al., 2013; Pianta et al., 2005). The Teacher Interactions factor consists of items related to the interactions that occur between children and teachers, how they communicate with each other, and the type of discipline that occurs in the classroom (α = .92). The Provisions for Learning factor focuses on the learning environment and physical features that are provided to the children in the classroom (α = .88). This measure of quality was included because several of the curricula in the PCER study included manipulatives, such as books or art materials, which may have had an influence on the learning environment.

Interactions between children and teachers were measured using the Caregiver Interaction Scale (CIS; Arnett, 1989). This is an observational tool in which trained observers rate caregivers on their interactions with children on 26 items, each being a 4-point indicator with anchors of 1 (*not at all true*) and 4 (*very much true*). The scale measures the extent to which teacher–child interactions were positive as well as the extent of harshness, detachment, and permissiveness in these interactions (Preschool Curriculum Evaluation Research Consortium, 2008). The total score is computed as the mean of the item scores, with negative items reversed (α = .94).

To examine specific teacher instructional practices, we used the Teacher Behavior Rating Scale (TBRS) as a measure of the quality and quantity of instruction in math or reading (Landry, Crawford, Gunnewig, & Swank, 2002). This measure captures both the frequency and quality of teacher behaviors surrounding literacy and math activities in the classroom. Observers rated the presence of a certain activity on a 4-point scale (0 = *activity not present*, 3 = *activity happened often or many times*). Similarly, the quality of the activity was rated on a 4-point scale, with anchors of 0 (*activity not present/conducted in classroom*) and 3 (*activity rated as high quality*). The total literacy domain, both the quality (α = .82) and quantity measure (α = .76), was composed of ratings related to four different literacy skills: describing instruction in written expression, print and letter knowledge, book reading, and oral language use. Four scores

Table 2
Preschool Curriculum Evaluation Research Consortium Study Participant Demographic and Background Characteristics, and Center Quality Measures

Variable	Treatment						Control					
	N	% of sample	Mean	*SD*	Min	Max	*N*	% of sample	Mean	*SD*	Min	Max
Child characteristics												
Gender–Female	1,540	48					1,160	49				
Race	1,410						1,080					
White		35						33				
Black		42						44				
Asian		1						1				
Hispanic		16						15				
Other		6						7				
Age (months)	1,520		54.66	3.76	45.50	71.59	1,140		54.74	3.87	45.10	66.03
Maternal characteristics baseline (fall 2003)												
Married	1,300	48					970	46				
Education level (years)	1,310		13.01	1.90	10	16	970		12.77	1.90	10	16
Employed	1,310	67					970	64				
Maternal/caregiver age (years)	1,300		31.67	7.68	16.00	68.00	960		31.55	7.72	19	74
Income (thousands)	1,160		31.02	24.47	2.50	87.50	850		29.31	23.11	2.50	87.50
Receiving welfare aid	1,300	13					960	17				
Child academic achievement baseline (fall 2003)												
WJ Letter Word	1,490		99.08	16.10	65	184	1,120		98.75	15.95	51	185
WJ Applied Problems	1,470		93.71	15.04	45	137	1,100		94.06	15.12	46	132
PPVT	1,510		88.60	15.54	40	135	1,140		88.42	16.11	40	131
Child academic achievement end of preschool year (spring 2004)												
WJ Letter Word	1,510		103.31	13.87	51	172	1,140		102.47	13.95	51	158
WJ Applied Problems	1,510		96.21	13.50	42	137	1,130		95.27	14.71	16	147
PPVT	1,530		93.25	15.00	40	134	1,160		92.26	15.45	40	132
Preschool center care quality measures (spring 2004)												
Center care quality composite	1,530		0.18	1.00	−2.02	2.67	1,150		−0.21	0.95	−2.55	2.65
ECERS−R Teacher Interactions	1,530		4.79	1.42	1.45	7.00	1,150		4.38	1.43	1.00	7.00
ECERS−R Provisions for Learning	1,530		4.00	1.08	1.73	6.18	1,150		3.72	1.08	1.27	6.82
CIS Total	1,480		3.21	0.56	1.24	3.92	1,120		3.06	0.62	1.12	3.88
TBRS–Math quantity	1,470		1.18	0.54	0.43	3	1,120		1.05	0.48	0.43	2.86
TBRS–Math quality	1,470		1.10	0.71	0	3	1,120		0.91	0.60	0	2.86
TBRS–Literacy quantity	1,470		0.14	0.80	−1.17	2.67	1,120		−0.16	0.67	−1.37	1.81
TBRS–Literacy quality	1,470		0.16	0.80	−1.33	2.17	1,120		−0.18	0.76	−1.62	1.91

Note. *N*s are rounded to the nearest 10 in accordance with National Center on Education Statistics data policies. Teacher Behavior Rating Scale (TBRS) literacy quantity and quality measures are composites of all literacy skills examined (written expression, print and letter knowledge, book reading, and oral language use). WJ = Woodcock–Johnson; PPVT = Peabody Picture Vocabulary Test; ECERS-R = Early Childhood Environment Rating Scale–Revised; CIS = Caregiver Interaction Scale.

from the TBRS were used in the analysis—math quality and quantity and total literacy quality and quantity.

