CHILD DEVELOPMENT

Child Development, March/April 2018, Volume 89, Number 2, Pages 539–555

What Is the Long-Run Impact of Learning Mathematics During Preschool?

Tyler W. Watts and Greg J. Duncan
University of California, Irvine

Douglas H. Clements and Julie Sarama
University of Denver

The current study estimated the causal links between preschool mathematics learning and late elementary school mathematics achievement using variation in treatment assignment to an early mathematics intervention as an instrument for preschool mathematics change. Estimates indicate (n = 410) that a standard deviation of intervention-produced change at age 4 is associated with a 0.24-*SD* gain in achievement in late elementary school. This impact is approximately half the size of the association produced by correlational models relating later achievement to preschool math change, and is approximately 35% smaller than the effect reported by highly controlled ordinary least squares (OLS) regression models (Claessens et al., 2009; Watts et al., 2014) using national data sets. Implications for developmental theory and practice are discussed.

An accumulating body of research suggests that early mathematical skills are critical to developing long-run success in school (Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Byrnes & Wasik, 2009; Claessens & Engel, 2013; Duncan et al., 2007; Geary, Hoard, Nugent, & Bailey, 2013; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Stevenson & Newman, 1986; Watts, Duncan, Siegler, & Davis-Kean, 2014). Among these studies, Duncan et al.'s (2007) analysis of six longitudinal data sets provides the most robust evidence of strong associations between early and later mathematics achievement. Their investigation of school readiness skills asked a seemingly straight-forward question: If one examined a broad range of child skills and behaviors at school entry, and controlled for a host of child and family background characteristics, which characteristics would emerge as the strongest predictors of the child's eventual school achievement? Among the candidates investigated were academic competencies, attention problems, and internal and externalizing problem behaviors. Across the data sets, a consistent pattern emerged: Mathematics achievement at school entry was the strongest predictor of later success in mathematics, and in some cases reading, even when all other characteristics tested were controlled. Since the publication of this study, other correlational studies have found similar results (Claessens, Duncan, & Engel, 2009; Claessens & Engel, 2013; Foster, 2010), including one that extended the outcome measurement into high school (Watts et al., 2014).

Developmental and cognitive theories predict that early mathematics knowledge is associated with later achievement because early numerical skills facilitate students' future mathematical skill acquisition (e.g., Aunola et al., 2004; Entwisle & Alexander, 1990; Gersten et al., 2009; Jordan et al., 2009). This skill-building framework rests on the idea that mathematics is a particularly hierarchical subject, in which mastery of simple concepts and procedures is required for understanding more difficult mathematics. For example, solving even a simple algebraic equation would be impossible without knowledge of operations such as division and multiplication, and this operational knowledge depends on understanding the basic principles of counting. Relatedly, Siegler, Thompson, and Schneider (2011) describe how students gradually broaden the class of numbers that they understand as they

This research was supported by the Institute of Education Sciences, U.S. Department of Education through Grants R305K05157 and R305A120813 and the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under Award Number P01HD065704. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education nor the officials views of the National Institutes of Health. The authors wish to express appreciation to the school districts, teachers, and students who participated in this research. We would like to acknowledge the helpful contributions of Ana Auger, Drew Bailey, Howard Bloom, Marianne Bitler, Peg Burchinal, Damon Clark, Pamela Davis-Kean, Dale Farran, George Farkas, Jade Jenkins, Tutrang Nguyen, Katerina Schenke, Mary Elaine Spitler, Jeff Smith, and Chris Wolfe.

Correspondence concerning this article should be addressed to Tyler W. Watts, School of Education, University of California, Irvine, 3200 Education, Irvine, CA 92697-5500. Electronic mail may be sent to twatts@uci.edu.

 0009-3920/2018/8902-0015
DOI: 10.1111/cdev.12713

progress through mathematics, with successful students moving from mastery of whole numbers in early grades to fractions in later elementary and middle school. Indeed, a well-developed body of empirical work documents the carefully sequenced cognitive steps students take as they expand their understanding of numbers and mathematics (e.g., Booth & Siegler, 2006; Gilmore, McCarthy, & Spelke, 2007; Laski & Siegler, 2007; Opfer & Thompson, 2008; Sarama & Clements, 2009).

Beyond the cognitive skill-building framework lie other developmental reasons to expect that early success in mathematics would set children on a successful trajectory throughout school. Complex interactions between the child and her environment in the early schooling years are likely to leave long-lasting influences on the child's developmental trajectory (Bronfenbrenner & Morris, 2006). For example, high-achieving children in kindergarten are more likely to receive positive feedback regarding their academic proficiency from teachers, parents, and peers, which in turn may boost their perception of their own math competence (Bong & Skaalvik, 2003; Meisels, 1998). Relatedly, early mathematics achievement could be a gateway to higher ability tracking in school, which would also support further academic development. Indeed, these pathways from early to later mathematics achievement have received empirical support, as evidence suggests that self-concepts and placement into gifted and talented programs both mediate the association between early and later mathematics (Watts et al., 2015).

From Level to Change in Early Mathematics

Much of the correlational evidence linking early and later mathematics ability is based on measures of early levels of math skills. Other studies show strong associations between early *gains* in mathematical ability and later success in school. For example, using longitudinal data, Watts et al. (2014) found that gains in mathematical skills during the first 2 years of school were more predictive of later achievement than were level measures of school-entry skills. Moreover, early math gains were just as predictive of high school achievement as Grade 3 math achievement, even after controlling for concurrent gains in other cognitive skills, such as working memory and reading achievement. Using nationally representative data, Claessens et al. (2009) found that change in mathematics achievement across kindergarten was highly predictive of both fifth-grade mathematics and reading achievement. Finally, using a growth curve modeling approach, Jordan et al. (2009) found that change in number competence, measured six times in kindergarten and first grade, strongly predicted third-grade mathematics achievement.

Taken together, these studies suggest that the process of *learning* mathematics during the early-grade years may set students on a higher achievement trajectory throughout their time in school. If the associations between early change and later achievement reported by these correlational studies approximate causal effects, then such long-run impacts could be expected from educational interventions that successfully promote early mathematics learning. Although past studies of early math change controlled for a host of child characteristics, including initial level of mathematics achievement, it is still unclear whether the regression-adjusted association between early change in mathematics and later achievement represents a causal effect. Here, we ask: Do early mathematics gains produced by random assignment to an intervention predict later math achievement as strongly as the naturally occurring gains used in past studies? If the associations reported in past studies are driven by unobserved characteristics, such as interest, motivation, parental support for mathematics, or cognitive aptitude, then even highly successful early mathematics interventions may have no detectable impact on later achievement.

Indeed, experimental and observational studies suggest that the regression-adjusted associations reported by correlational research overstate the potential long-run impacts of early mathematics intervention. Bailey, Watts, Littlefield, and Geary (2014) hypothesized that the stable correlation observed between measures of early mathematical ability and the sequence of later mathematics measures may be due to stable but unobserved factors that heavily influence mathematics achievement throughout development. Using a latent-factor state–trait model, they separated the variance in longitudinal measures of mathematics achievement into time-variant (state) and time-invariant (trait) components. They found that most of the variation in repeated measures of mathematics achievement was trait-like, as variation in individual differences in mathematics achievement were highly stable over time. Conversely, changes in any single measure of mathematics ability had relatively small effects on subsequent achievement scores once the stable variance was partitioned into a single, latent factor. They concluded that correlational studies investigating the association between early and

later measures of achievement fail to take into account the multitude of stable environmental and individual factors that likely influence achievement over time, and this omission leads to an overstatement of the importance of early measures of achievement on later measures.

