







INTERVENTION, EVALUATION, AND POLICY STUDIES

The Impact of Every Classroom, Every Day on High School Student Achievement: Results From a School-Randomized Trial

Diane M. Early ^{a,*}, Juliette K. Berg ^{b,*}, Stacey Alicea ^{c,*}, Yajuan Si ^{d,*},
J. Lawrence Aber^c, Richard M. Ryan ^e, and Edward L. Deci ^e

ABSTRACT

Every Classroom, Every Day (ECED) is a set of instructional improvement interventions designed to increase student achievement in math and English/language arts (ELA). ECED includes three primary components: (a) systematic classroom observations by school leaders, (b) intensive professional development and support for math teachers and instructional leaders to reorganize math instruction, assessment, and grading around mastery of benchmarks, and (c) a structured literacy curriculum that supplements traditional English courses, with accompanying professional development and support for teachers surrounding its use. The present study is a two-year trial, conducted by independent researchers, which employed a school-randomized design and included 20 high schools (10 treatment; 10 control) in five districts in four states. The students were ethnically diverse and most were eligible for free or reduced-price lunch. Results provided evidence that ECED improved scores on standardized tests of math achievement, but not standardized tests of ELA achievement. Findings are discussed in terms of differences between math and ELA and of implications for future large-scale school-randomized trials.

KEYWORDS

instruction
school-randomized trial
high school
math achievement
English/language arts
achievement

High-quality classroom instruction is the core of effective schooling. Indeed, the National Research Council's Committee on Increasing High School Students' Engagement and Motivation to Learn (2004) argued forcefully that, although school-level policies and efforts to restructure schools may benefit students in myriad ways, student learning is most directly and deeply affected by how and what teachers teach. Every Classroom, Every Day (ECED) is an instructional improvement approach to increasing student achievement. This is the first report of the ECED school-randomized trial. ECED was designed and implemented by the Institute for Research and Reform in Education (IRRE) to systematically improve instruction, without employing a full

CONTACT Diane M. Early  diane_early@unc.edu  FPG Child Development Institute, University of North Carolina at Chapel Hill, CB #8180, Chapel Hill, NC, 27599-8180, USA.

^aUniversity of North Carolina–Chapel Hill, Chapel Hill, North Carolina, USA

^bAmerican Institutes for Research, Washington, DC, USA

^cRamapo for Children, New York, New York, USA

^dUniversity of Wisconsin–Madison, Madison, Wisconsin, USA

^eUniversity of Rochester, Rochester, New York, USA

*When this work was conducted Diane M. Early was at the University of Rochester, Juliette K. Berg and Stacey Alicea were at New York University, and Yajuan Si was at Columbia University.

comprehensive school reform model, which would typically alter broader aspects of schools, such as school climate, management, class and school size, instructional time, and/or parent involvement (Desimone, 2002). ECED is based on three key components of high-quality instruction that are linked to academic achievement (Early, Rogge, & Deci, 2014): (a) engagement of all students in their learning, (b) alignment of what is taught with state and national standards and high-stakes assessments, and (c) rigor in the content and methods of instruction. ECED's components are supported by the instructional reform and professional development literatures reviewed herein.

Engagement, Alignment, and Rigor as the Basis For High-Quality Instruction

Each component of the ECED intervention is designed to increase engagement, alignment, and/or rigor, collectively called EAR. ECED focuses on increasing student engagement because engagement has been consistently linked to student learning (National Research Council, 2004). Students high in intrinsic motivation are more likely to be engaged in learning, which may lead to deeper conceptual understanding of the material (Grolnick & Ryan, 1987). Further, when students have fully internalized the regulation of learning particular topics, they tend to be more engaged in learning and perform better than when learning is controlled by external or internal contingencies (e.g., Black & Deci, 2000; Grolnick & Ryan, 1989). Thus, intrinsic motivation and fully internalized extrinsic motivation are key predictors of engagement and positive educational outcomes (Ryan & Deci, 2000). Research has also shown that when teachers are supportive of students, interested in the material, and enthusiastic about teaching, students tend to be more engaged (e.g., Deci & Ryan, 1985).

ECED's second key focus is on alignment. In aligned classrooms, what is being taught and what students are being asked to do are in line with standards and curricula, are "on time" and "on target" with the scope and sequence of the course of study, and provide students with opportunities to experience high-stakes assessment methodologies (Connell & Broom, 2004). Polikoff (2012) referred to alignment of instruction with standards and assessments as the "key mediating variable separating the policy of standards-based reform to the outcome of improved student achievement" (p. 341). Porter (2002) pointed out that despite widespread understanding that students are most likely to learn what they are taught, instructional content and its alignment with standards and assessments have received little empirical attention, in part due to measurement difficulties. He posited that alignment is an important predictor of student achievement, and is crucial for accountability, ensuring that students are exposed to a logical progression of instruction, and monitoring reform and professional development efforts.

Rigor, as defined by IRRE, reflects the idea that students will only achieve high levels if such levels of work are expected and supported. Rigor reflects instructional strategies deployed by teachers to ensure that coursework provides optimal challenges for all students to move from where they are toward higher standards. ECED aims to increase rigor because the extant literature shows a strong connection between challenge and students' intrinsic motivation (e.g., Csikszentmihalyi, 1975; Danner & Lonky, 1981; Deci & Ryan, 1985; Harter, 1978), which has been shown to promote student achievement. Indeed, rigorous high school curriculum was linked to higher achievement and lower dropout, after controlling for

students' background and past performance (Lee & Burkam, 2003; Lee, Croninger, & Smith, 1997).

Instructional Reform in Middle and High Schools

ECED is in keeping with several recent meta-analyses of research regarding middle and high school instruction. Findings suggest broad-based reform models designed to improve instruction using multiple methods and extensive professional development result in better outcomes than models that narrowly target a single curriculum, technology, or instructional technique. For example, in a meta-analytic review of secondary mathematics programs, Slavin, Lake, and Groff (2009) found that effect sizes were greatest for instructional process programs that focused on changing teacher and student behaviors during daily lessons. Rakes, Valentine, McGatha, and Ronau (2010) found that among a variety of math interventions resulting in increased student achievement in algebra, interventions that focused on instructional strategies (i.e., cooperative learning, mastery learning, multiple representations, and assessment strategies) produced larger effect sizes than those focused on curricula or technology. Regarding ELA, Slavin, Cheung, Groff, and Lake (2008) likewise found that interventions designed to change daily teaching practices had substantially greater impacts on student reading comprehension than those focused on curriculum or technology alone.

Importantly, all three meta-analyses of instructional practice highlighted similar limitations in the literature (Rakes et al., 2010; Slavin et al., 2008; Slavin et al., 2009). First, the majority of interventions lasted a relatively short time (i.e., 12 weeks to 1 year). Second, many of the interventions and evaluations included small numbers of classrooms or students, often in a single school district. Third, most studies employed matching and randomization at the student level, rather than the school level, limiting the ability of evaluators to account for nested data structures, thereby prohibiting rigorous schoolwide assessment of intervention impacts. The current ECED school-randomized trial addresses these shortcomings by lasting two years, including schools in five districts with a large sample of students, randomizing at the school level, and employing an analytic strategy that accounted for the nested nature of the design.

Professional Development to Change Teacher Practices and Increase Student Achievement

Professional development, traditionally thought of as workshops, courses, and study groups for teachers, has recently been more broadly defined to encompass any activity aimed at improving instruction or teachers' skills and knowledge (Desimone, 2009). Using this broad definition, ECED is a professional development approach. Educational scholars have reached near consensus regarding critical features of effective professional development (Darling-Hammond, Wei, Andree, Richardson, & Orphanos, 2009; Desimone, 2009; Elmore, 2002; Garet, Porter, Desimone, Birman, & Yoon, 2001). Specifically, the form of professional development (e.g., workshop versus coaching) is less important than the extent to which it embodies five critical features. The same core ideas are espoused across the literature despite variation in nomenclature. Desimone's (2009) nomenclature is used here.