To estimate the global effect of process quality on children's preschool academic achievement, we created an omnibus measure of process quality from the average of the standardized scores on all the quality measures ($\alpha = .91$). This alpha, reflecting the relatively high correlations among the quality measures (range of correlations: .40–.93; see Table 3 for a correlation matrix of the quality measures), indicates a single dimension of quality and justifies forming a quality composite. Such a composite has been used in other studies such as the Cost, Quality, and Child Outcomes Study (Peisner-Feinberg & Burchinal, 1997) and recent analyses of the Head Start Impact Study. Psychometric analyses in those studies also suggested that the various measures of observed quality were measuring a single dimension. All of the quality measures used in the present study were collected at the end of the 2003 preschool year in the spring of 2004.

Child outcomes. Three academic outcomes were examined in this study. Children's vocabulary skills were measured with the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1981). The second and third assessments were taken from the Woodcock–Johnson Psycho-Educational Battery–Revised (Woodcock, McGrew, & Mather, 2001) and measured math (Applied Problems), and reading ability (Letter–Word Recognition). These three outcomes were chosen because of their well-documented psychometric properties and their widespread use in other studies examining the association between child care quality and children's academic achievement (e.g., Burchinal et al., 2011; Burchinal, Vandergrift, Pianta, & Mashburn, 2010; Keys et al., 2013; Preschool Curriculum Evaluation Research Consortium, 2008). Each study site administered the child assessments at the beginning and end of their preschool year (fall 2003 and spring 2004). Table 2 shows the means and standard deviations for the baseline and posttreatment assessments. No significant baseline

Table 3
Correlations of Preschool Quality Measures

Measure	TBRS				ECERS–R		
	Literacy quality	Literacy quantity	Math quality	Math quantity	Teacher Interactions	Provisions for Learning	CIS total
TBRS–literacy quality	1						
TBRS–literacy quantity	.92	1					
TBRS–math quality	.61	.56	1				
TBRS–math quantity	.57	.56	.93	1			
ECERS–R–Teacher Interactions	.60	.56	.49	.44	1		
ECERS–R–Provisions for Learning	.56	.53	.43	.40	.77	1	
CIS Total	.60	.57	.48	.46	.82	.58	1

Note. Teacher Behavior Rating Scale (TBRS) literacy quantity and quality measures are composites of all literacy skills examined (written expression, print and letter knowledge, book reading, and oral language use). ECERS–R = Early Childhood Environment Rating Scale–Revised; CIS = Caregiver Interaction Scale. All correlations significant at the $p < .05$ level.

achievement differences between the treatment and control conditions were found.

Site. Site was included to account for differences across the sites at which random assignment to treatment occurred. As a categorical factor, a separate dummy variable represented each of these units. The majority of preschools judged random assignment of classrooms within the school to be infeasible. Accordingly, most sites consisted of groups of preschools managed by a particular grantee, with each preschool being randomly assigned to either the treatment or control condition. However, for the few grantees that were able to implement random assignment of classrooms *within* the preschool, the individual preschool was taken to be the study site. This leads to an analysis sample composed of a total of approximately 40 sites, of which 30 were individual preschools and 10 were grantee locations containing multiple schools.

Treatment. The PCER grantees examined the effects of one or more curricula, randomly assigning each school/classroom within their grantee to either a treatment or control condition. For the purposes of this study, we created a treatment dummy variable that estimated the difference between the treatment and control programs across sites by assigning it to have a value of 1 if the site was a treatment site regardless of which curricula was being implemented and a value of 0 for control sites.

Covariates. To account for any treatment–control imbalances remaining after random assignment, as well as to reduce the standard errors of the coefficient estimates, we controlled a number of child and maternal characteristics in the analyses. Child-level characteristics included gender (female = 1); race (White [omitted category], Black, Hispanic, Asian, and other); and age in months. Maternal characteristics obtained through parent interviews included marital status (married = 1), education level in years, whether employed (employed = 1), age in years, income in thousands, and whether receiving welfare assistance (yes = 1). For descriptive statistics of the covariates, quality measures, and child outcomes, see Table 2.

Analytic Strategy

The analysis plan involved conducting IV analyses to estimate the effect of center care quality on children's reading, math, and language skills, and, if significant, examining the individual components of the quality composite. The IV analyses consisted of two-stage least squares regressions to estimate the effect of center care quality on children's achievement (e.g., Angrist & Pischke, 2008; Duncan et al., 2011; Gennetian et al., 2008). In the first-stage regression, center care quality was analyzed as a function of site, treatment, site by treatment, the pretest score, and selected child and family characteristics. For the *i*th child in the *j*th site, the first stage equation is:

$$\text{Quality}_{ij} = a + b_1\text{Trtmt}_{ij} + b_2\text{Site}_i + b_3\text{Trtmt}_{ij} * \text{Site}_i + b_4\text{BaselineAch}_{ij} + b_5\text{Covs}_{ij} + e_{ij}.$$

As shown in the following equation, the second-stage regression then used the predicted value of quality from the first stage, site fixed effects (d_2), child baseline academic achievement (d_3), and child and family covariates (d_4) to predict the academic outcome. Treatment and the interactions of site and treatment were omitted from the second-stage regression because they were the instrumental variables used to predict a portion of the variation in the quality variable, which was then used to estimate the effect of quality on the academic outcomes. In order to obtain correct standard errors, we used a two-stage least squares estimation command in Stata, Version 11. For the *i*th child in the *j*th site, the second stage equation is:

$$\text{Achievement}_{ik} = c + d_1\,\text{Predicted Quality}_{ij} + d_2\text{Site}_i + d_3\text{BaselineAch}_{ij} + d_4\text{Covs}_{ij} + m_{ij}.$$

A strong first stage regression is necessary for the IV estimator to accomplish its goals. Our first stage regressions met this criterion, since the F statistics for the increment to R^2 when the treatment and site by treatment variables were added to the regression ranged from 13.60 to 22.11 (all of which were significant at the .001 level), which are above the recommended minimum level of 10 (e.g., Angrist & Pischke, 2008). Separate IV analyses were undertaken for each of the math, language, and vocabulary outcomes. For each of these outcomes, separate IV analyses were performed using predicted (from the first-stage regression) values of the quality composite and then each of the individual quality measures. The IV method essentially isolated the within-site variation in center care quality and child outcomes generated by random assignment. Therefore, potential problems with underestimating standard errors related to the use of multilevel data (i.e.,

multiple children in the same randomly assigned unit) were not an issue with IV analyses.