Furthermore, experimental evidence from intervention studies also suggests that long-run correlational models may not accurately represent causal impacts. *Building Blocks*, a preschool mathematics curriculum designed by Clements and Sarama (2008), was evaluated as part of a multisite scale-up evaluation of an intervention model called TRIAD (Technology-Enhanced, Research-Based, Instruction, Assessment, and Professional Development; see Clements, Sarama, Spitler, Lange, & Wolfe, 2011; Clements, Sarama, Wolfe, & Spitler, 2013). In the TRIAD evaluation study, state-preschool programs were randomly assigned to either a curriculum implementation condition or a business-as-usual control condition. Although the intervention produced a large impact on mathematics achievement at the end of preschool (Hedge's $g = .72$), this effect faded by over 60% by the end of first grade (Clements, Sarama, Spitler, et al., 2011; Clements et al., 2013). The fadeout pattern reported by Clements et al. resembles the results of a meta-analysis of early childhood education interventions (Leak et al., 2010), which found that most early interventions faded substantially in the years immediately following the end of treatment. Moreover, recent evidence from a large-scale middle childhood mathematics intervention has shown similar fadeout effects (Taylor, 2014).

Although these intervention findings dim hopes that producing gains in early mathematical skills might transform long-run academic trajectories, analysis of intervention effects do not directly test the causal returns of early skill gains. Even if an early intervention such as TRIAD produced a large boost in skills during the treatment period, estimates of the intervention's impact on later-grade math achievement would merely test the effect of being assigned to the treatment group on later achievement not the effect of students' math skill gains across the treatment period. Furthermore, traditional "treatment on the treated" analyses in such contexts test the effect of actually participating in the program on later outcomes, but this analysis still falls short of directly examining the long-run effects of growth in early skills.

If we want to understand how long-run developmental trajectories might be altered as a result of spurring early gains in academic skills, a different analytic approach is needed. To be effective, this approach would need to separate variation in early mathematics change from sources of unobserved characteristics (e.g., child IQ, parental investment, interest) that might induce an upward bias in the estimated relationship between early skill gains and later achievement. Yet, unlike long-run analyses of intervention effects, this approach should also directly test the effect of early mathematics change on later achievement, not the effect of program participation, or assignment to a program, on later measures of math ability.

Current Study: Instrumental Variables

To obtain a causal estimate of the association between early mathematical skill change and later achievement, the current study employs instrumental variables (IV) techniques, which are widely used in applied econometric studies (see Angrist & Pischke, 2008; Murnane & Willett, 2010). IV methods have recently garnered considerable attention from developmental scientists; Gennetian, Magnuson, and Morris (2008) demonstrated the potential utility of the method for answering questions of theoretical importance in developmental psychology. Auger, Farkas, Burchinal, Duncan, and Vandell (2014) employed IV for estimating the causal impact of child-care quality on later academic outcomes, and Crosby, Dowsett, Gennetian, and Huston (2010) used IV to examine the impact of child-care type on child behavioral problems.

The intuition behind an IV approach is relatively simple: If the variation in a theoretically interesting predictor variable can be purged of the portion of its variation that stems from unobserved factors (i.e., selection bias), then the "clean" variation left can be used to estimate a causal effect. To generate this clean variation, the observational data set must contain a variable (i.e., instrument) that satisfies two conditions. First, the instrument must have a strong effect on the predictor of interest (in our case, early math gains). Second, the instrument can only affect the eventual dependent variable of interest (in our case later-grade achievement) through the main predictor. In other words, the effect of the instrument on the dependent variable should be completely mediated by the key endogenous predictor. Both requirements are essential to the success of the IV analysis, and finding instruments that satisfy these criteria in developmental research can be difficult (Gennetian et al., 2008).

In the current study, we seek to identify the causal impact of early mathematical skill change on

later mathematics achievement. We test this causal relation by leveraging random assignment within the TRIAD scale-up evaluation as an instrument for preschool mathematics change. We then relate this "exogenously produced" change (i.e., the change in mathematics learning that is only due to random assignment to the intervention, not other personal or environmental factors such as cognitive ability or parenting) to mathematics achievement measured in fourth and fifth grades. We chose the fourth- and fifth-grade outcome measures because they closely align with the time at which outcomes were measured in previous correlational work (e.g., Duncan et al., 2007) and because they were the most distal measures of mathematics achievement available in the data.

To produce exogenous variation in preschool math change, we take advantage of the fact that the *Building Blocks* intervention randomly assigned treatment to classrooms within clusters of preschools (called "blocking groups" and described below). The intuition behind our IV approach is that, to the extent that the relationship between early math change and later math achievement is causal, preschool clusters showing particularly large treatment impacts on math gains across the preschool year should also show larger-than-average impacts on later-grade achievement. The IV estimate is essentially the ratio of the later-grade impacts to early-gain impacts—both of which are produced by random assignment to treatment status. Mechanically, we use blocking group and treatment status interactions as instruments for early mathematics change in a two-stage least squares (2SLS) model (e.g., Duncan, Morris, & Rodrigues, 2011). The 2SLS estimator is a common technique for IV analyses (see Murnane & Willett, 2010).

If the IV criteria mentioned above are satisfied (i.e., the instrument strongly predicts preschool math gains, and the instrument only affects later achievement through its effect on preschool math learning), then the 2SLS model should provide an unbiased estimate of the causal effect of preschool mathematics change on later mathematics achievement. Prior research leads us to hypothesize that early mathematics change will have a causal effect on later achievement, because early success in mathematics is likely to improve the chances of later mathematics achievement through both skill acquisition and other related personal and environmental processes (e.g., boosting positive self-concepts, placement in higher achievement tracks in school). However, we expect that the causal impact will be smaller than the relations reported by correlational studies, as recent evidence suggests that omitted factors probably bias estimates of the association between early and later measures of mathematics achievement.

Method

Study Design

The design of the TRIAD scale-up evaluation is crucial to our analytic model. The intervention evaluation study researchers recruited 42 elementary schools with state-funded preschool programs serving low-income communities in New York and Massachusetts to participate in the evaluation, and they then grouped these schools into eight blocks. The blocking groups were determined based on fourth-grade state-collected achievement test scores alone and were not linked to district or other shared characteristics. This process was done to help ensure that schools in the treatment and control condition were balanced on unobserved characteristics (see Clements, Sarama, Spitler, et al., 2011).

Within each block, schools were randomly assigned to one of three conditions: (a) control condition (business as usual), (b) *Building Blocks* curriculum during preschool only, (c) *Building Blocks* curriculum during preschool with extended pedagogical development (PD) in kindergarten and first grade. Schools assigned to either treatment condition (i.e., Conditions 2 and 3) implemented the *Building Blocks* curriculum along with aspects of the TRIAD model that included PD and extensive instructional support (described below). Thus, the TRIAD evaluation study tested the success of the *Building Blocks* preschool curriculum in comparison with other preschool approaches to teaching mathematics, as students in the control condition still received mathematical instruction in their preschools (see Clements, Sarama, Spitler, et al., 2011). As explained below, our analysis focuses just on the first and second groups.

The *Building Blocks* curriculum (Clements & Sarama, 2013), implemented during preschool, was based on theory and research on early childhood learning and teaching. The basic approach was finding the mathematics in, and developing mathematics from, children's activities by helping children extend and mathematize these activities. All components were based on learning trajectories for each core topic. First, empirically based models of children's thinking and learning were synthesized to create a developmental progression of levels of thinking in the goal domain that emphasized

conceptual understanding, procedural skill, and problem-solving competencies. Second, sets of activities were designed to engender those mental processes or actions hypothesized to move children through a developmental progression.