Effective professional development is *content focused*, meaning it extends and intensifies teacher knowledge of a subject area (e.g., math) and how children learn subject specific

content (Garet et al., 2001), rather than focusing on general pedagogy, abstract educational principals, or noncontent issues (e.g., team-building). Effective professional development uses *active learning* in which teachers engage with and analyze material through activities such as reviewing student work or discussing a videotaped lesson (Garet et al., 2001), as opposed to passively listening as material is presented. *Coherent* professional development is aligned with participating teachers' other professional development activities and is endorsed by the school and district leadership. High-quality professional development is of long enough *duration* for teachers to deeply explore new ideas, and it involves multiple sessions devoted to related concepts to allow teachers time to practice and receive feedback (Desimone, 2009; Garet et al., 2001). Last, high-quality professional development requires *collective participation* by all teachers from a department, school, or district, thereby promoting collaboration and allowing teachers to support one another in changing practice and sustaining that change through shared goals and language.

We have observed that schools and districts often undermine critical features of effective professional development by creating an array of opportunities from which teachers select those that appeal to them. Indeed, in a national study of professional development, Desimone, Porter, Garet, Suk Yoon, and Birman (2002) reported greater within-school variability in the quality of professional development received by teachers than between-school variability, indicating a lack of coherent schoolwide planning. Such a system may weaken the experiences of all teachers by preventing them from working toward a common set of goals over an extended period. As outlined below, ECED embodies each of the key components of effective professional development that together are posited to lead to changes in teacher practice and student learning.

Every Classroom, Every Day (ECED): The Intervention

Every Classroom, Every Day (ECED) provided ninth- and tenth-grade math and literacy teachers and instructional leaders with two years of intensive professional development and curricular support, using tools and processes developed by IRRE, with the ultimate goal of increasing student learning and achievement. ECED blends theory about engagement, alignment, and rigor with empirical work on effective school reform and professional development and is heavily informed by IRRE's extensive fieldwork in helping low-achieving schools improve student outcomes (e.g., Connell, Klem, Lacher, Leiderman, & Moore, 2009).

ECED has three major components: Use of the Engagement, Alignment, and Rigor (EAR) Classroom Visit Protocol, ECED Math, and Literacy Matters. The Method section provides details about the components and how variation in implementation was measured. In keeping with findings from the meta-analyses on school reforms (Rakes et al., 2010; Slavin et al., 2008; Slavin et al., 2009) and professional development literature (Desimone, 2009), ECED employs a broad range of strategies, including instructional coaches, content-focused professional development that encourages active participation, and curricular and assessment support to improve instruction. ECED teachers implement specific instructional strategies, including both small- and large-group instruction. All literacy and math teachers take part in the same activities, lasting two years. Coherence is created by a continual focus on engagement, alignment, and rigor.

ECED provides a literacy curriculum, but not a math curriculum. This is because the state of teaching and learning in these two instructional areas differs. What is taught in math

classes does not show as much variation across classrooms, schools, districts, and states as what is taught in ELA classes. Further, the alignment among curricula, standards, and standardized tests is typically greater for math than for English (Stotsky, 2005). Last, in mathematics, the process of teaching has been cited as the greatest shortfall, whereas in language arts both the content and the pedagogy have been implicated (Serdyukov & Hill, 2008). IRRE therefore focused the design of the mathematics intervention on the “how” of teaching mathematics, and focused the literacy intervention on “what” is being taught, as well as “how” it is being taught.

Past Research on ECED

Every Classroom, Every Day evolved out of the instructional improvement component of IRRE’s comprehensive school reform model called First Things First (FTF). Prior to the current study, no research had focused solely on ECED, but two quasi-experimental studies concluded that FTF was promising. In the first study, Gambone, Klem, Summers, Akey, and Sipe (2004) compared outcomes in Kansas City, Kansas—the first district to implement FTF—to those from all schools in other districts in the state. They found that the percentage of students scoring proficient or above on the high school math and reading tests went up across the three years of FTF implementation and the gap between Kansas City, Kansas, and other districts in the state diminished. In a second study of FTF, Quint, Bloom, Black, Stephens, and Akey (2005) used interrupted time-series analyses to investigate FTF in Kansas City, Kansas, and four districts from other states at varied stages of FTF implementation. Comparison schools were matched on preintervention test scores and student demographics. Findings indicated that academic outcomes in Kansas City, Kansas, high schools and middle schools improved substantially over those of comparison schools. The findings were inconclusive in the other four districts where FTF had been implemented for a shorter period of time.

ECED grew out of the lessons learned by IRRE during the original FTF evaluations (Gambone, Klem, Summers, & Akey, 2004; Quint et al., 2005). When those studies were conducted, the instructional improvement components of FTF were much less structured than those of the ECED intervention that was examined in the current study. For example, rather than having a set literacy curriculum or math benchmarking system as in ECED, the instructional improvement component of FTF included goals such as “Set high, clear, and fair academic and conduct standards” and “Provide enriched and diverse opportunities to learn, by making learning more authentic (active, cooperative, integrated and real-world based).” (Gambone et al., 2004, p. 18). The means of achieving the goals were identified and implemented by IRRE and instructional leaders from the schools as part of the reform agenda. Since completing those evaluations, IRRE has partnered with over 30 middle and high schools in seven states to improve student achievement. The practices specified in ECED result from the original FTF research, IRRE’s ongoing work at schools across the country, and IRRE’s continued focus on engagement, alignment, and rigor as means of increasing student achievement (Connell, Early, & Deci, 2014).

Although there are no published studies of ECED’s effects, Kansas City, Kansas, has been implementing most of ECED’s current math strategies for the past several years, in addition to the original FTF model. Since implementation of these revised strategies, the percentage of students meeting or exceeding the state proficiency standard on the tenth-grade math

achievement test increased between 27 and 41 percentage points across four schools over a four-year period (Connell, 2010).

Current Study

In order to evaluate the ECED approach to instructional improvement, a school-randomized trial was conducted in which 20 schools, stratified within five districts, were randomly assigned to receive either all ECED supports for two years ($j = 10$) or to a “business as usual” control group ($j = 10$). Although this was the first randomized test of ECED and such tests are typically efficacy trials, the current one is best characterized as an effectiveness trial. Implementation took place in “real-world” settings, rather than the tightly controlled settings that characterize an efficacy trial; the outcome measures were scores from standardized tests administered by the participating districts; and an intent-to-treat analytic strategy was employed.

The effectiveness design allowed tests of two main hypotheses: (a) ECED will increase student achievement in math and English language arts (ELA), as measured by standardized tests; and (b) the extent to which the ECED components are implemented as intended will be associated with greater increases in student math and ELA achievement.

Method

The ECED Approach

As noted earlier, ECED has three major components: Use of the EAR Classroom Visit Protocol, ECED Math, and Literacy Matters.

Use of the EAR Classroom Visit Protocol

Use of the EAR Protocol by instructional leaders—such as instructional coaches and school administrators—is a cornerstone of the ECED process. The EAR Protocol is a 15-item observational tool completed by trained observers following a 20-minute observation of an instructional session. It measures the extent to which students are actively and intellectually *engaged*; the extent to which learning materials, work expectations, and classwork are *aligned* with relevant standards and assessments; and the *rigor* of the material and expectations for student work (see Early et al., 2014 for details).

As part of ECED, trained instructional leaders are asked to make at least five EAR Protocol visits per week throughout the project and upload data to a secure server. The visits are designed to provide a structure for conducting classrooms observations that replaces the informal classroom visits and walk-throughs typically expected of instructional leaders. The data are used to generate reports about teaching and learning at different levels (e.g., teacher, department, school). The reports are used to inform school-level discussions about improving teaching and learning, conference calls between school leaders and IRRE consultants, and instructional coaches’ work with teachers. They also guide IRRE site visits and professional development sessions and provide a common language and lens for identifying high-quality instruction.