In addition, three-level hierarchical linear models (HLM; Raudenbush & Bryk, 2002) were tested for comparison purposes. This model was selected as the comparison because it is the method that developmental researchers often use to account for the nesting of children within classrooms and classrooms within study sites in studies of child care quality and child outcomes (e.g., final report; Preschool Curriculum Evaluation Research Consortium, 2008).

The same control variables, including children's baseline academic achievement, plus site fixed effects, were included in the HLM analysis. In the HLM analysis, the observed center care quality was used, whereas in the IV analysis, the predicted value of quality was used. Similar to the IV models, hierarchical linear models were tested separately, predicting each academic outcome from the different quality measures and the center care quality composite. All quality measures, the quality composite, and academic outcomes were standardized to have a mean of 0 and standard deviation of 1.

Missing data. Only children with at least one academic outcome were retained in the analyses. This requirement dropped 230 children who were missing all outcomes, which resulted in 10 schools not being included in the final analyses. A total of 2,700 children of the full sample (N = 2,911) were included in the final sample. The children without outcomes were more disadvantaged than the retained sample; their mothers were younger, and their parents reported working significantly less, having lower income, and being on welfare at higher rates. Children without outcome data were also more likely than those with outcomes to be female and to be in the "other race category" that included Native Americans, Alaskans, and other races.

Missing data were handled by setting missing values for the covariates and baseline achievement measures to the mean of the variable and adding a dummy variable into the prediction equations for each covariate and baseline achievement indicating if the variable was missing (1 = missing, 0 = not missing). This approach has been noted as being effective for handling missing data in randomized control trials (Puma, Olsen, Bell, & Price, 2009). We also chose this approach because methods that incorporate data imputation techniques into instrumental variables estimation models are not yet available, thus, we were hesitant to combine these two techniques in an analysis.

Results

There are two ways to demonstrate whether the first stage was successful. We first turn to the site by treatment interaction coefficients from the first stage analysis presented in Table 4. The second and fourth columns of Table 4 show the estimated coefficients for each site by treatment interaction in the first-stage regression predicting the omnibus quality composite variable. Again, these interactions serve as the instruments producing the variation that was used to estimate center care quality effects. For this analysis to be successful, the first-stage results displayed in Table 4 must show a range of impacts, from negative to positive, of the different site by treatment interactions on the quality measures. This differential impact is important for a successful first stage in our IV analysis—without variation across the sites, this IV analysis would not have a sufficiently strong instrument to proceed to the next stage (Reardon, Raudenbush, & Bloom, 2013). Examining these coefficients, we see that 13 coefficients are statistically significant and positive, indicating that the treatment condition in each of these sites experienced significantly higher quality center care than the control condition after implementation of the treatment curriculum. Three sites also have significantly negative coefficients, indicating that the treatment curriculum resulted in lower process quality than the "as is" curriculum. Results for the separate quality measures are similar.

We also see that the F statistic for the improvement in R^2 when the interaction terms were added to the equation shown in the table for the quality composite is well above the recommended level of 10 and is so for the other quality measures also. The large F statistic in the first stage indicates that the instruments—in this case, the site dummy variables interacted with the treatment variable—were sufficiently strong predictors of the quality measure to provide adequate power to estimate center care quality effects. That is, significant variability across sites in treatment effects on quality implies that predicting quality under this model would be capturing sufficient variability in quality to have a meaningful predictor to use in our second-stage analyses.

To graphically depict the variation between each site's average control and treatment quality scores, we plotted them against the average PPVT score for each site. Overall, there was reasonably large variation across study sites in the effectiveness of the treatment curricula in raising center care quality. Figure 1 shows the site average of the quality composite plotted against average PPVT scores. This figure shows the variation in treatment and control conditions for each of the random assignment sites. The important message from this graph is that there is a reasonable amount of variation in the treatment impacts across the random assignment sites. On the graph for every black indicator (treatment), there is a black-outlined hollow indicator (control) in approximately the same location in the opposite quadrant. The numbers next to the indicators refer to the random assignment site location. Black indicators are clustered in the upper right quadrant, indicating a positive effect of the center care quality composite on children's vocabulary (as measured by the PPVT) achievement in the spring of their preschool year. The slope of the indicators is nearly identical to the IV analysis coefficient estimate for the effect of the quality composite on children's PPVT scores with the exception being the fitted line displayed in the figure does not take into consideration covariates.

IV regression analyses were performed to estimate the effects of both the quality composite and the individual quality measures. HLM analyses were performed for comparison to the IV estimates. It was expected that the p value for the HLM estimates would be smaller than the IV estimates, as the two-stage approach almost always increases the standard errors. Table 5 presents the results from the HLM and IV analyses for these measures. All quality measures were entered separately into the regressions. Each estimate presented in the table was from one individual regression. The quality measures and child achievement assessments were standardized, so the magnitudes of their estimated coefficients can be understood as effect sizes.