Preschool teachers working in schools assigned to either treatment condition attended 13 PD sessions over the course of 2 years. The PD sessions were designed to help teachers understand the developmentally sequenced learning trajectories that form the basis of the *Building Blocks* curriculum, and teachers also learned the core mathematics procedures and concepts for each topic. Teachers were also trained to use formative assessment and the *Building Blocks* software, called Building Blocks Learning Trajectories (BBLT). BBLT was an individually paced program for students that was aligned with the curriculum and intended to provide additional instructional support. Finally, throughout the preschool year, teachers interacted with program mentors who offered instructional guidance and also assessed the fidelity of implementation. Analyses showed that teachers taught the curriculum with adequate fidelity (mode and mean of 1, "agree" on −2 to +2 Likert scale; see Clements & Sarama, 2011; Clements, Sarama, Spitler, et al., 2011). On an observational instrument focused on mathematics, *Building Blocks*, compared to control, teachers had significantly higher scores on the classroom culture scale, total number of mathematics activities observed, and the number of computers on and working for students to use. However, there were no observed statistically significant differences in the number of minutes mathematics was taught (Clements, Sarama, Spitler, et al., 2011).

The current study only considers children attending schools assigned to the preschool only treatment condition or control (school $N = 30$). Unfortunately, we were not able to use the alternative treatment condition in our current analyses as the requirements for a viable instrument (described below) were not met by this third condition. We describe our attempts to use the third, follow-on treatment arm in more detail in the Supporting Information.

The key component of our analyses, the IV, is derived by generating treatment by block interactions, which we then relate to preschool mathematics change. We use these interactions because we expect that some blocks were more successful at producing preschool mathematics change than others, and these block differences should produce more variation in intervention-caused preschool math learning. As explained above, our IV procedure assesses whether blocks with the largest treatment-induced gains in early math also produced the largest impacts on measures of fourth- and fifth-grade achievement.

Data

We use data drawn from the TRIAD evaluation study, which randomly selected 880 students from the preschool classrooms of the schools assigned to either the preschool curriculum intervention or the control condition. Students' mathematical knowledge was assessed at the beginning and end of preschool, spring of kindergarten and first grade, fall and spring of Grade 4, and the spring of Grade 5, and data were collected from the fall of 2006 (preschool year) through the spring of 2013 (fifth-grade year). The current study relies on data collected during preschool and Grades 4 and 5. As described below, we employ two separate model specifications. The first group of models uses a balanced panel, which only includes students with nonmissing test score data during preschool and Grades 4 and 5 (subsequently referred to as the "grade-pooled" sample; $n = 410$). The second group of models considers students that had data on any of the respective follow-up measures (fall of fourth grade $n = 469$, spring of fourth grade $n = 543$, spring of fifth grade $n = 502$). The missing cases in the grade-pooled sample are missing due to study attrition. Of the baseline characteristics assessed, only free or reduced price lunch (FRPL) status contains any nonresponse (approximately 20%), and nonresponse was not related to treatment status ($p = .30$). In the regression models that follow, FRPL was included as a covariate, and missing cases were set to 0. A dummy variable was then included in each regression indicating whether an observation had missing data on the FRPL indicator.

Table 1 presents sample characteristics for participants in the full sample, grade-pooled sample, treatment, and control. As Table 1 reflects, half of the students recruited for participation in preschool were African American, 23% were Hispanic, and 21% were White. Furthermore, 85% of the sample qualified for FRPL (only the 773 nonmissing cases were considered here). Students who are included in the pooled sample (i.e., those that did not attrit) were more likely to attend a New York school ($p<.001$). They were also more likely to be Hispanic ($p = .063$) and less likely to be White ($p = .027$). However, students in the pooled sample did not

Table 1
Sample Characteristics

	Full sample (1)	Pooled sample (2)	*p* value (3)	Treatment (4)	Control (5)	*p* value (6)
PreK entry math	−3.210 (.830)	−3.210 (.808)	.984	−3.249 (.856)	−3.164 (.795)	.467
Site						
New York	.725	.815	.001	.702	.753	.756
Massachusetts	.275	.185	.001	.298	.247	.756
Ethnicity						
African American	.502	.488	.275	.519	.482	.814
Hispanic	.231	.198	.063	.198	.270	.523
White–non-Hispanic	.211	.249	.027	.246	.169	.506
Other	.0557	.0659	.055	.0372	.0783	.237
Female	.497	.556	.003	.496	.497	.893
Age at PreK entry	4.359 (.352)	4.339 (.352)	.302	4.331 (.353)	4.392 (.348)	.382
Special education	.167	.156	.476	.173	.159	.678
Free/reduced lunch	.849	.850	.951	.824	.881	.25
Limited English proficiency	.167	.163	.417	.124	.220	.279
Observations	880	410	—	484	396	—

Note. The mean values are displayed for each variable. Standard deviations are in parentheses. Column 2 displays mean characteristics for students included in the primary analysis model, in which only participants who had nonmissing test score data in fall or spring of fourth grade and spring fifth grade were considered. Column 3 displays *p* values from regressions comparing students who were included in the pooled sample with those who were not. The *p* values listed in Column 6 indicate the extent to which treatment participants differed from controls. In each regression, standard errors were adjusted for clustering at the school level (30 schools).

statistically significantly differ on the preschool entry test, and were not more or less likely to be in the treatment or control group.

A comparison of Columns 4 and 5 from Table 1 shows that treatment and control groups were balanced on baseline observable characteristics, as no statistically significant differences were detected between the two groups.

Measures

Mathematics Achievement

During preschool, mathematics achievement was assessed at the beginning and end of the preschool year with the Research-based Early Math Assessment (REMA; Clements, Sarama, & Liu, 2008; Clements, Sarama, & Wolfe, 2011). The REMA was designed specifically for use with children ages 3 through 8, and it was administered through 2 one-on-one interviews with a trained administrator. The test was administered in two sections: number and geometry. Topics found on the number portion of the exam included counting, subitizing, number sequencing, cardinality, number composition and decomposition, place value, and adding and subtracting. Topics on the geometry part of the exam included shape recognition, congruence, measurement, patterning, and shape composition and decomposition.

The REMA included 225 items that were ordered according to difficulty. The study administrator stopped the exam once a student incorrectly answered four consecutive items. The testing process was videotaped and subsequently coded for correctness and strategy use. Approximately 10% of the assessments were double coded, and assessors and coders were blind to study condition. The REMA scores were then converted to Rasch–item response theory (IRT) scores to account for random guessing and item difficulty. The measure was validated in three diverse samples of young children, and it has been shown to have a .86 correlation with the Child Math Assessment: Preschool (see Clements et al., 2008), a .74 correlation with the *Applied Problems* subtest of the Woodcock–Johnson III (see Weiland et al., 2012), and strong internal reliability (Cronbach's $\alpha = .94$; see Clements et al., 2008). The REMA was also administered in the spring of kindergarten and first grade. The current study employs both the standardized Rasch–IRT scores and simple raw counts of the number of items correctly answered (subsequently referred to as "raw scores").

During the fall and spring of Grade 4 and spring of Grade 5, an extension of the REMA, called the

Tools for Early Assessment in Mathematics (TEAM) 3–5, was administered (Clements, Sarama, Khasanova, & Van Dine, 2012). The TEAM 3–5 is a paper-and-pencil assessment that can be administered in a group setting. Similar to the REMA, it is aligned with developmental progressions; although some topics are "retired" (e.g., simple counting, subitizing, shape recognition), others, similarly drawn from research-based developmental progressions (see Maloney, Confrey, & Nguyen, 2014; Wilson, Mojica, & Confrey, 2013), are introduced or receive greater emphasis (e.g., multiplication and division, fractions and decimals, measurement of area and volume, coordinate systems, and more sophisticated analysis of geometric shapes). In the current sample, the TEAM 3–5 was found to have good internal reliability (Cronbach's $\alpha = .91$). Furthermore, correlations between the assessment and state Grade 5 achievement tests in New York ($r(351) = .82$, $p < .001$), and Massachusetts ($r(110) = .76$, $p < .001$), were high for the subset of students for which state tests were available (approximately 40% of the full sample). As with the REMA, the TEAM 3–5 was also converted to a standardized Rasch–IRT score.