ECED Math

ECED’s second component—ECED Math—is based on the work of James Henderson and Dennis Chaconas in Kansas City, Kansas, in the late 1990s. ECED Math is not a curriculum; it is a system for delivering instruction and assessing student progress targeted to local, state, and national standards. ECED Math could be used in any math course; ECED treatment schools used it in Algebra 1 and Geometry.

IRRE consultants work with math teachers and coaches from all ECED schools within a district to identify key standards students must be able to demonstrate on high-stakes tests and to be successful at the next level of coursework. Teachers, coaches, and consultants work together to group those standards into meaningful sequences of skills, referred to as benchmarks. Each day’s instruction is focused on a specific benchmark, phrased in student-friendly terms called “I Can...” statements. After each unit, each student should be able to make an “I Can...” statement, such as “I Can solve quadratic equations.” Teachers work as teams—with support from IRRE consultants and local math coaches—to develop pacing guides to ensure that all benchmarks are addressed. Because the teams are made up of teachers from all ECED schools in a district, course content is similar among schools in a district, but might vary considerably between districts.

To check that all students have understood the benchmarks, teacher teams develop and administer five-question benchmark assessments. Students who do not pass or “master” a benchmark assessment are given additional support, followed by an alternate form of the assessment. Additionally, teacher teams develop capstone assessments, which integrate several related individual benchmarks into a coherent application of logically related concepts and skills. Students are graded solely on the number of benchmarks and capstones mastered. Those who have not passed enough benchmarks to attain a C at the end of the grading period receive an Incomplete (I) for the course. They have multiple opportunities, including tutoring and summer school, to change that I to a C or higher. If they do not succeed, the I is changed to an F.

IRRE intentionally designed ECED Math to focus on the three core instructional goals of EAR. “I Can...” statements make math engaging and personally relevant by showing students what skills they will acquire during each lesson. Each school or district creates its own pace and sequence based on local, state, and national standards, supporting alignment. Mastery grading promotes rigor by holding all students to the same high standards. The frequent benchmark assessments and the larger capstone assessments provide continual feedback and ensure that students are incorporating the new information into their larger base of mathematics knowledge.

Literacy Matters: ECED’s Literacy Component

The final component of ECED is the English/Language Arts (ELA) component, called Literacy Matters. Secondary students’ literacy skills—their ability to read, write, speak, and listen—form a fundamental building block for high school achievement and lifelong success. To address these pressing needs, ECED provides a research-based, structured literacy curriculum that uses authentic, real-world expository texts and engaging activities. This two-year curriculum is delivered in a required, stand-alone class that is in addition to the regular ninth- and tenth-grade English courses, doubling the amount of ELA exposure. The first year of the curriculum aims to strengthen students’ abilities to comprehend and gather information, helping them identify ways to make learning easier. The second year aims to

strengthen students' abilities to share and communicate information with others, helping them identify ways to express and personalize their knowledge. In both years, teachers use a set of interdisciplinary instructional strategies, called the Power 10, that equip students with transferable skills for comprehending, organizing, and remembering information.

Literacy Matters, like ECED Math, was designed with a focus on EAR. Making the material personally relevant is a well-established path to encourage engagement (National Research Council, 2004). To this end, Literacy Matters uses texts and assignments that address personal responsibility, positive societal change, and one's impact on the world. IRRE works with school districts and states to map the Literacy Matters curriculum onto state and local standards to ensure alignment. Rigor is pursued through appropriately challenging texts and the Power 10 strategies. Assessment rubrics provide ongoing information about mastery.

Supports for Changing Practice

ECED Math and Literacy Matters require substantial changes in the daily practice of teachers. Teachers are supported to make those changes through instructional coaching, weekly meetings to discuss emerging issues, summer workshops, and four annual IRRE site visits during which ECED teachers participate in half-day professional development sessions. Instructional coaches are employed by the school but trained by IRRE, starting in the summer before implementation. Coaches are supported via conference calls and IRRE's site visits, during which coaches make EAR Protocol visits with IRRE consultants, debrief about what they saw and what supports teachers need, discuss reflective coaching strategies, and plan how to best support each teacher.

Study Design

Twenty high schools (five districts, four schools per district) were assigned to either the treatment ($j = 10$) or control ($j = 10$) condition, using a stratified random approach in which two schools from each district were assigned to each condition. School recruitment began at the district level. To be considered for participation a district needed to include at least four high schools that each enrolled at least 220 ninth graders and where a minimum of 30% of students were eligible for free or reduced-price lunch (FRPL). The Common Core of Data (U.S. Department of Education, n.d.) was used to create a list of over 150 potentially eligible districts. Each was contacted via e-mail, regular post, and/or telephone. Interested districts participated in several phone calls with IRRE's leadership and the research team, followed by a site visit that provided extensive information about the intervention and research requirements. After the visit, interested districts signed a memorandum of understanding outlining implementation and research requirements, including the random assignment procedures.

The five participating districts were those that agreed to participate and were seen by IRRE and the research team as a good fit for this type of in-depth, two-year intervention and data-collection effort. They were spread across four states: two in California, one in Arizona, one in Tennessee, and one in New York. Schools in the first recruitment group ($j = 8$) participated in 2009–10 and 2010–11. Schools in the second recruitment group ($j = 12$) participated in 2010–11 and 2011–12. Treatment schools received all ECED supports, free of charge, although they were required to have some elements in place such as instructional

coaches and professional development time for teachers. Control-group schools were given a \$5,000-per-year honorarium to thank them for their participation in the data collection activities. They were asked only to continue with “business as usual,” using whatever instructional supports they had in place. There were no staffing or professional development requirements for control schools, but six of the 10 did employ instructional coaches and all had some professional development time.

As seen in Table 1, the participating schools were generally large, with an average enrollment of over 1,300 ($SD = 690$; $median = 1,151$), but the range of school sizes was also large (156 to 2,553). Only schools with over 220 ninth graders were initially recruited, but in the end five participating schools had fewer ($range = 156$ to 206), due largely to the open enrollment in several districts allowing students to attend any school in the district and making it difficult for districts to predict enrollment. Most schools were quite ethnically diverse. On

Table 1. Demographic characteristics of study schools and students.

Demographic characteristics	Overall	Treatment	Control
School characteristics^a	$j = 20$	$j = 10$	$j = 10$
Mean school enrollment	1,357	1,337	1,378
Mean % free/reduced-price lunch	70.4	65.3	75.5
Mean pupil/teacher ratio	18.3	18.3	18.3
Mean race/ethnicity			
% Hispanic	42.4	42.0	42.7
% Black, non-Hispanic	31.2	27.7	34.8
% White, non-Hispanic	16.7	20.9	15.4
% Asian/Pacific Islander	7.3	8.3	9.9
Student characteristics^b	$n = 8,250$	$n = 3,935$	$n = 4,315$
Grade in school (%)			
9th in Y1, 10th in Y2	75.2	76.3	74.2
9th both years	4.5	4.0	5.1
9th in Y1, not enrolled Y2	10.8	9.8	11.8
Not enrolled Y1, 10th in Y2	9.4	9.9	9.0
Race/ethnicity (%)			
Hispanic	50.9	49.7	52.0
Black, non-Hispanic	24.3	22.5	26.0
White, non-Hispanic	14.3	17.1	11.8
Asian/Pacific Islander	8.0	8.2	7.9
Other	2.5	2.6	2.3
Male (%)	52.7	53.1	52.3
Free/reduced-price lunch (either year%)	75.8	72.7	78.6
ELL (Y1%)	22.2	19.9	24.3
Special education (Y1%)	5.5	5.4	5.6
Age at baseline [Mean (SD) in years]	14.70 (.68)	14.69 (.66)	14.71 (.70)
Terms enrolled (%)			
1	16.6	16.5	16.7
2	17.7	17.2	18.1
3	11.4	10.1	12.5
4	54.4	56.2	52.7
Terms enrolled [Mean (SD)]	3.03 (1.18)	3.06 (1.18)	3.01 (1.17)
Baseline math score [Mean (SD)]	-0.10 (0.97)	-0.11 (0.94)	-0.09 (1.00)
Baseline ELA score [Mean (SD)]	-0.05 (0.96)	0.03 (0.97)	-0.12 (0.94)