Table 4
First Stage Instrumental Variables Results–Instruments (Treatment and Treatment by Site) Predicting the Preschool Center Care Quality Composite

Site	Site interacted with treatment	Site	Site interacted with treatment
NC		(17) CA	0.604 (0.260)*
(1) School 1	(omitted)	(18) School 1	0.085 (0.327)
(2) School 2	−0.079 (0.374)	(19) NY	0.357 (0.256)
(3) School 3	0.377 (0.379)	(20) FL–1	0.154 (0.286)
GA		(21) FL–2	0.484 (0.275)
(4) School 1	2.426 (0.362)***	(22) FL–3	0.748 (0.287)**
(5) School 2	3.043 (0.300)***	(23) NH	0.226 (0.290)
NJ		(24) School 1	1.450 (0.443)**
(6) School 4	−2.388 (0.397)***	(25) TN	1.086 (0.244)***
(7) School 7	0.161 (0.434)	(26) TX	0.790 (0.246)**
(8) School 8	−0.018 (0.276)	(27) WI	0.241 (0.250)
(9) School 10	0.315 (0.434)	(28) MO	0.390 (0.251)
(10) School 11	0.217 (0.514)	(29) FL–FSU	0.331 (0.245)
(11) School 12	0.261 (0.358)	(30) KS–SFA	0.587 (0.269)*
(12) School 14	0.210 (0.488)	(31) FL–SFA	0.892 (0.308)**
		(32) NJ–SFA	−0.876 (0.304)**
VA			
(13) School 1	0.044 (0.334)	Treatment (Intervention Curriculum Classroom)	0.156 (0.228)
(14) School 2	−0.782 (0.343)*	Model R^2	.50
(15) School 3	0.599 (0.342)	Model F	28.32***
(16) School 4	1.028 (0.279)***		

Note. N = 2,670. Preschool quality composite is the dependent variable: F Treatment × Site (instruments) = 20.479***. Standard errors are in parentheses. Number of observations is rounded to the nearest 10 in accordance with National Center for Education Statistics data policies. Omitted sites from the table were dropped due to missing data. All models presented in the table include the following child-level covariates: age in months, gender, race (Black, Asian, Hispanic, or other; White is the comparison group), and baseline achievement as measured in the fall 2003. Parent- (mother-)level covariates include (for fall 2003): age in years, education level, whether married, whether working (full or part time), whether receiving welfare aid, and annual income in thousands. Level of random assignment depends on whether the school allowed both treatment and control condition to be present. Most sites did not, so level of random assignment is at the study site location level. Missing data were handled by setting missing cases in variables to the mean except for dichotomous variables. An additional variable was entered into the model for whether the variable was missing. The quality measure is a composite of the following quality measures: Early Childhood Environment Rating Scale–Revised (ECERS-R; Provisions for Learning and Teacher Interactions), Caregiver Interaction Scale (CIS), Teacher Behavior Rating Scale (TBRS)–math quality, TBRS–literacy quality (composite of all literacy activities), TBRS–math quantity, and TBRS–literacy quantity (composite of all literacy activities). TBRS literacy quantity and quality measure are composites of all literacy skills examined (written expression, print and letter knowledge, book reading, and oral language use). FSU = Florida State University; SFA = Success for All.
* $p < .05$. ** $p < .01$. *** $p < .001$.

We begin with the estimated effects of the quality composite presented in Table 5. In the IV analyses, higher quality was significantly related to higher language scores (d = 0.07), and marginally significantly related to math scores (d = 0.08), but not to reading scores (d = 0.04). Findings from HLM analyses also yielded significant associations between composite quality and children's academic school readiness skills: language, d = 0.05, math, d = 0.04, and reading, d = 0.10. The Appendix displays the coefficients for the quality composite and all covariates in the first and second stage analyses in the IV regressions and the HLM analyses.

Next, we examined the relations between the individual quality measures and the academic skills to determine whether specific quality measures have reliable associations with these outcomes. As shown in Table 5, only the ECERS–R scales were significantly related to these outcomes in the IV analyses. More frequent and stimulating teacher–child interactions as measured by the ECERS–R Teacher Interactions factor predicted higher math skills (d = 0.10), whereas greater access to developmentally appropriate activities according to the ECERS–R Provisions for Learning factor predicted higher language (d = 0.14) and math (d = 0.11) scores. Neither the other measure of teacher–child interactions, the CIS, nor the measures of the quality and quantity of literacy or math instruction, the TBRS, were significantly related to children's academic outcomes in these IV analyses. In contrast to the IV analyses, the HLM analyses of language and math skills as a function of the quality composite or individual quality measures tended to yield smaller coefficients, indicating HLM results may be downward biased, which would occur if there is measurement error in the predictors.