The key measure in the study, mathematics change, was constructed by taking the simple difference between the standardized post–preschool IRT-scored REMA and the standardized preschool entry IRT-scored REMA. Thus, model coefficients should be interpreted as "a standard deviation of change," which makes the effects most comparable to effect sizes reported in both intervention and correlational literature. However, because IRT scores can be difficult to interpret, we have also calculated a simple measure of the change in the raw number of items correctly answered on the pre- and posttests. When considering this measure in comparison with the IRT scores, recall that the IRT score takes into account correctness, as well as strategy use and item difficulty. Thus, the raw scores reflect a much simpler, and less comprehensive, measure of mathematics knowledge that do not have the characteristics of measurement that the IRT scores possess.

Table 2 presents descriptive statistics for both IRT-scaled and raw score measures of the pretest, posttest, and change measure for both the treatment and control groups. On average, students in the treatment group correctly answered approximately 11 items on the pretest, and students in the control group answered 12 items, a statistically nonsignificant difference ($p = .526$). By the end of preschool, students in the treatment group correctly answered approximately 21 more questions than on the pretest measure, and students in the control group correctly answered roughly 16 more items than on the pretest ($p < .01$). Thus, both groups grew substantially in their mathematics knowledge. The standardized IRT scores also reflect the substantial change students made in both the treatment and control groups. The REMA IRT scores were standardized to have a mean of zero at approximately first grade, thus the change from an average score of −3.25 for the treatment group at pretest to a score of −1.87 at the posttest reflects positive growth toward the normed first-grade mean.

Table 2
Math Change Descriptives

	Treatment	Control	*p* value
PreK entry math			
IRT score	−3.249 (0.856)	−3.164 (0.795)	.467
Number correct	11.46 (7.493)	12.10 (7.781)	.526
PreK post math			
IRT score	−1.872 (0.672)	−2.245 (0.749)	.004
Number correct	32.70 (12.11)	28.02 (12.07)	.022
PreK change			
IRT score	1.376 (0.705)	0.919 (0.650)	.001
Number correct	21.25 (8.647)	15.92 (8.053)	.001
Observations	456	378	

Note. Entries show means and standard deviations are shown in parentheses. The IRT scores were scaled such that a score of "0" approximates the achievement level of a student in first grade. The *p* value column lists *p* values from regressions in which each variable listed was regressed on treatment status. *p*s < .001 were rounded to .001.

Covariates

Information regarding child ethnicity, gender, age, limited English proficiency, special education status, and FRPL status were collected at baseline from the study schools' administrative data. The measures are included as controls in the following analyses.

IV Model

We used a 2SLS modeling procedure in Stata 13.0 (StataCorp; College Station, TX) to estimate the causal effect of preschool mathematical skill change on later mathematics achievement. In the first-stage regression, we regressed our key predictor, preschool mathematics change, on treatment status, blocking group, preschool-entry mathematics achievement, baseline measures of student characteristics, and most importantly, the interaction

between treatment status and blocking group. The resulting equation for the *i*th child in the *j*th block is as follows:

$$\begin{aligned}\text{MathChange}_{ij} = \; & a_1 + \beta_1 Tx_{ij} + \sum_{j=1}^{8} \beta_2 \text{Block}_j \\ & + \beta_3 \text{Block}_j * Tx_{ij} + \beta_4 \text{EntryMath}_{ij} \\ & + \beta_5 \text{Covariates}_{ij} + e_{ij} \end{aligned} \quad (1)$$

where MathChange_{ij} is the posttest math score subtracted from the pretest math score of the *i*th student in the *j*th block, and the instruments are represented by the treatment dummy variable (Tx_{ij}) and the treatment and block interactions (Block × Tx_{ij}). The use of interactions between random assignment design characteristics (such as site) and treatment status as instruments has been used in other quasi-experimental studies of educational settings (Auger et al., 2014; Duncan et al., 2011; Taylor, 2014). The second-stage regression, which estimated the impact of preschool math change on later achievement, then used the predicted values for preschool math change generated in the first equation:

$$\begin{aligned}\text{MathAchievement}_{ijt} = \; & a_1 + \theta_1 \text{PredictedMathChange}_{ij} \\ & + \sum_{j=1}^{8} \theta_2 \text{Block}_j + \theta_3 \text{EntryMath}_{ij} \\ & + \theta_4 \text{Covariates}_{ij} + z_{ij} \end{aligned} \quad (2)$$

where $\text{MathAchievement}_{ijt}$ represents the math achievement test score for the *i*th child, in blocking group *j*, at time *t* (either fall or spring of fourth grade, or spring of fifth grade). In this equation, the instruments from the first equation (treatment status, and treatment and block interactions) do not appear, and θ_1 represents the causal impact of preschool mathematical skill change on later achievement. If the key IV assumptions described below are satisfied, then z_{ij}, the error term, should only represent random shocks, and should not include the sources of omitted variable bias that typically plague correlational models.

Whenever IV methods are employed, the instrumented parameter of interest (θ_1 in Equation 2) should be interpreted as the local average treatment effect (LATE), where "local" describes compliant students (see Angrist & Pischke, 2008; Murnane & Willett, 2010). In other words, IV methods only identify the effect for participants who were compelled to participate in the treatment based on random assignment. In our setting, this means that we identify the effect of preschool mathematics change for students who grew in mathematics only as a result of random assignment to the treatment.

As described in more detail below, we estimated separate 2SLS models for fall and spring of fourth-grade and spring of fifth-grade measures of mathematics achievement, respectively. However, we also estimated models in which we pooled mathematics achievement scores across these three grades. All models presented included robust standard errors that were adjusted for clustering at the school level.

Correlations Between Instruments and Mathematics Change

To be effective in an IV analysis, an instrument must have a strong effect on the endogenous predictor variable. In this case, the Treatment × Block interactions need to produce enough variation to reliably predict mathematics change in Equation 1. Indeed, in the intervention considered here, the treatment was specifically designed to affect mathematics change during the preschool year. However, some blocks may have been more successful at this goal than other blocks. To assess the correlation between the instruments and preschool mathematics change, we ran a regression predicting our key measure of preschool mathematics change on baseline characteristics (including preschool-entry mathematics score), block and treatment dummies, and interactions between treatment and block. With standard errors adjusted for clustering at the school level, the joint test for the set of treatment and block interactions produced a large-enough *F* statistic, $F(8) = 41.46$, $p < .001$, to confidently conduct 2SLS analyses, as an *F* statistic of 10 is usually considered the threshold for an effective instrument (e.g., Angrist & Pischke, 2008). Column 1 of Table 3 displays the coefficients produced by this model, including the block and treatment interactions. Block 5 was omitted from the regression as the comparison group, as this was the block with the most students ($n = 162$). In this model, the treatment had a large main effect ($\beta = .699$, $SE = .138$), and some blocks produced positive interactions with treatment status, whereas others produced negative coefficients. This indicates considerable variability between blocks on the effect of the treatment on mathematics change.