^a School characteristics come from the U.S. Department of Education, National Center for Education Statistics (NCES) Common Core of Data (CCD), Public Elementary/Secondary School Universe Survey Data, 2009–10 and 2010–11. Values represent the first year the school participated in the study and include all students in the school. There were no statistically significant demographic differences between treatment and control schools.^b Student characteristics refer only to the students in the current study. Data come from school records and were 94% complete in each condition (excluding test scores). See the section entitled Student Achievement for a description of how the baseline test scores were calculated. There were no statistically significant baseline differences between students in control and treatment conditions on any variable when standard errors were adjusted for clustering of students within schools.

average they were 42% Hispanic, 31% Black, and 17% White, but again the range was large, with one school having no Hispanic students and another having only 2% White students. On average, 70% of students in these schools were eligible for FRPL (*range* = 46% to 98%; *median* = 69%).

All control group schools completed the two years of data collection, but two treatment schools stopped participating in ECED due to changes in leadership and low teacher commitment. One stopped after a single semester and one after the first year. The districts continued to provide student records, including test scores. With the intent-to-treat design, these two schools remained in all analyses, and missing data were imputed. This is an intentionally conservative approach that mirrors what might be expected in a typical implementation, but likely underestimates the effects that could be attained under more ideal circumstances.

Study Students

Because ECED was intended as a school-level intervention, the target population was almost all students who were in ninth grade in the study's first year and/or tenth grade in the second year. Over the two years, the 20 schools enrolled 8,786 such students. The only students who were excluded were: (a) those whose parents had returned a form indicating they did not want their children's school records released (184; 2.1%); (b) in a self-contained special education class (231, 2.6%), or (c) "newcomers" to the country with such limited English that they were excluded from the regular curriculum (121, 1.4%). These were small subgroups of the special education and English language learners (ELL) in the schools. After excluding these groups, the final student sample was 8,250 (3,935 in treatment schools and 4,315 in control schools).

Table 1 presents the demographic information for schools and study students. Although most of the students were in the ninth grade in Year 1 and the tenth grade in Year 2, a substantial minority was retained in the ninth grade or was only enrolled in a study school during one of the two study years. The sample was quite diverse with regard to race/ethnicity, and a large percentage of the students were from low-income families, as indicated by their FRPL eligibility.

Table 1 also indicates how many terms students were enrolled in study schools. For students in treatment schools to have the full benefit of ECED they would need to be enrolled for all four terms of the project. Just over half of students were enrolled all four terms, with the remainder arriving after the first semester of their school's participation, leaving prior to the last semester, or both. This high level of mobility means a large portion of the students did not receive the full treatment, potentially lowering the intervention's impact.

Following What Works Clearinghouse (WWC, 2014) guidelines calling for establishment of baseline equivalence, students with valid outcome scores in the treatment versus control groups were compared on all characteristics presented in **Table 1**, after adjusting the standard errors for clustering of students within schools. No statistically significant differences were found. Moreover, there were no statistically significant demographic differences at the school level, suggesting that randomization successfully produced equal groups of schools at baseline.

Table 2. Testing systems and test included in analyses, by district.

District (State)	Stakes	n	Baseline Math	Baseline ELA ^a	Year 1 Math	Year 1 ELA	Year 2 Math	Year 2 ELA
1 (CA)	Baseline, Y1 & Y2: CST ^b ; used to calculate AYP ^c ; no individual student stakes	2,615	<ul style="list-style-type: none"> 73% 7th math 4% 8th math 5% Alg 1 17% missing 	<ul style="list-style-type: none"> 70% 7th & 8th ELA 3% 7th ELA 9% 8th ELA 17% missing 	<ul style="list-style-type: none"> 60% Alg 1 17% Geom 22% missing 	<ul style="list-style-type: none"> 82% 9th ELA 18% missing 	<ul style="list-style-type: none"> 24% Alg 1 33% Geom 13% Alg 2 29% missing 	<ul style="list-style-type: none"> 75% 10th ELA 25% missing
2 (CA)	Baseline, Y1 & Y2: CST; used to calculate AYP; no individual student stakes	1,512	<ul style="list-style-type: none"> 68% 7th math 5% 8th math 8% Alg 1 19% missing 	<ul style="list-style-type: none"> 73% 7th & 8th ELA 3% 7th ELA 6% 8th ELA 19% missing 	<ul style="list-style-type: none"> 57% Alg 1 18% Geom 4% Alg 2 20% missing 	<ul style="list-style-type: none"> 80% 9th ELA 20% missing 	<ul style="list-style-type: none"> 15% Alg 1 33% Geom 17% Alg 2 8% H.S. math 23% missing 	<ul style="list-style-type: none"> 80% 10th ELA 20% missing
3 (TN)	Baseline: TCAP; used to calculate AYP; Y1 & Y2: TCAP EOC ^d ; used to calculate AYP and contribute to student course grades	1,217	<ul style="list-style-type: none"> 59% 7th math 14% 8th math 2% Alg 1 26% missing 	<ul style="list-style-type: none"> 2% 7th rdg 54% 7th rdg & 8th rdg & wrt 2% 7th rdg & 8th wrt 13% 8th rdg & wrt 26% missing 	<ul style="list-style-type: none"> 53% Alg 1 6% Geom 41% missing 	<ul style="list-style-type: none"> 66% Eng 1 4% Eng 2 30% missing 	<ul style="list-style-type: none"> 13% Alg 1 39% Geom^e 11% Alg 2 37% missing 	<ul style="list-style-type: none"> 5% Eng 1 56% Eng 2 40% missing
4 (AZ)	Baseline: AIMS ^f ; used to calculate AYP; Y1: Stanford 10; low stakes; Y2: AIMS Exit; students must eventually pass to graduate	2,181	<ul style="list-style-type: none"> 28% 7th math 33% 8th math 39% missing 	<ul style="list-style-type: none"> 27% 7th rdg & wrt 33% 8th rdg 39% missing 	<ul style="list-style-type: none"> 70% math 30% missing 	<ul style="list-style-type: none"> 68% rdg & lang 2% rdg 30% missing 	<ul style="list-style-type: none"> 73% math 28% missing 	<ul style="list-style-type: none"> 74% rdg & lang 26% missing
5 (NY)	Baseline: NYSTP ^g ; used to calculate AYP; Y1: Regents; students must pass to graduate; Y2: Regents & GMRT ^h	725	<ul style="list-style-type: none"> 50% 7th math 20% 8th math 2% Alg 1 28% missing 	<ul style="list-style-type: none"> 46% 7th & 8th ELA 20% 8th ELA 4% 7th ELA 30% missing 	<ul style="list-style-type: none"> 51% Alg 1 3% Geom 46% missing 	<ul style="list-style-type: none"> 100% missing 16% Geom 	<ul style="list-style-type: none"> 31% Alg 1 2% Alg 2 50% missing 	<ul style="list-style-type: none"> 27% GMRT 5% ELA 68% missing

^aWhen more than one test (e.g., reading and writing) was administered in 7th and/or 8th, scores from a single year were first averaged together, then that value was averaged with the other year, to avoid giving extra weight to a single year. ^bCalifornia Standards Test. ^cAdequate yearly progress, as required by No Child Left Behind. ^dTennessee Comprehensive Assessment Program and Tennessee Comprehensive Assessment Program End of Course. ^eTCAP EOC does not include a Geometry test; scores from a districtwide geometry test were used instead. ^fArizona Instrument to Measure Standards. ^gNew York State Testing Program. ^hGates-MacGinitie Reading tests; added by the study because 10th graders in NY do not typically take an ELA exam.