Discussion

Although several developmental articles have advocated for use of instrumental variables within the field of child development (Crosby et al., 2010; Foster, 2010; Gennetian et al., 2008), the use

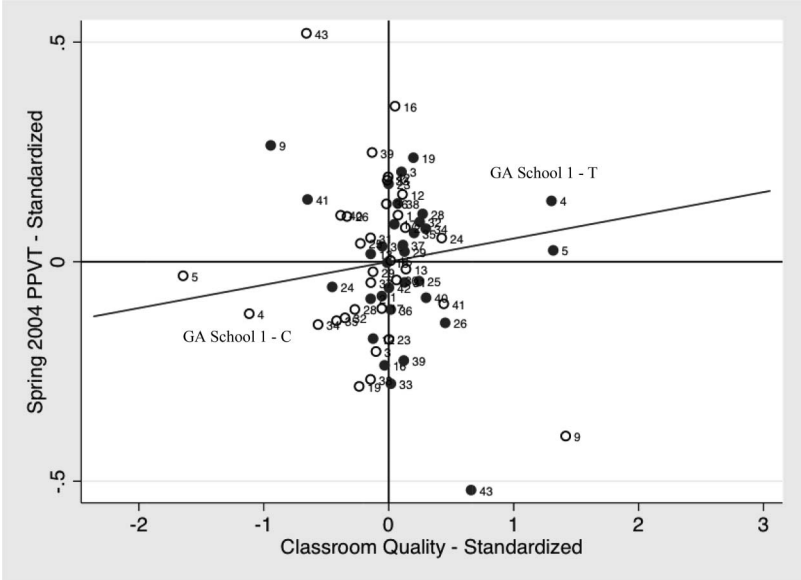

Figure 1. Center care quality composite by children's spring 2004 Peabody Picture Vocabulary Test (PPVT) score. Black indicators are treatment condition, and hollow indicators are control condition. Each random assignment site retained in the analysis sample is represented on the graph, with more treatment conditions being clustered in the upper right quadrant, indicating a positive relationship between children's vocabulary achievement and the center care quality composite. An example of this is Georgia (GA) School No. 1 (Site No. 4). The treatment condition is in the upper right quadrant, and the control condition is located in the lower left quadrant. The slope is indicated by the diagonal line—this slope is nearly identical to the regression coefficient from the instrumental variable (IV) analysis, with the exception being the fitted line displayed in the figure does not take into consideration covariates.

of this econometric method in developmental science is limited. The aim of the present study was to demonstrate the use of the instrumental variables technique to investigate the causal association between variations in preschool center care process quality and preschool children's academic school readiness. Whereas most child care research and evaluations have been concerned about bias due to potential confounds related to which children receive higher or lower quality child care, this may be the first study that directly estimates the causal effect of child care quality using statistically rigorous methods. Using an IV approach, our findings indicate that process quality—both the composite measure and the global and domain specific measures—positively affects preschool children's language skills, with a small effect size of approximately .08 and has a similar small, albeit not significant, effect on math and reading skills. This effect size is similar to previous work examining the immediate association between center care quality and children's school readiness (e.g., Burchinal et al., 2011; Keys et al., 2013; Mashburn et al., 2008; NICHD Early Child Care Research Network, 2006).

Burchinal et al. (2011) conducted a meta-analysis of the relations between child care quality and academic outcomes across multiple large-scale data sets and found an overall association between child care quality and academic and social outcomes that was modest (correlation range: −.09 to .15). Similarly, Keys et al. (2013) using four national data sets and a meta-analytic technique found small effects of center care quality on preschool children's academic achievement, with effect sizes ranging from .03 to .05. An important future research direction would be to consider these estimates within a benefit–cost framework. The cost of moving programs from medium to high quality has not been accurately estimated, nor have there been studies testing whether the benefits of doing so outweigh the costs. Despite numerous calls to conduct cost–benefit analyses, little evidence exists regarding the cost–benefit of large-scale center-based care, such as Head Start (Duncan & Magnuson, 2007; Haskins, 1989), or public preschool centers such as the Oklahoma or Boston pre-K programs (Gormley, Gayer, Phillips, & Dawson, 2005; Weiland & Yoshikawa, 2013).

Application of Instrumental Variables in Developmental Research

One of the reasons IV is so useful to developmental science is that few other tools are available apart from random assignment studies that allow researchers to draw causal claims. Random assignment is clearly the "gold standard," but in some cases it is infeasible. IV may be of particular importance for estimating causal effects of features of care settings, such as center-based care, that have policy implications but cannot be estimated through other experimental or quasi-experimental study designs. In this study, we were able to take advantage of the multisite evaluation data from the PCER study. The multisite data allowed us to use treatment and the treatment by site interactions in the first stage to estimate center care quality, which worked well because of the differential impacts of curricula on center care quality. Other researchers have noted the effectiveness of utilizing multisite ex-

Table 5
Preschool Center Care Quality Predicting Children's Academic Outcomes Using Hierarchical Linear Models and Two-Stage Least Squares Regression (Instrumental Variables)

Variable	Peabody Picture Vocabulary Test		Woodcock–Johnson Psycho-Educational Battery–Revised			
			Applied Problems		Letter–Word Recognition	
	HLM	IV	HLM	IV	HLM	IV
Child care quality composite	.05** (.02)	.07* (.03)	.04* (.02)	.08† (.04)	.10*** (.02)	.04 (.04)
Individual quality measures						
ECERS–R–Teacher Interactions	.02 (.02)	.07 (.04)	.01 (.02)	.10* (.05)	.04 (.02)	.06 (.04)
ECERS–R–Provisions for Learning	.03 (.02)	.14** (.04)	−.01 (.02)	.11* (.05)	.02 (.02)	.04 (.05)
CIS	.03 (.02)	.06 (.04)	.03 (.02)	.07 (.04)	.04 (.02)	.05 (.04)
TBRS literacy quality	.06*** (.02)	.04 (.03)	.05* (.02)	.07† (.04)	.13*** (.02)	.05 (.04)
TBRS literacy quantity	.05** (.02)	.05 (.03)	.04* (.02)	.05 (.04)	.11*** (.02)	.05 (.04)
TBRS math concepts quality	.04** (.02)	.06 (.04)	.05** (.02)	.08 (.04)	.11*** (.02)	.03 (.04)
TBRS math concepts quantity	.04** (.02)	.06 (.03)	.06** (.02)	.07 (.04)	.10*** (.02)	.02 (.04)