Exclusion Restriction

To produce only exogenous variation in the endogenous predictor, the instrument should not be

Table 3
OLS Models Predicting Preschool Change and Late Elementary School Math Achievement

	Later achievement			
	Math change (1)	Fall of fourth grade (2)	Spring of fourth grade (3)	Spring of fifth grade (4)
Math change		.568 (.044)***	.582 (.042)***	.529 (.043)***
Treatment	.699 (.138)***	−.313 (.047)***	−.371 (.041)***	−.234 (.061)***
Controls	Inc.	Inc.	Inc.	Inc.
Blocking group	Inc.	Inc.	Inc.	Inc.
Block × Treatment				
1	−.127 (.262)			
2	−.320 (.135)*			
3	−.281 (.165)			
4	.102 (.162)			
6	−.262 (.186)			
7	−.189 (.187)			
8	.045 (.182)			
Observations	834	469	543	502
R^2	.425	.499	.496	.448

Note. Robust standard errors were adjusted for clustering at the school level and are displayed in parentheses. In each model, the dependent variable was standardized, as was math change and age. Column 1 displays coefficients produced by treatment and Block × Treatment group interactions (the main component of the instrumental variable analysis) predicting math change during preschool. Columns 2 through 4 display the results of OLS models predicting standardized math achievement in Grades 4 and 5, respectively, with baseline characteristics and preschool math change. "Inc." denotes the inclusion of various sets of control variables. Coefficients produced by control variables (PreK entry math, gender, race, whether limited English proficient, age, whether designated for special education, whether free or reduced price lunch, site, and blocking group) can be found in the Supporting Information. *$p < .05$. ***$p < .001$.

correlated with the error term in Equation 2. In other words, the instrument should not have an effect on the dependent variable (late elementary school mathematics achievement) except through the endogenous predictor (preschool mathematics change).

Theoretically, this should be the case in the current analysis. The model by which the intervention was designed conceptualizes the impact of the intervention on elementary school mathematics achievement through a skill-building framework that hinges upon gains made in preschool mathematics achievement (Clements, Sarama, Spitler, et al., 2011; Clements et al., 2013). Thus, future mathematical skill production relies on the mathematics skills children carry at the end of preschool, as the preschool mathematical competencies allow them to learn and master new, more difficult material. Furthermore, we found no differences in baseline observables between the treatment and control groups (see Table 1), indicating that at baseline, the treatment group was not advantaged in a way that would have improved their chances of becoming high achievers later on.

However, it is possible that the intervention could have affected later elementary school mathematics achievement through other mechanisms, such as boosts in language skills, motivation, or executive functioning. Furthermore, treatment students could have been sorted into higher quality classrooms after preschool, which could have, in turn, boosted their later mathematics achievement. Our data include observational measures of classroom instructional quality from the children's kindergarten and first-grade classrooms (observations were recorded for approximately 73% of the current analysis sample; see Clements et al., 2013 for full description of the observational measure). We found no indication that treatment status was correlated with kindergarten or first-grade instructional quality. We also found that treatment status was not related to the likelihood of staying in the same school through kindergarten, first grade, or fifth grade.

Unfortunately, we lack the broad measures of child characteristics needed to rule out unexpected changes in child functioning due to the preschool mathematics intervention. However, language skills were measured at the beginning of the kindergarten year, and Sarama, Lange, Clements, and Wolfe (2012) reported a standardized statistically significant treatment impact of approximately .10 on the measure of language achievement (measure

included the ability to recall key words, use of complex utterances, willingness to reproduce narratives independently, and inferential reasoning). We tested whether this boost in language skills could bias our models by running our primary OLS and IV models with, and without, the kindergarten entry language score. Including the language measure did not change our estimates (results shown in the Supporting Information), indicating that although the treatment impacted language functioning, this boost in language did not affect later mathematics achievement.

Given that the intervention was only the implementation of a preschool mathematics curriculum (that ran for approximately 15 min per day; Clements, Sarama, Spitler, et al., 2011) not a global program targeted at a wide array of socioemotional and cognitive skills; it seems most plausible that the primary mechanism through which the intervention affected students was through preschool mathematical skill development. Still, we cannot rule out whether the treatment might have caused changes in unobserved child characteristics, such as motivation or executive functioning. In both cases, changes in these unobserved skills could bias our estimates if boosts in these skills also impacted later mathematics achievement. Previous correlational studies that have examined relations between mathematics achievement and various socioemotional and cognitive skills suggest that any likely bias-causing candidate would probably have a small effect on our model (e.g., Claessens & Engel, 2013; Duncan et al., 2007; Jordan et al., 2009; Watts et al., 2014). Nevertheless, if such biases were present in our models, they would likely have positive correlations with later mathematics achievement and preschool change and would then bias our key estimate in an upward direction. Because we lack the measures to totally rule out this potential threat, our findings should be considered upper-bound estimates of the causal relation between preschool mathematical skill change and later mathematics achievement.

Grade-Pooled Estimates

In the analyses that follow, we rely primarily on estimates generated from a grade-pooled data set. In these models, we pooled observations across the fall and spring of fourth grade and the spring of fifth grade, such that each student was observed three times, and students were only included in this sample if they had nonmissing data on both fourth-grade measures and the fifth-grade test (n = 410). We chose this path for two reasons. First, IV models typically generate relatively large standard errors, because IV models depend only on variation produced by the instruments and thus have less variation with which to produce estimates (Angrist & Pischke, 2008). Thus, to generate precise estimates, more statistical power is required.

Second, this model is justified by the high correlations between the fourth- and fifth-grade test scores, as these measures each had an average correlation of .84. Furthermore, after pooling across grades, we regressed fall of fourth-grade, spring of fourth-grade, and spring of fifth-grade mathematics achievement on preschool mathematics change and covariates. In this model, we included dummies for grade level and interactions between grade and change. This set of interactions, which test whether the relation between change and later achievement differs between grade levels, were jointly not statistically significantly different from 0, $F(2) = 0.50$, $p = .610$.

However, because the impact of preschool change on achievement at different grade levels is of theoretical interest, we also present models that were estimated using nonpooled data. In these models, fall and spring of fourth-grade and spring of fifth-grade achievement were each regressed independently on instrumented-preschool mathematics change.

Results

We begin with results from OLS models in which we regressed our later measures of mathematics achievement (fall and spring of fourth grade and spring of fifth grade) on preschool mathematics change, preschool entry mathematics achievement, and other baseline characteristics. Columns 2 through 4 of Table 3 presents results from nonpooled OLS models in which we examined the relation between preschool mathematics change and fourth- and fifth-grade mathematics achievement, respectively. Key independent and dependent variables were standardized, and all models presented included the full list of control variables (correlations for all predictor variables are shown in the Supporting Information). Columns 2 through 4 show the relatively stable predictive relation between preschool mathematics change and later achievement, as a standard deviation of change had approximately a one-half standard deviation effect on fall and spring of fourth-grade and spring of fifth-grade achievement. The effects reported in

Columns 2 through 4 are larger than the OLS-adjusted effects of early mathematical skill change reported by Claessens et al. (2009) and Watts et al. (2014), as their studies produced standardized effects of approximately .35. This discrepancy probably reflects the greater availability of cognitive control measures available in the data sets employed by those studies.

Grade-Pooled IV Estimates

Next, we turn to estimates generated from pooled models that used block and treatment interactions as instruments for preschool mathematics change. Recall that in the pooled models, each student's fourth- and fifth-grade tests were considered as separate observations in one model. In each model, standard errors were adjusted for school-level clustering, but we also tested models that adjusted for student-level clustering to account for the panel structure of the data set, and results did not qualitatively differ.

In Column 1 of Table 4, we begin with the reduced form estimates, which show the effect of the instrument on the eventual outcome variable of interest. In our study, the reduced form model can be interpreted as a basic treatment impact model, as we show the average treatment impact of random assignment to the TRIAD intervention on mathematics achievement in fall and spring of fourth grade and spring of fifth grade. Across the grades, the average treatment impact was positive but not significant ($\beta = .094$, $SE = .064$, $p = .154$). However, our IV results suggest that the simple treatment impact estimate masks the effect of treatment-induced change in mathematics on later achievement.