Measures

Student Achievement

School-administered standardized tests are the primary way schools measure student progress, thus math and ELA achievement as measured by standardized tests was this trial's primary outcome of interest. The districts were spread across four states, each with its own testing system. In order to mirror real-world conditions in which already overburdened schools do not have resources to administer additional tests, existing tests were used rather than adding a common test across districts. Additional testing would have likely harmed district recruitment efforts, would not have been feasible in most school calendars, and would be less relevant to schools than the school-administered tests. In order to include the tests from different districts in the same models, scores had to be combined into comparable variables indicating student performance, relative to peers, in each subject. This involved overcoming several challenges: there were four different sets of standardized tests; different students within the same district and grade often took different tests depending on their course enrollment; some students took more than one math or ELA test during a single year; and some districts did not routinely administer standardized tests each year in both math and ELA.

Table 2 summarizes the testing system in each state and indicates which tests were included in the scores used in the analyses. Six test scores were calculated for each student: baseline math, Year 1 math, Year 2 math, baseline ELA, Year 1 ELA, and Year 2 ELA. The rules used to combine different tests into these six variables are detailed below. It is important to note that because random assignment took place within district, the rules for combining test scores were applied identically to treatment and control schools. Thus, this system for combining test scores poses little threat to internal validity. All decisions regarding what tests to include and how to combine test scores were made prior to conducting any analyses.

Math Scores. The general rules for combining math test scores onto a common scale were: (a) standardize each test within test subject (e.g., eighth-grade math, Algebra 1) and district, but across administration years (after confirming there had been no major changes across the study years), and (b) when students had more than one score at baseline, Year 1, or Year 2, use the lowest-level test (e.g., if a student had Algebra 1 and Geometry scores, use Algebra 1 score). Additionally, a control variable was created to indicate the level of the test (e.g., Algebra/ninth grade, Geometry/tenth grade) used for each student's math score.

The decision about how to treat math scores at the end of Years 1 and 2 stemmed from three interrelated concerns. First, there were some tests that only a very few students took, such as Calculus, causing concern about standardizing scores within the test. Nonetheless, within-test standardized scores were retained for such tests because omitting them would typically have resulted in missing data for those students, which would have meant using imputed scores for cases where actual data had been provided. Second, it is possible that taking certain math courses—and therefore taking certain math tests—was influenced by the intervention itself. If, for instance, instruction improved in ECED schools in Year 1, then students in those schools might have taken more advanced math courses and tests in Year 2. However, their scores on the more advanced tests might be lower than they would have been had they taken a lower level test. Thus, the intervention could lower scores by increasing advanced math course enrollment. This concern was addressed by reviewing test-taking

in each district and finding no clear pattern across treatment and control schools. In some districts, students took higher tests in treatment schools, in some the opposite was true, and in some there was no between-group difference. Third, combining different level math tests could introduce error because a student might have received a higher score if she or he had taken a lower level test. This concern was partially addressed by selecting the score from the lowest-level math test for students who took more than one in a given year. However, this still means that for different students, different levels of tests have been combined into the same variable, after standardizing within test and district.

The approach for combining math baseline scores was the same as for Year 1 and 2 scores, but the rationale was slightly different. Often there was a score for the same student in both seventh and eighth grade. In those cases, the seventh-grade score was used for baseline. This was to minimize the range of tests included. In all districts, seventh graders took a seventh-grade math test, but in many districts the math test taken in eighth grade depended on the course the student took, with a relatively large group taking the Algebra 1 test. Eighth-grade students enrolled in Algebra were likely more advanced than those enrolled in eighth-grade math, but because they took a harder test, they may actually have scored lower. In order to minimize the number of different tests being combined into a single score, the lowest math tests at baseline were used, so the baseline score often came from the seventh grade (see Table 2). When there was no seventh-grade score, the eighth-grade score was used, so the baseline variable included different tests for different students, thereby minimizing, but not eliminating, the concern.

To check baseline equivalence, after creating these combined scores, baseline math scores in treatment versus control schools were compared, after adjusting for clustering of students within school, using only students who had outcome scores (WWC, 2014). No treatment versus control differences were found. Likewise, average school-level math baseline scores did not vary between conditions.

ELA Scores. As with math, the goal for ELA was to obtain a single score for each student at baseline, Year 1, and Year 2. Most students had only one ELA test in Year 1 and one ELA test in Year 2. A few students, however, had more than one. When the two tests were at different levels (e.g., ninth and tenth grade), the one that matched the students' grade in school that year was used, to avoid combining tests within grade cohort. When the two tests were at the same level (e.g., ninth-grade reading and ninth-grade language), the mean of the two scores was used, to represent the broadest conceptualization of ELA and minimize error.

At baseline, within a district, most eighth-grade students took the same ELA test(s) and the level of the test(s) was not linked to the student's course-taking or past ELA achievement. Thus, unlike math, averaging across years would likely result in the least error. When students had both a seventh- and eighth-grade ELA score, or a seventh-grade reading and a seventh-grade writing score, each was first standardized within test and district and then the standardized values were averaged together. In a small percentage of cases (8%), a district provided three test scores (seventh-grade reading, eighth-grade reading, and eighth-grade writing). In those cases, in order to weight seventh and eighth grade equally, each test was standardized within test and district, the two eighth-grade tests were averaged together, and then that value was averaged with the seventh-grade test.

After computing these combined scores, baseline ELA scores were compared for students in treatment versus control schools, after adjusting for clustering of students within school, using only students who had outcome scores (WWC, 2014). No treatment versus control differences were found, and average ELA baseline scores at the school level did not vary by condition.

Variation in Implementation

As with any intervention, schools in the treatment condition varied with regard to how faithfully they implemented the ECED components, and schools in the control condition varied in the extent to which similar types of supports were in place. Because the full ECED intervention had never been implemented prior to this study, there was no preexisting measure of implementation fidelity. In order to include this variation in nonexperimental analyses, the independent research team created a system for measuring this variation as implementation began. There were four major steps involved: (a) creating indicators and operational definitions to describe full implementation; (b) gathering data from multiple sources, including key-informant interviews, and linking them to operational definitions, (c) reliably coding the key-informant interviews, and (d) combining all information to create final scores. These steps are similar to the first four steps advocated by Hulleman, Rimm-Kaufman, and Abry (2013), although data collection in the present study was less structured. Their fifth and final step—linking the measure of implementation to outcomes—is addressed in the results.

To create indicators and operational definitions, IRRE senior staff worked with the ECED research team to create a list of the specific activities that would define “full implementation.” The research team then identified ways to measure the 30 indicators they identified, using a scoring rubric that combined multiple sources of information. Most of the information came from semi-structured, open-ended interviews conducted each spring with instructional coaches (or department chairs in the four control schools without instructional coaches) and principals or assistant principals at each participating school. The interviews were conducted by members of the research team who had little knowledge of the implementation’s success. Two individuals worked independently to reduce each interview into a series of very brief (i.e., yes/no) responses that directly addressed the full implementation activities. They compared their responses regularly, ensuring over 90% agreement. Next, one of those two individuals transformed the brief responses into numeric scores using the scoring rubric created by the research team. As a check, the research project director completed two rounds of scoring, each time scoring 10% of the responses. In the first round, the project director’s codes agreed 81% of the time. After some discussion of coding rules and inconsistencies, the research project director coded an additional 10% and the codes matched 92% of the time. In addition to the interviews, some of the indicators were scored using information from teacher questionnaires, student records, and data maintained by IRRE. The same sources of information and coding systems were used for both treatment and control schools, making the scores directly comparable. Coders were not, however, blind to the school’s treatment condition because the responses from individuals at treatment schools often referred to the ECED supports.