Note. Standard errors are in parentheses. Each estimate comes from a separate regression. Number of observations range from 2,580 to 2,670. *F* statistics for the instrumental variables (IV) models range from 13.60 to 22.11. All models presented in the table include the following child-level covariates: age in months, gender, race (Black, Asian, Hispanic, or other; White is the comparison group), and baseline achievement as measured in the fall 2003. Parent-(mother-)level covariates include (for fall 2003): age in years, education level, whether married, whether working (full or part time), whether receiving welfare aid, and annual income in thousands. Level of random assignment depends on whether the school allowed both treatment and control condition to be present. Most sites did not, so level of random assignment is at the study site location level. Independent and dependent variables were standardized to have a mean of 0 and a standard deviation of 1. In the IV models, quality was instrumented by treatment and random assignment level interacted with treatment condition. Missing data were handled by setting missing cases in variables to the mean except for dichotomous variables. An additional variable was entered into the model for whether the variable was missing. The quality measure is a composite of the following quality measures: Early Childhood Environment Rating Scale–Revised (ECERS-R; Provisions for Learning and Teacher Interactions), Caregiver Interaction Scale (CIS), Teacher Behavior Rating Scale (TBRS)–math quality, TBRS–literacy quality (composite of all literacy activities), TBRS–math quantity, and TBRS–literacy quantity (composite of all literacy activities).
† $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

perimental data (e.g., Duncan et al., 2011; Gennetian et al., 2008) in an IV analysis because of how well these interactions work as instruments.

Recently, there has been an increase of multisite experiments in the child development field, and several data sets from these experiments are prime candidates for using the IV method to estimate causal effects. Specifically the Head Start Impact Study and National Early Head Start Evaluation data sets are excellent examples of existing data sets where this method could be applied. Researchers have already begun to take advantage of these data sets; for example, Li (2013) used the Head Start Impact Study to estimate the effect of quantity of center care on children's school readiness skills.

An important note about the IV method is that unless multiple instruments are used, one must assume that the effect of the instrument on the outcome of interest works *only* through the mediator. As Gennetian et al. (2008) and others have discussed, this is a critical assumption of the method as many developmental psychologists examine multiple paths of influence using methods such as structural equation modeling. Also, this assumption of the method is untestable, so adequate justification is needed as to why the instrument(s) work only through the mediator of interest.

Although IV analyses account for differential selection bias, the method is unable to solve issues related to the validity of the observational or survey measures in terms of their ability to measure important aspects of a care setting or parenting, for example. This is something that should be noted when applying this method to research questions. IV does not address problems related to the conceptualization of what should be measured, nor the psychometrics of a measure. If a measure is not valid, IV will not be able to adjust for this type of measurement error.

The present study builds off prior studies demonstrating how the use of instrumental variables can aid the developmental field in estimating causal effects (e.g., Crosby et al., 2010; Foster & McLanahan, 1996; Gennetian et al., 2008). Even with these previous articles describing the need for the application of this method to developmental research questions, the method has been infrequently employed in this research area. The goal of our study was to demonstrate how the method could be used to examine how center care quality causally impacts children's academic school readiness, a question of much interest to the field and one with important policy implications. This research question was used as an example to demonstrate how the method can be applied using data from multisite experiments to answer developmental research questions that have policy implications but cannot be addressed with random assignment research designs. With the influx of more multisite experiments, it is likely, and we are hopeful, that more researchers will employ this method to estimate causal effects.

References

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics, 106,* 979–1014. doi:10.2307/2937954

Angrist, J. D., & Krueger, A. B. (2001). *Instrumental variables and the search for identification: from supply and demand to natural experiments* (National Bureau of Economic Research Working Paper No. 8456). Retrieved from http://www.nber.org/papers/w8456

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology, 10,* 541–552. doi:10.1016/0193-3973(89)90026-9

Bloom, H., Zhu, P., & Unlu, F. (2010). *Finite sample bias from instrumental variables analysis in randomized trials* (MDRC Working Paper). New York, NY: MDRC.

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner (Ed.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (6th ed., pp. 793–828). Hoboken, NJ: Wiley.

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). Baltimore, MD: Brookes.

Burchinal, M., Magnuson, K., Powell, D., & Hong, S. S. (in press). Early child care and education and child development. In R. Lerner, M. Bornstein, & T. Leventhal (Eds.), *Handbook of child psychology and developmental science: Vol. 4. Ecological settings and processes* (7th ed.). New York, NY: Wiley.

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. doi:10.1016/j.ecresq.2009.10.004

Côté, S. M., Mongeau, C., Japel, C., Xu, Q., Séguin, J. R., & Tremblay, R. E. (2013). Child care quality and cognitive development: Trajectories leading to better preacademic skills. *Child Development, 84,* 752–766. doi:10.1111/cdev.12007

Crosby, D. A., Dowsett, C. J., Gennetian, L. A., & Huston, A. C. (2010). A tale of two methods: Comparing regression and instrumental variables estimates of the effects of preschool child care type on the subsequent externalizing behavior of children in low-income families. *Developmental Psychology, 46,* 1030–1048. doi:10.1037/a0020384

Duncan, G. J., & Gibson-Davis, C. M. (2006). Connecting child care quality to child outcomes: Drawing policy lessons from nonexperimental data. *Evaluation Review, 30,* 611–630. doi:10.1177/0193841X06291530

Duncan, G. J., & Magnuson, K. (2007). Penny wise and effect size foolish. *Child Development Perspectives, 1,* 46–51. doi:10.1111/j.1750-8606.2007.00009.x

Duncan, G. J., Magnuson, K., & Ludwig, J. (2004). The endogeneity problem in developmental studies. *Research in Human Development, 1,* 59–80. doi:10.1080/15427609.2004.9683330

Duncan, G. J., Morris, P. A., & Rodrigues, C. (2011). Does money really matter? Estimating impacts of family income on young children's achievement with data from random-assignment experiments. *Developmental Psychology, 47,* 1263–1279. doi:10.1037/a0023875

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test* (Rev. ed.). Circle Pines, MN: American Guidance Service.