For the purposes of comparison, Column 2 of Table 4 presents grade-pooled OLS results comparable with the estimates displayed in Columns 2 through 4 of Table 3, as a standard deviation of preschool mathematics change was related to a 0.535-*SD* gain in later mathematics achievement ($SE = .044$, $p < .001$). Column 3 displays the 2SLS-estimated (IV) impact of standardized mathematics change on later achievement with only site blocking group and preschool entry math score controlled. In this model, the effect fell by over 50% when compared with the OLS models, though the estimate was still substantively and statistically significant ($\beta = .236$, $SE = .113$, $p = .037$). In Column 4, we added the full list of background characteristics, and the coefficient was nearly unchanged, though the standard error fell, reflecting the control

Table 4
Instrumental Variable (IV) Estimates Relating Preschool Change to Late Elementary School Achievement

	Reduced form (1)	OLS (2)	IV Reduced control (3)	IV Full controls (4)	IV Fall of fourth grade (5)	IV Spring of fourth grade (6)	IV Spring of fifth grade (7)
Math change		.535 (.041)***	.236 (.113)*	.242 (.081)**	.132 (.109)	.039 (.096)	.257 (.079)**
Treatment	.094 (.064)						
Controls							
Entry math score	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Site	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Block	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Background characteristics	Inc.	Inc.		Inc.	Inc.	Inc.	Inc.
Observations	1,230	1,230	1,230	1,230	469	543	502

Note. Robust standard errors, shown in parentheses, were adjusted for clustering at the school level. IV estimates were generated using two-stage least squares. In the models presented in Columns 1 through 4, students were observed three times (fall and spring of fourth grade and spring of fifth grade). In the models presented in Columns 5 through 7, students were observed only once, and the measurement point of the dependent variable varies in each model. "Inc." denotes the inclusion of various sets of control variables. The dependent variable, late elementary school math achievement, was within-grade standardized, and the main independent variable, preschool math change, was also standardized. See Table 3 note for the full list of background controls. Coefficients produced by background controls can be found in the Supporting Information. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

variables' added utility for increasing precision (β = .242, *SE* = .081, *p* = .003). The lack of change in the coefficient on preschool change after the addition of these control variables provides some degree of confidence that the exclusion restriction assumption is fairly safe in our models, as this indicates that the relation between instrument-produced change and later achievement was not correlated with baseline observables.

Additional Models

Column 5 through 7 present 2SLS estimates generated from non pooled models in which every student was only observed one time, and the fall of fourth-grade, spring of fourth-grade, and spring of fifth-grade scores were considered individually. We present these models because they can provide theoretically interesting information regarding whether the relation between exogenously produced mathematics change and later achievement may differ by grade. However, we hesitate to draw strong inferences based on these models because our sample sizes drop considerably in each of them, and this limits our ability to generate precise estimates when using IV (see Angrist & Pischke, 2008). Thus, these models merely inform the primary estimates presented in Columns 3 and 4 of Table 4, but drawing strong conclusions based solely on these models would be inadvisable.

As Columns 5 and 7 demonstrate, the significant and positive effect detected in the pooled models was not found in models relating change to either measure of fourth-grade achievement. Although the fall of fourth-grade model presented a positive coefficient with a large standard error (β = .132, *SE* = .109, *p* = .223), the spring of fourth-grade model produced a coefficient of nearly zero (β = .039, *SE* = .096, *p* = .683). However, we were surprised to find that preschool math change strongly predicted fifth-grade mathematics achievement in our disaggregated IV model (β = .257, *SE* = .079, *p* = .001). It would seem that the fifth-grade effect was largely driving the positive grade-pooled estimate, as the grade-pooled estimate roughly represents an average of the three disaggregated effects.

In the Supporting Information, we present results from additional analyses in which we estimated our grade-pooled IV model in only key subgroups (i.e., African Americans, limited English-proficient students, high and low achieving students, FRPL students). Across all groups, we found positive effects within the confidence interval of our key estimate shown in Column 3 of Table 4. We found the largest effect for African American children (β = .379, *SE* = .104, *p* < .001), but we did not find that this effect was statistically significantly different from the effect for non-African American students (*p* = .150).

In analyses presented in the Supporting Information, we also tested the sensitivity of our primary findings to various model specifications. As mentioned above, we tested whether controlling for kindergarten measures of language and literacy skills changed our results, and we found no indication that our models were affected by these measures. Furthermore, we examined models that did not control for baseline mathematics achievement and found that this did not substantively change our estimates. Next, we tested whether controlling for grade level changed the grade-pooled IV estimates, and again found that our results were robust to this specification. Finally, we tested whether changing our IV estimation procedures affected our results. We found that using only the single treatment status indicator as an instrument produced a positive, marginally statistically significant coefficient of .154 (*SE* = .094, *p* = .104), and using "limited information maximum likelihood" IV estimator instead of the 2SLS estimator produced a coefficient quite similar to the one reported in Column 3 of Table 4.

Discussion

The current study tested the extent to which learning mathematics during preschool improves mathematics achievement in late elementary school. We leveraged variation in preschool learning produced by a preschool mathematics intervention to generate causal estimates of the impact of gains in preschool mathematics knowledge. In our main models, we found that a 1-*SD* boost in preschool math learning produced approximately a quarter *SD* gain in late elementary school achievement. However, we were surprised that this relation was only detected between preschool math learning and fifth-grade achievement, and we found no such association between preschool gains and fourth-grade achievement.

Taken together, these results lead us to make two primary conclusions. First, correlational approaches to questions regarding longitudinal achievement patterns should be approached with great caution. Second, early learning does not appear to be an "inoculation" that necessarily

produces later achievement gains, and consequently, theories regarding skill-building processes probably require some amount of revision.

Comparisons With Correlational Literature

Our results suggest that the correlational literature, based primarily on OLS models that controlled for a host of family and child background characteristics, probably overstated the long-run effects of preschool mathematics achievement. When compared with OLS models estimated in the current study, the IV models reduced the effect of preschool change on later mathematics achievement by nearly 50%. When considered alongside the intervention literature, perhaps this finding should not be surprising, as preschool interventions often show steady fadeout patterns as time after the end of treatment elapses. Yet, why did the correlational literature fail to predict the modesty of the causal relation between early math skill gains and later achievement?

The answer could simply be that it is nearly impossible to control for all of the potential confounds between early and later test scores. Indeed, previous correlational investigations (Claessens et al., 2009; Watts et al., 2014) included a wide array of cognitive, academic, and socioemotional skills not included in our study, but these controls apparently failed to account for all of the underlying sources of bias. Watts et al. (2014) even controlled for *gains* in reading achievement and domain-general cognitive skills, and still found a 1-*SD* gain in early math achievement was associated with a 0.37-*SD* boost in late elementary school achievement. When compared with our grade-pooled models, these estimates are approximately 35% larger than the 0.24-*SD* effect that we found using IV (it should be noted that the 95% confidence interval for our primary grade-pooled model ranged from .08 to .39).

Furthermore, compared with previous examinations, we did not find the IV-produced effect of preschool change to be consistent across grades, as we found no evidence of a strong relation between change and achievement in fourth grade, but we detected a substantial link between change and achievement in fifth grade. Certainly, the developmental period over which change was measured should be considered when drawing such comparisons, as Claessens and colleagues measured mathematical skill change during kindergarten, and Watts et al. measured mathematics change from preschool through the end of first grade. It is possible that change during kindergarten or first grade could be a stronger predictor of later achievement than change during preschool. Yet, given that we found a comparably large, OLS-adjusted, relation between preschool change and later achievement, we find it unlikely that this difference accounts entirely for the discrepancies between our IV estimates and the associations reported in previous correlational research.

If previous correlational models simply lacked the necessary set of controls, what factors might need to be controlled if correlational models stand a chance of replicating causal estimates? Indeed, future work should seek to find the set of measures that can fully reduce bias in analyses of longitudinal academic achievement data, and it is likely that such measures would need to include indicators of a wide variety of environmental and personal characteristics that could influence the development of math achievement over time. However, a few recent investigations also demonstrate that alternative approaches to modeling correlational data may provide a more productive path forward. Bailey et al. (2014) found that a state–trait model, which accounted for omitted-variables bias by modeling the stable variation present in repeated measures of mathematics achievement as a single, latent factor, substantially reduced the predictive relation between gains in an early measure of math ability and later measures of achievement.