Once each indicator had been scored, the scores were combined using weights devised by IRRE senior staff. Variation-in-implementation scores had a potential range of 0 to 100. The average final score across the 20 schools was 38.06 ($SD = 29.60$). For the 10 treatment schools the average was 65.20 ($SD = 14.42$, $range = 38.74$ to 78.34), with the mean score for the two schools that stopped implementation during the project being much lower than for the eight that continued

implementation (39.42 vs. 71.64). The average final score for the 10 control schools was 10.92 ($SD = 2.15$, $range = 8.72$ to 14.52), markedly lower than the treatment schools.

Attrition and Imputation of Missing Data

According to WWC (2014) “attrition occurs when an outcome variable is not available for all participants initially assigned to the intervention and comparison groups” (p. 11). For cluster randomized-control trials, such as the current study, it is important to have low attrition at both the cluster (i.e., school) and subcluster (i.e., student) levels. Further, it is important that differential attrition in the treatment versus control conditions not be too high.

In the current study, there was no cluster-level attrition. Two treatment schools stopped receiving the supports before the study was finished; however, the districts did provide test scores for students in those schools. At the student level, after creating math and ELA scores as described previously, overall attrition for the four key outcomes was 30% for Year 1 math and 32% for Year 2 math, Year 1 ELA, and Year 2 ELA. According to WWC (2014), the conservative boundary for low differential attrition is 4.1% when overall attrition is 30% and 3.8% when overall attrition is 32%. In the current study, the largest differential attrition was for Year 2 Math where it was 2%. Thus, according to WWC, student-level attrition in this study was low. Much of the missing data resulted from the fact that some students were not enrolled (i.e., had left the school or had not yet enrolled) and therefore had not taken the assessment. Additionally, as noted on Table 2, there was considerable missing data in District 5 due to their testing system.

To handle missing values, multiple imputation, which generates multiple complete data sets and mitigates the uncertainties introduced by missing data, was employed (Rubin, 1987). A two-step imputation procedure was used in which first student demographic and questionnaire data¹ were imputed using the latent class approach (Si & Reiter, 2013), which can flexibly and efficiently deal with a large number of categorical variables with complex dependency structures. This approach assumes that the students are divided into several latent classes. Within each class, the variables are conditionally independent, meaning the variables are independent and have the same distributions. The number of classes and class assignment is determined by data. The missing values are then imputed within each class. School, district, and treatment condition were included as background variables, along with the following student demographic variables: grade, gender, ethnicity, age at baseline, ELL, special education, FRPL. Five completed data sets were generated.

For the second step, the R package “mi” (Su, Gelman, Hill, & Yajima, n.d.) was used to impute students’ test scores. Imputed demographic information from one randomly chosen data set of the five generated above were used as covariates, treating them as categorical. An interaction term between state and treatment status was also included as a covariate. Again, five multiply-imputed data sets were created. Values were imputed for all eligible students (i.e., all students except those in self-contained special education, newcomers, or parent refusals).

These missing data were mainly due to item nonresponse and design issues. The current imputation generated plausible results assuming that data are missing at

¹Student questionnaires regarding students’ experiences in school were collected as part of this project but are not included in the current analyses.

random (MAR) conditioned on all observed information, including baseline demographic covariates and previous test scores. MAR is an inherently untestable assumption because it requires that missingness not depend on the outcome, yet by definition the outcome is not observed when it is missing. Indirect testing of the MAR assumption would have been difficult because of the large number of missing items and the complex data structure that included repeated measures and nesting. As described in the Results section, as an additional check, the main impact analyses using unimputed data were conducted and similar results were obtained, bolstering confidence that the imputed findings were not biased.

Unconditional Models

Variance in the outcomes was partitioned into within-school and between-school components by fitting an unconditional two-level model with no predictors (Bryk & Raudenbush, 1992), and intraclass correlation coefficients (ICC) and ICC(2) were calculated for each outcome. ICC is a measure of the ratio of the variance that lies between schools to the total variance. For math and ELA achievement, the ICCs ranged from .06 to .09 across the study's two years, indicating modest between-school variance. The magnitude of these ICCs resembles the average ICC of .08 for low-achieving schools, based on nationally representative samples across grades K–12 (Hedges & Hedberg, 2007). ICC(2) is an estimate of the reliability of the group-mean rating that takes group sample size into account.² An ICC(2) between .70 and .85 is considered acceptable (Ludtke, Trautwein, Kunter, & Baumert, 2006). Here, the ICC(2)s for both ELA and math achievement were .97 in both years, indicating a high level of reliability.

Data Analytic Strategy

There were two main outcomes of interest: (a) students' math achievement and (b) students' ELA achievement. For each type of outcome, two main types of analyses were conducted: (a) intent-to-treat analyses of the impact of ECED and (b) nonexperimental analyses of variation in implementation. For each of these four types of analyses, a series of two-level hierarchical linear models (HLM 6.02; Raudenbush & Bryk, 2002) with random intercepts was estimated. The models accounted for nesting of students within schools. In all models, maximum likelihood parameter estimates were used to estimate the parameters. All covariates were grand-mean-centered, following guidelines by Enders and Tofghi (2007) for cluster-randomized studies where a Level 2 treatment effect is of interest. In interpreting the results we consider an alpha level of $p < .05$ as statistically significant. Effects up to the .10 level are noted as marginally significant because the design resulted in relatively low power to estimate the intervention effects (i.e., only 14 df), particularly in the case of interactions (McClelland & Judd, 1993). Effect sizes were calculated by dividing the estimate of the intervention effect by the raw standard deviation of the dependent variable for the control group (Hedges' g).

² $ICC(2) = k \times ICC(1)/1 + (k - 1) \times ICC(1)$, with k being the average group sample size.

Intent-to-Treat

The intent-to-treat impact analyses considered the impact of ECED on outcomes at the end of Year 1 and the end of Year 2, using imputed data. (Note that growth curves were not estimated because there were only three data points). Year 1 analyses included all study students who were in ninth grade and were enrolled in a target school at any point during the first year of the study ($n = 7,184$). The Year 2 analyses included all study students who were enrolled in a target school during either year of the study and were in ninth grade in the first year and/or tenth grade in the second year ($n = 8,250$).

Model 1 included student baseline test score as a covariate at Level 1, school-level treatment condition, and four dummy codes for the five school districts at Level 2. In Model 2, student baseline demographic covariates (i.e., gender, race/ethnicity, FRPL, special education, receipt of ELL services) were added at Level 1. In Model 3, a variable indicating the number of semesters the student was enrolled in a study school ($range = 1$ to 4) was added at Level 1 to account for variation in students' potential exposure to the treatment. For the math analyses, in Model 4 a variable indicating the level of math test taken (e.g., Algebra 1, Geometry) was added to account for the fact that different districts administered different tests, and test level often depended on students' course schedules. It is important to note that these last two models are not exogenous because enrollment and test-taking could have been affected by the intervention. Thus, Models 3 and 4 were considered nonexperimental and were meant to complement the main impact analyses. Finally, to see if the treatment differentially impacted different students, moderation effects of gender, race/ethnicity, FRPL, special education, receipt of ELL services, baseline test score, number of semesters in study schools, and math test taken (in math models only) were tested by including cross-level interactions between the covariates and treatment.

Variation in ECED Implementation

In order to test the degree to which variation in implementation affected the impact of the intervention, follow-up analyses were conducted in which Models 1 through 4 were examined, replacing the treatment/control variable with the overall variation in implementation score at Level 2. Because variation in implementation was not randomly assigned, these analyses are nonexperimental and are meant to complement the experimental results. In the interest of parsimony, only the main effects were tested.