Foster, E. M. (2010). Causal inference and developmental psychology. *Developmental Psychology, 46,* 1454–1480. doi:10.1037/a0020204

Foster, E. M., & McLanahan, S. (1996). An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods, 1,* 249–260. doi:10.1037/1082-989X.1.3.249

Gennetian, L. A., Magnuson, K., & Morris, P. A. (2008). From statistical associations to causation: What developmentalists can learn from instrumental variables techniques coupled with experimental data. *Developmental Psychology, 44,* 381–394. doi:10.1037/0012-1649.44.2.381

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS–R with implications for measures of child care quality and relations to child development. *Developmental Psychology, 49,* 146–160. doi:10.1037/a0027899

Gormley, W. T., Jr., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology, 41,* 872–884. doi:10.1037/0012-1649.41.6.872

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early Childhood Environment Rating Scale–Revised Edition*. New York, NY: Teachers College Press, Columbia University.

Haskins, R. (1989). Beyond metaphor: The efficacy of early childhood education. *American Psychologist, 44,* 274–282. doi:10.1037/0003-066X.44.2.274

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. doi:10.1016/j.ecresq.2007.05.002

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

Landry, S. H., Crawford, A., Gunnewig, S. B., & Swank, P. R. (2002). *Teacher Behavior Rating Scale*. Unpublished instrument, University of Texas Health Science Center at Houston.

La Paro, K. M., Thomason, A. C., Lower, J. K., Kintner-Duffy, V. L., & Cassidy, D. J. (2012). Examining the definition and measurement of quality in early childhood education: A review of studies using the ECERS–R from 2003 to 2010. *Early Childhood Research & Practice, 14*(1). Retrieved from http://ecrp.uiuc.edu/v14n1/laparo.html

Li, W. (2013, March). *Effects of Head Start hours on children's cognitive, pre-academic, and behavioral outcomes: An instrumental variable analysis.* Paper presented at spring 2013 conference of the Society for Research on Educational Effectiveness, Washington, DC.

Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. W. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. *Economics of Education Review, 26,* 52–66. doi:10.1016/j.econedurev.2005.11.005

Loeb, S., Fuller, B., Kagan, S. L., & Carrol, B. (2004). Child care in poor communities: Early learning effects of type, quality, and stability. *Child Development, 75,* 47–65.

Ludwig, J., & Kling, J. R. (2006). *Is crime contagious?* (National Bureau of Economic Research Working Paper No. 12409). Retrieved from http://www.nber.org/papers/w12409

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., 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. doi:10.1111/j.1467-8624.2008.01154.x

National Institute of Child Health and Human Development Early Child Care Research Network. (2002). Child-care structure process outcome: Direct and indirect effects of child-care quality on young children's development. *Psychological Science, 13,* 199–206. doi:10.1111/1467-9280.00438

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,* 1454–1475. doi:10.1111/1467-8624.00617

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,* 99–116. doi:10.1037/0003-066X.61.2.99

Peisner-Feinberg, E. S., & Burchinal, M. R. (1997). Relations between preschool children's child-care experiences and concurrent development: The Cost, Quality, and Outcomes Study. *Merrill-Palmer Quarterly, 43,* 451–477. doi:10.2307/23093333

Phillips, D. A., & Lowenstein, A. E. (2011). Early care, education, and child development. *Annual Review of Psychology, 62,* 483–500. doi:10.1146/annurev.psych.031809.130707

Pianta, R. C., Barnett, W. S., Burchinal, M., & Thornburg, K. R. (2009). The effects of preschool education: What we know, how public policy is

or is not aligned with the evidence base, and what we need to know. *Psychological Science in the Public Interest, 10,* 49–88. doi:10.1177/1529100610381908

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,* 144–159. doi:10.1207/s1532480xads0903_2

Preschool Curriculum Evaluation Research Consortium (PCER). (2008). *Effects of preschool curriculum programs on school readiness (NCER 2008–2009)*. Washington, DC: Institute of Education Sciences, National Center for Education Research.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). Washington, DC: Institute of Educational Sciences, National Center for Educational Evaluation and Regional Assistance. doi:10.1037/e600782011-001

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed., *Advanced Quantitative Techniques in the Social Sciences Series* Vol. *1*). Thousand Oaks, CA: Sage.

Reardon, S. F., Raudenbush, S. W., & Bloom, H. (2013, October). *Multiple-site, multiple-mediator instrumental variables (MSMM–IV) methods: Assumptions and design issues*. Paper presented at a William T. Grant meeting, Chicago, IL.

Sabol, T. J., Hong, S. L., Pianta, R. C., & Burchinal, M. R. (2013, August). Can rating pre-K programs predict children's learning? *Science, 341,* 845–846. doi:10.1126/science.1233517

Stata. (Version 11) [Computer software]. College Station, TX: StataCorp.

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. doi:10.1111/cdev.12099

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III*. Itasca, IL: Riverside.