Alternatively, the current article provides another possible approach for generating more accurate causal predictions. If researchers can find instruments that satisfy the criteria described above, then such analyses could better improve our understanding of many developmental processes, as this approach is not necessarily limited to investigations of cognitive and academic development. Finding viable instruments is no easy task, but other quasi-experimental approaches can also provide more robust causal estimates (see Murnane & Willett, 2010 for an approachable review of a variety of quasi-experimental methods). For example, Cortes and Goodman (2014) found that students who were approximately randomly assigned to an extra mathematics course in high school (generated from a regression discontinuity in assignment based on prior-year math scores) had higher graduation rates and were more likely to attend college. Such findings provide robust causal evidence of the possible benefits of mathematics education and offer an important test of developmental theories that would predict better outcomes for students with enhanced math learning opportunities. Thus,

although quasi-experimental methods may be difficult to pursue, the benefits of generating more accurate causal estimates should make such efforts worthwhile.

Implications for Developmental Theory and Practice

Our most surprising result, perhaps, was that we found a strong impact of instrument-produced change on fifth-grade mathematics achievement, but we found no impact on achievement in our 2 fourth-grade measures of math ability. We did not hypothesize this pattern of results, and because these models were less precisely estimated than our grade-pooled models, we do not wish to overstate these findings. Nevertheless, when considering what processes might have given rise to these results, recall that the same test was administered at both fourth-grade measurement points and at the spring of fifth-grade measurement point. Thus, changes in the measure should not account for differences in the pattern of findings. However, it is likely that the curriculum students encountered in school changed substantially between the fourth- and fifth-grade years. During the fifth-grade year, the schools in Massachusetts and New York both switched to the Common Core Standards, which emphasizes conceptual understanding of mathematics (Common Core Standards Initiative, 2010). Furthermore, it has been argued that this shift toward conceptually focused math would especially alter the way math was taught in low-income schools (Schmidt & Burroughs, 2013).

It is quite possible that the knowledge gained from the intervention during preschool only benefited students once the more conceptually rich content was emphasized in fifth grade. Certainly, this finding warrants further investigation and replication before major conclusions can be drawn. Yet, it should be noted that even if preschool math change only positively impacted mathematics achievement in fifth grade, but not fourth grade, then this finding strongly contradicts the predictions made by correlational models. Previous studies (e.g., Claessens & Engel, 2013; Duncan et al., 2007; Watts et al., 2014) have all reported stable relations between early mathematics achievement and later measures of achievement, no matter when the dependent variable was measured. Indeed, these findings led previous studies to predict that early intervention efforts would have stable long-run effects (Duncan et al., 2007; Watts et al., 2014). Our findings suggest that this is not likely to be the case.

Our pattern of results has implications for developmental theory. If our fifth-grade finding is found to be robust to replication, then this would suggest that skill-building processes do not necessarily unfold in a monotonic manner. In other words, early math skills might not reliably lead to the development of later mathematical knowledge across all settings. Rather, early mathematical knowledge may only lead to the production of later knowledge when this early knowledge base is paired with the correct mix of content and teaching. This suggests that subsequent environments play a critical role in sustaining cognitive development in the wake of early investments in cognitive skills. This also suggests that skill-building theories that predict that early knowledge gains will necessarily lead to advantages in later achievement (e.g., Cunha & Heckman, 2008) may need some revision, as our results imply that skill development may be a more complex process that relies on many factors other than the mere possession of early skill advantages.

However, we also wish to underscore that our preferred estimates, the grade-pooled models, suggested that intervention-spurred early gains in mathematics led to approximately a fifth of a *SD* gain in mathematics across fourth and fifth grades. This implies that early skill gains do matter for developing long-run achievement trajectories. Although the effect was not as large as was previously predicted by correlational work (e.g., Duncan et al., 2007), our results do demonstrate the long-run utility of early skills advantages. When considering what these results imply for developmental theory and practice, we should recall the "LATE" interpretation of IV results (see Angrist & Pischke, 2008). IV techniques identify effects for the "complier" population within the sample. In our study, compliers are students who responded to the intervention and gained in mathematics knowledge as a result of participation in the program. This is perhaps intuitive, as this means that we identified the effect of early math gains for students that, for whatever reason, were able to particularly benefit from participation in *Building Blocks.* Understanding what types of students respond best to early academic programs, like *Building Blocks*, presents a promising avenue for further research, as it opens the door for targeting programs toward students that might stand to benefit the most from early cognitive investments.

Although our results imply that early gains in mathematics ability should lead to moderate advantages in math achievement later in elementary school, for interventions, it is important to consider the amount of change that would be required of a

program to replicate the effect reported here. For an intervention effect to produce a 1-*SD* end-of-treatment effect on mathematics gains, students in the treatment group would need to gain a full standard deviation *more* in mathematics achievement than students in the control group. Although our raw score measure compares imperfectly to the standardized Rasch–IRT scores (recall that IRT scores take into account strategy use and item difficulty), the raw scores presented in Table 2 show that students in the control group still learned a considerable amount of mathematics during preschool. If we trace the raw score means back to the test items, our results suggest that students would need to move from simple number recognition to addition and subtraction by the end of preschool to produce a full standard deviation of change beyond the control group. Although such a progression in average mathematical ability during preschool may not be impossible, current data from nationally representative samples indicates that addition and subtraction are taught far less than more simple mathematics topics in even kindergarten, and only 5% of students have mastered adding and subtracting at kindergarten entry (Engel, Claessens, & Finch, 2013). Thus, our results likely reflect an upper bound estimate of the probable long-run effects of successful early math interventions.

Limitations and Conclusion

The results should also be considered against the limitations of the study. As was discussed previously, the exclusion restriction assumption could be violated if the intervention affected later mathematics achievement through unknown pathways unaccounted for by the present models. Unfortunately, we lack the data to extensively test for extraneous treatment–effect pathways. Yet, we found no evidence that boosts in language skills might have also affected later mathematics achievement, and our results did not change with the inclusion of background control variables. We also tested whether students in the treatment group were more likely to remain in the same school throughout the elementary school years, and whether they entered into higher quality kindergarten and first-grade classrooms. In both cases, we found no evidence that treatment students' schooling environments changed after the treatment year. This also suggests that peer effects should not bias our results, as students in the treatment group were not more likely to remain in school with the same peers than students in the control condition.

Furthermore, although we employed fairly comprehensive measures of mathematics achievement, it is likely that these measures still failed to capture all dimensions of children's mathematics knowledge. Thus, it remains possible that the benefits of gains in early math skills were not fully detected by the later mathematics measures. Finally, when interpreting our results, one should recall that our models were only tested within a relatively low-income sample of children. Thus, it is unclear how our results might relate to students from different socioeconomic backgrounds. This further implies the need for replicating our results in diverse settings and samples.

Nevertheless, the threat of omitted variable bias was not completely eradicated, meaning the current estimates produced by the 2SLS models likely reflect upper-bound estimates of the effect of intervention-caused mathematics change on later math achievement. Thus, although we found some indication that a standard deviation of change during preschool might lead to approximately a quarter of a standard deviation gain in later mathematics achievement, intervention fadeout is likely to be substantial even in the years following a treatment successful enough to produce an average treatment effect of a full standard deviation. As a result, if educational practitioners and policymakers wish to produce early childhood interventions that sustain effects in the years following the end of preschool, time and attention might be better placed on developing methods designed to build upon preschool gains during the early elementary school years (see Clements et al., 2013 for description of a follow-through treatment that abated early intervention fadeout effects to a degree).