Results

ECED's Impact on Students' Achievement

In Model 1, students in treatment schools had higher math test scores than their counterparts in control schools. This finding was marginally significant in Year 1 ($b = .18$, $SE = .09$, $p = .053$, $ES = .18$) and statistically significant in Year 2 ($b = .15$, $SE = .07$, $p = .036$, $ES = .16$). When the demographic covariates were added to the models, once again the effect of treatment was marginally significant at Year 1 and statistically significant at Year 2 (see Table 3, Year 1 $ES = .17$; Year 2 $ES = .14$).

Cross-level interactions between treatment status and baseline demographic covariates were largely not statistically significant. There was a positive, marginally significant interaction between treatment and baseline score at the end of Year 1 ($b = .04$, $SE = .02$, $p = .051$) such that students who

Table 3. Impact analyses predicting math and ELA achievement.

	Math achievement						ELA achievement					
	Year 1 (n = 7,184)			Year 2 (n = 8,250)			Year 1 (n = 7,184)			Year 2 (n = 8,250)		
	Estimate	SE	p	Estimate	SE	p	Estimate	SE	p	Estimate	SE	p
Condition	0.164	0.080	0.060	0.133	0.061	0.048	0.035	0.126	0.788	0.063	0.052	0.248
Covariates												
District 2	0.124	0.122	0.330	-0.069	0.089	0.453	0.044	0.123	0.723	-0.024	0.083	0.773
District 3	0.744	0.129	0.200	0.181	0.091	0.065	0.017	0.129	0.898	-0.130	0.083	0.147
District 4	0.249	0.125	0.065	0.390	0.092	0.001	-0.054	0.123	0.668	0.009	0.085	0.916
District 5	0.205	0.130	0.137	0.332	0.097	0.004	0.249	0.319	0.471	-0.147	0.105	0.186
Baseline	0.521	0.015	0.000	0.395	0.015	0.000	0.676	0.011	0.000	0.608	0.021	0.000
Gender	-0.005	0.020	0.786	-0.071	0.020	0.001	0.023	0.019	0.237	-0.009	0.023	0.691
Hispanic	-0.153	0.040	0.001	-0.136	0.047	0.013	-0.099	0.041	0.028	-0.170	0.050	0.009
Black	-0.195	0.039	0.000	-0.243	0.050	0.000	-0.156	0.052	0.016	-0.151	0.049	0.012
Asian/Pac. Isl.	0.005	0.052	0.921	0.047	0.071	0.528	-0.004	0.055	0.939	-0.030	0.049	0.548
Other ethnicity	-0.212	0.080	0.011	-0.115	0.081	0.168	-0.075	0.081	0.366	-0.175	0.087	0.065
FRPL	-0.003	0.037	0.932	-0.009	0.032	0.795	-0.041	0.031	0.199	-0.101	0.026	0.000
Spec. ed.	-0.314	0.052	0.000	-0.229	0.057	0.001	-0.291	0.051	0.000	-0.173	0.071	0.040
ELL	-0.054	0.031	0.086	-0.021	0.029	0.463	-0.090	0.028	0.002	-0.047	0.033	0.170

Note. Unstandardized estimates shown. Students were in 9th grade in Year 1. Most were in 10th grade in Year 2, but some were still in 9th grade. Condition: 0 = control, 1 = treatment; Districts 2–5 are compared to District 1; Gender: 0 = male, 1 = female; each race/ethnicity group is compared to White students; Free/reduced-price lunch: 0 = received neither year of the study, 1 = received one or both years of the study; special education and English language learner: 0 = service not received during Year 1, 1 = service received during Year 1. Effect sizes: Math Year 1: .17; Math Year 2: .14, ELA Year 1: .04, ELA Year 2: .06.

started off with higher scores performed slightly better at the end of the school year when they were in treatment schools compared to when they were in control schools.

The subsequent nonexperimental models indicated that the number of semesters enrolled in a study school was not a statistically significant predictor of math achievement (Model 3), nor was the level of math test statistically significantly associated with math 2 achievement (Model 4). However, there was a statistically significant negative interaction between treatment and level of math test at the end of Year 2 ($b = -.08$, $SE = .03$, $p = .034$) such that students who took a more advanced math test performed better when they were in the control schools compared to when they were in the treatment schools.

There was no effect of ECED treatment on students' ELA scores at the end of Year 1 or Year 2, with or without controlling for the demographic characteristics (see Table 2). Model 3 indicated that the number of semesters enrolled in a study school was not a significant predictor of ELA achievement. The cross-level interactions between treatment and covariates were not statistically significant, indicating that ECED did not affect ELA scores in any subgroup.

Variation in ECED Implementation

When the variation-in-implementation variable replaced the treatment-condition variable in Model 1, there was a significant association between level of implementation and math achievement in Year 1 ($b = .40$, $SE = .14$, $p = .013$, $ES = .40$) and Year 2 ($b = .26$, $SE = .11$, $p = .038$, $ES = .28$). Students in schools that implemented ECED-like activities to a greater extent had higher math achievement. As seen in Table 4, when student baseline demographic covariates were added (Model 2), the relationship between level of implementation and math achievement remained significant in Year 1

Table 4. Variation in implementation predicting math and ELA achievement.

	Math achievement						ELA achievement					
	Year 1 (n = 7,184)			Year 2 (n = 8,250)			Year 1 (n = 7,184)			Year 2 (n = 8,250)		
	Estimate	SE	p	Estimate	SE	p	Estimate	SE	p	Estimate	SE	p
Variation in Implementation	0.365	0.129	0.014	0.225	0.108	0.055	0.013	0.196	0.950	0.081	0.096	0.412
Covariates												
District 2	0.095	0.114	0.419	-0.087	0.090	0.353	0.043	0.126	0.737	-0.031	0.084	0.717
District 3	0.143	0.122	0.261	0.163	0.092	0.099	0.016	0.130	0.906	-0.137	0.086	0.134
District 4	0.206	0.119	0.104	0.364	0.095	0.002	-0.055	0.127	0.668	-0.001	0.086	0.995
District 5	0.212	0.122	0.104	0.337	0.098	0.004	0.249	0.319	0.471	-0.145	0.106	0.194
Baseline	0.521	0.015	0.000	0.396	0.015	0.000	0.676	0.011	0.000	0.607	0.021	0.000
Gender	-0.005	0.020	0.786	-0.071	0.020	0.001	0.023	0.019	0.237	-0.009	0.023	0.690
Hispanic	-0.153	0.040	0.001	-0.136	0.047	0.013	-0.099	0.041	0.027	-0.171	0.050	0.009
Black	-0.195	0.039	0.000	-0.243	0.050	0.000	-0.156	0.052	0.015	-0.151	0.049	0.012
Asian/Pac. Isl.	0.004	0.052	0.934	0.046	0.071	0.531	-0.005	0.055	0.936	-0.030	0.049	0.543
Other ethnicity	-0.213	0.080	0.010	-0.115	0.081	0.168	-0.075	0.081	0.365	-0.175	0.087	0.064
FRPL	-0.003	0.037	0.937	-0.008	0.032	0.799	-0.041	0.031	0.198	-0.101	0.026	0.000
Spec. ed.	-0.314	0.052	0.000	-0.229	0.057	0.001	-0.291	0.051	0.000	-0.173	0.071	0.041
ELL	-0.054	0.031	0.086	-0.021	0.029	0.457	-0.090	0.028	0.002	-0.048	0.033	0.169

Note. Unstandardized estimates shown. Students were in 9th grade in Year 1. Most were in 10th grade in Year 2, but some were still in 9th grade. Districts 2–5 are compared to District 1; Gender: 0 = male, 1 = female; each race/ethnicity group is compared to White students; Free/reduced-price lunch: 0 = received neither year of the study, 1 = received one or both years of the study; special education and English language learner: 0 = service not received during Year 1, 1 = service received during Year 1. Effect sizes: Math Year 1: .37; Math Year 2: .24; ELA Year 1: .01; ELA Year 2: .08.