Appendix

Full Results for the Instrumental Variables and Hierarchical Linear Modeling Analyses

	Instrumental variables analyses						Hierarchical linear modeling analyses		
			Woodcock–Johnson Psycho-Educational Battery–Revised					Woodcock–Johnson Psycho-Educational Battery–Revised (N = 2,630)	
	Peabody Picture Vocabulary Test		Applied Problems		Letter–Word Recognition		Peabody Picture Vocabulary Test (N = 2,670)		
Variable	1st Stage (N = 2,680)	2nd Stage (N = 2,670)	1st Stage (N = 2,680)	2nd Stage (N = 2,630)	1st Stage (N = 2,680)	2nd Stage (N = 2,630)		Applied Problems	Letter–Word Recognition
Quality composite		0.07* (0.03)		0.08 (0.04)		0.04 (0.04)	0.05** (0.02)	0.04* (0.02)	0.10*** (0.02)
Child–Female	−0.03 (0.03)	0.04 (0.02)	−0.03 (0.03)	0.09** (0.03)	−0.03 (0.03)	0.09*** (0.03)	0.03 (0.02)	0.09** (0.03)	0.10*** (0.03)
Race–Black	−0.13** (0.04)	−0.11** (0.04)	−0.13** (0.04)	−0.06 (0.04)	−0.14** (0.04)	0.00 (0.04)	−0.12** (0.04)	−0.09 (0.04)	0.02 (0.04)
Race–Asian	0.14 (0.14)	0.01 (0.12)	0.12 (0.14)	0.35** (0.14)	0.11 (0.14)	0.33* (0.13)	−0.00 (0.12)	0.28* (0.14)	0.29* (0.14)
Race–Hispanic	−0.04 (0.05)	−0.14** (0.05)	−0.05 (0.05)	0.03 (0.05)	−0.06 (0.05)	−0.01 (0.05)	−0.17*** (0.04)	−0.04 (0.05)	−0.04 (0.05)
Race–other	−0.12 (0.07)	−0.07 (0.06)	−0.12 (0.07)	0.16* (0.07)	−0.12 (0.07)	0.04 (0.06)	−0.07 (0.06)	0.13 (0.07)	0.05 (0.06)
Parent education	0.02** (0.01)	0.03*** (0.01)	0.03** (0.01)	0.05*** (0.01)	0.03** (0.01)	0.04*** (0.01)	0.04*** (0.01)	0.06*** (0.01)	0.04*** (0.01)
Whether working	−0.05 (0.03)	−0.03 (0.03)	−0.05 (0.03)	0.01 (0.03)	−0.05 (0.03)	−0.03 (0.03)	−0.03 (0.03)	0.02 (0.03)	−0.03 (0.03)
Mother age	−0.00 (0.00)	−0.00 (0.00)	−0.00 (0.00)	0.00 (0.00)	−0.00 (0.00)	0.00 (0.00)	−0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Child age	0.00 (0.00)	−0.01 (0.00)	0.00 (0.00)	−0.02*** (0.00)	0.00 (0.00)	−0.04*** (0.00)	−0.01* (0.00)	−0.02*** (0.00)	−0.04*** (0.00)
Family income	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Welfare assistance	−0.03 (0.05)	−0.02 (0.04)	−0.03 (0.05)	0.08 (0.05)	−0.03 (0.05)	−0.05 (0.05)	−0.04 (0.04)	0.08 (0.05)	−0.06 (0.05)
Married	−0.05 (0.04)	−0.02 (0.03)	−0.06 (0.04)	0.09* (0.04)	−0.06 (0.04)	−0.00 (0.04)	−0.01 (0.03)	0.09* (0.04)	0.01 (0.03)
Baseline achievement	0.02 (0.02)	0.71*** (0.01)	0.02 (0.02)	0.60*** (0.02)	−0.00 (0.02)	0.60*** (0.02)	0.71*** (0.01)	0.59*** (0.02)	0.61*** (0.02)
Constant	−0.42 (0.30)	−0.36 (0.23)	−0.47 (0.30)	−0.19 (0.27)	−0.47 (0.30)	1.08*** (0.26)	0.09 (0.21)	0.12 (0.25)	1.53*** (0.25)

Note. Standard errors are in parentheses. Number of observations is rounded to the nearest 10 in accordance with National Center for Education Statistics data policies. Level of random assignment depends on whether the school allowed both treatment and control condition to be present. Most sites did not, so level of random assignment is at the study-site-location level. Schools and sites that were omitted did not contain enough data for the models to test. Independent and dependent variables were standardized to have a mean of 0 and a standard deviation of 1. In the instrumental variable (IV) model, quality was instrumented by treatment, and random assignment level interacted with treatment condition. Two-level (classrooms nested with evaluation sites): Hierarchical linear modeling (HLM) was conducted to compare the results from the IV models with HLM results. Missing data were handled by setting missing cases in variables to the mean except for categorical variables. An additional variable was entered into the model for whether the variable was missing. The quality measure is a composite of the following quality measures: Early Childhood Environment Rating Scale–Revised (ECERS-R; Provisions for Learning and Teacher Interactions), Caregiver Interaction Scale (CIS), Teacher Behavior Rating Scale (TBRS)–math quality, TBRS–literacy quality (composite of all literacy activities), TBRS–math quantity, and TBRS–literacy quantity (composite of all literacy activities).
* $p < .05$. ** $p < .01$. *** $p < .001$.

Received January 29, 2013
Revision received July 7, 2014
Accepted July 30, 2014