In sum, the current article demonstrated the use of a quasi-experimental method for better understanding how mathematics skills develop during the early and middle childhood years. Our results illustrate that previous correlational approaches overstated the long-run benefits of early math intervention and that more robust approaches are necessary for generating better causal estimates. Furthermore, such approaches are also fundamental to our ability to test developmental theories, as the current findings imply that early math skills do not automatically lead to future academic success.

References

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

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

Aunola, K., Leskinen, E., Lerkkanen, M. K., & Nurmi, J. E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology*, *96*, 699. doi:10.1037/0022-0663.96.4.699

Bailey, D. H., Watts, T. W., Littlefield, A. K., & Geary, D. C. (2014). State and trait effects on individual differences in children's mathematical development. *Psychological Science*, *25*, 2017–2026. doi:10.1177/0956797614547539

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, *15*, 1–40. doi:10.1023/A:1021302408382

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, *42*, 189. doi:10.1037/0012-1649.41.6.189

Bronfenbrenner, U., & Morris, P. (2006). The bioecological model of human development. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (6th ed., pp. 793–828). New York, NY: Wiley.

Byrnes, J. P., & Wasik, B. A. (2009). Factors predictive of mathematics achievement in kindergarten, first and third grades: An opportunity–propensity analysis. *Contemporary Educational Psychology*, *34*, 167–183. doi:10.1016/j.cedpsych.2009.01.002

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review*, *28*, 415–427. doi:10.1016/j.econedurev.2008.09.00

Claessens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record*, *115*, 060306. Retrieved from http://eric.ed.gov/?id=EJ1020177

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, *45*, 443–494. doi:10.3102/0002831207312908

Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, *333*, 968–970. doi:10.1126/science.1204537

Clements, D. H., & Sarama, J. (2013). *Building blocks* (Vols. 1 and 2). Columbus, OH: McGraw-Hill Education.

Clements, D. H., Sarama, J., Khasanova, E., & Van Dine, D. W. (2012). *TEAM 3–5—Tools for elementary assessment in mathematics*. Denver, CO: University of Denver.

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The research-based early maths assessment. *Educational Psychology*, *28*, 457–482. doi:10.1080/01443410701777272

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *42*, 127–166. Retrieved from http://www.jstor.org/stable/10.5951/jresematheduc.42.2.0127

Clements, D. H., Sarama, J., & Wolfe, C. B. (2011). *TEAM—Tools for early assessment in mathematics*. Columbus, OH: McGraw-Hill Education.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies persistence of effects in the third year. *American Educational Research Journal*, *50*, 812–850. doi:10.3102/0002831212469270

Common Core State Standards Initiative (2010). *About the standards*. Retrieved from http://www.corestandards.org/about-the-standards

Cortes, K. E., & Goodman, J. S. (2014). Ability-tracking, instructional time, and better pedagogy: The effect of Double-Dose Algebra on student achievement. *The American Economic Review*, *104*, 400–405. doi:doi: 10.1257/aer.104.5.400

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. doi:10.1037/a0020384

Cunha, F., & Heckman, J. J. (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources*, *43*(4), 738–782. Retrieved from http://www.jstor.org/stable/40664515

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . & Sexton, H. (2007). School readiness and later achievement. *Developmental Psychology*, *43*, 1428–1446. doi:10.1037/0012-1649.43.6.1428

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. doi:10.1037/a0023875

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (mis) alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis*, *35*, 157–178. doi:10.3102/0162373712461850

Entwisle, D. R., & Alexander, K. L. (1990). Beginning school math competence: Minority and majority comparisons. *Child Development*, *61*, 454–471. doi:10.1111/j.1467-8624.1990.tb02792.x

Foster, E. M. (2010). The value of reanalysis and replication: Introduction to special section. *Developmental Psychology*, *46*, 973. doi:10.1037/a0020183

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PLoS ONE*, *8*, e54651. doi:10.1371/journal.pone.0054651

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. doi:10.1037/0012-1649.44.2.381

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, *79*, 1202–1242. doi:10.3102/0034654309334431

Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2007). Symbolic arithmetic knowledge without instruction. *Nature*, *447*, 589–591. doi:10.1038/nature05850

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*, 850–867. doi:10.1037/a0014939

Laski, E. V., & Siegler, R. S. (2007). Is 27 a big number? Correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. *Child Development*, *78*, 1723–1743. doi:10.1111/j.1467-8624.2007.01087.x

Leak, J., Duncan, G. J., Weilin, L., Magnuson, K., Schindler, H., & Yoshikawa, H. (2010). *Is timing everything? How early childhood education program impacts vary by starting age, program duration, and time since the end of the program*. UC-Irvine working paper, presented at the fall 2010 meetings of the Association for Public Policy Analysis and Management, Boston, MA. Retrieved from http://education.uci.edu/docs/Leak_Duncan_Li_Timing_Paper_APPAM_102810.pdf

Maloney, A. P., Confrey, J., & Nguyen, K. H. (Eds.). (2014). *Learning over time: Learning trajectories in mathematics education*. New York, NY: Information Age.

Meisels, S. J. (1998). *Assessing readiness* (Report no. 3-002). Ann Arbor, MI: Center for the Improvement of Early Reading Achievement. Retrieved from http://www.ciera.org/prod-ucts/meisels-1998/reports32.html

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. New York, NY: Oxford University Press.

Opfer, J. E., & Thompson, C. A. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development*, *79*, 788–804. doi:10.1111/j.1467-8624.2008.01158.x

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. New York, NY: Routledge.

Sarama, J., Lange, A. A., Clements, D. H., & Wolfe, C. B. (2012). The impacts of an early mathematics curriculum on oral language and literacy. *Early Childhood Research Quarterly*, *27*, 489–502. doi:10.1016/j.ecresq.2011.12.002

Schmidt, W. H., & Burroughs, N. A. (2013). How the common core boosts quality and equality. *Educational Leadership*, *70*, 54–58.

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*, 273–296. doi:10.1016/j.cogpsych.2011.03.001

Stevenson, H. W., & Newman, R. S. (1986). Long-term prediction of achievement and attitudes in mathematics and reading. *Child Development*, *57*, 646–659. Retrieved from http://www.jstor.org/stable/1130343

Taylor, E. (2014). Spending more of the school day in math class: Evidence from a regression discontinuity in middle school. *Journal of Public Economics*, *117*, 162–181. doi:10.1016/j.jpubeco.2014.06.002

Watts, T. W., Duncan, G. J., Chen, M., Claessens, A., Davis-Kean, P. E., Duckworth, K., . . . Susperreguy, M. I. (2015). The role of mediators in the development of longitudinal mathematics achievement associations. *Child Development*, *86*, 1892–1906. doi:10.1111/cdev.12416

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, *43*, 352–360. doi:10.3102/0013189X14553660

Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. H., Sarama, J. H., & Yoshikawa, H. (2012). Early mathematics assessment: Validation of the short form of a prekindergarten and kindergarten mathematics measure. *Educational Psychology*, *32*, 311–333. doi:10.1080/01443410.2011.654190

Wilson, P. H., Mojica, G. F., & Confrey, J. (2013). Learning trajectories in teacher education: Supporting teachers' understandings of students' mathematical thinking. *The Journal of Mathematical Behavior*, *32*, 103–121. doi:10.1016/j.jmathb.2012.12.003

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Table S1. Correlations Among Key Independent and Dependent Variables

Table S2. OLS Models Predicting Preschool Change and Late Elementary School Math Achievement

Table S3. Instrumental Variable (IV) Estimates Generated From Grade-Pooled Models Predicting Fourth- and Fifth-Grade Math Achievement—Additional Model Specifications

Table S4. Pooled Instrumental Variable (IV) Estimates—Subgroup Effects