($ES = .37$) and was marginally significant in Year 2 ($ES = .24$). Neither the addition of the math test level nor the variable indicating the number of semesters students spent in a study school altered the findings in either model.

Variation in implementation was not statistically significantly related to students' ELA scores in Year 1 (Model 1: $b = .03$, $SE = .20$, $p = .868$; Model 2: $b = .01$, $SE = .20$, $p = .950$.) or in Year 2 (Model 1: $b = .11$, $SE = .10$, $p = .287$; Model 2: $b = .08$, $SE = .10$, $p = .412$). In follow-up models, adding the indicator variable for number of semesters spent in a study school did not alter the findings.

Robustness Checks

Two sets of checks were conducted to ensure that the findings were robust with regard to the missing data and multiple imputation strategies. First, Models 1 and 2 were repeated using the unimputed data. The pattern of results was largely the same. A table presenting those findings appears in the Appendix. Additionally, the analyses were repeated, using the imputed data but excluding the district that had the highest amount of missing data (District 5). Again, the pattern of findings was largely the same.

Discussion

The current trial examined Every Classroom, Every Day, an instructional improvement approach designed by the Institute for Research and Reform in Education. It is one of the few randomized field trials in the area of educational reform that has involved multiple school districts and high school-level randomization. The experiment was longitudinal and

involved implementation and data collection over two consecutive school years. The analyses used a multilevel design accounting for students nested within schools.

Summary of Findings

Using a conservative, intent-to-treat approach, these findings provide evidence that ECED improved student achievement in math. After two years of implementation, students in treatment schools scored statistically significantly higher on standardized tests of math than did students in control schools, controlling for preintervention math achievement, school district, and student demographics. In general, the effect of ECED Math on achievement was the same across subgroups. The effects of math were slightly smaller after two years of implementation than they were after the first year, an unusual finding likely due to two of the treatment schools opting out of the intervention prior to the second year.

In contrast, ECED did not affect students' ELA scores in either year of the study, and there was no evidence that it affected ELA achievement for any subgroup. There was evidence that fuller implementation of ECED was linked to higher math achievement, but there was no evidence of a link between ECED implementation and ELA achievement.

Limitations

The ECED school-randomized trial experienced a number of challenges in retention, implementation, and data collection. Of the 10 treatment schools, two dropped out before the second year of implementation, dramatically weakening the intervention. Using a conservative approach, the current analyses included those two schools with missing data imputed. The association between variation in implementation and student achievement was tested, but even these nonexperimental analyses likely underestimated how severely this cessation of participation affected the impact of ECED. This intent-to-treat approach strengthens the study's ability to draw conclusions about ECED's likely impact in the real world of large-scale school reform, but it may underestimate ECED's potential to improve student outcomes in schools that implement with high fidelity.

Additionally, the project suffered from considerable missing data resulting from imperfect record keeping in the school districts, high student mobility, and one district that did not routinely administer ELA tests in ninth or tenth grade. Missing data was addressed primarily through multiple-imputation. The main impact analyses were repeated in two steps: first using unimputed data and then excluding the district with the highest level of missing data. The findings were essentially the same in the two analyses.

The difficulty in combining test scores across state systems and across content is another study limitation. The fact that schools were randomized to treatment or control condition within districts means that test scores were treated identically in intervention and control schools, protecting the study's internal validity. External validity was strengthened by conducting the study in four different states across the country. However, each state had its own testing schedule and system, necessitating the combination of scores across systems. Systems may have been testing different types of material, or instruction may not have been equally linked to all tests. Further, within states, different students took different tests depending on their course enrollments. Students in more advanced courses are often given more advanced tests; it is not possible to know how those students would have scored had they been given

the less advanced tests taken by their peers. This is the nature of high school testing, but it poses a problem for researchers looking for a common metric. As outlined in the methods, steps were taken to mitigate this challenge, but some findings, particularly nonsignificant ones, may be related to less-than-perfect outcome measures.

The lack of a preexisting fidelity measure was an additional limitation. Because this exact set of supports had never been implemented prior to this study, fidelity had not been measured previously. Instead, the research team worked with the intervention developers to create interviews and coding-rubrics after the intervention was underway. Further, the nature of the interviews meant that the coding team could not be blind to a school's treatment condition. Thus, the rigor of the measurement of variation in implementation was less than would be ideal.

Finally, the participating schools were all fairly large and situated in districts that included at least four high schools. Some of the supports required by ECED (e.g., instruction coaches, professional development time) were already in place in most of the participating schools. Smaller and more isolated schools might have few resources, making ECED implementation more difficult. Thus, the current results may not generalize to such schools.

Interpreting Effect Sizes

The effect sizes for the math impacts, ranging from .14 to .17, do not reach the WWC (2014) definition of substantive (.25), but may be meaningful when put in the context of the intervention, grade level, and outcome under study. Hill, Bloom, Black, and Lipsey (2008) suggested several strategies for interpreting effect sizes of an educational intervention. One strategy is to compare effects of the intervention to expectations for growth in a typical year without an intervention. Hill and colleagues present average annual gains in effects sizes from nationally normed tests, which are relevant here because the tests used were normed on large or national populations. Using differences in mean scale scores for students in adjacent grades and converting them to standardized effect sizes, they found that the mean annual gain in effect size for math from grades 8 to 9 was .22 with a margin of error of plus or minus .10. The mean annual gain in effect size from grades 9 to 10 was .25 with a margin of error of plus or minus .07. When using those values as a comparison, ECED represents an improvement that is about 66% greater than the annual gains that would be expected for students in these grades. It is not possible to know if the current, low-income student sample would have larger or smaller gains than the national average during a typical year without intervention because our system for combining tests relied on standardized scores. Bloom, Hill, Black, and Lipsey (2008) considered this question for reading and found that gains among students eligible for FRPL were sometimes larger and sometimes smaller than the national averages. These values nonetheless provide a way to assess the magnitude of the effect sizes in this study and when framed in this context, these effects sizes seem quite meaningful.

A second way that Hill and colleagues (2008) suggest interpreting effect sizes is in comparison to research on similar interventions. In their meta-analysis of effective middle and high school math programs, Slavin et al. (2009) found a weighted mean effect size of .07 for all studies, .13 for the studies that used a randomized-control design, and .18 for instructional process interventions. Thus, ECED is a stronger than average math intervention and is

roughly equal to others that use intensive work with math teachers to improve classroom instruction.

On the other hand, Hill and colleagues' (2008) third recommendation for interpreting effect sizes indicates that ECED produced only small effects. They recommended comparing the effects of the intervention with policy-relevant achievement gaps, because interventions are typically designed to address such gaps. The current study specifically recruited schools serving high proportions of low-income students, so the gap between lower and higher income students is particularly relevant here. Using math data from the National Assessment of Educational Progress, Hill and colleagues reported mean performance gaps (in effect sizes) between students who were and were not eligible for FRPL of .80 for eighth graders and .72 for twelfth graders. Placed in that context, the current effect sizes for ninth and tenth graders are clearly small.

Thus, ECED's effects on math are medium-to-large in comparison to a typical year's growth at this age and in comparison to other math interventions, but small in comparison to the gap that needs to be filled. This contrast demonstrates the difficulty in altering math achievement at the high school level, especially for low-income students, many of whom have experienced years of low math achievement. ECED appears to make inroads in addressing that difficult problem, but is far from a full solution.

Differences Between Math and ELA Findings

There is considerable overlap between the ECED math and literacy interventions, raising questions as to why ECED Math affected student achievement but Literacy Matters did not. A similar pattern whereby an intervention affects math but not ELA achievement has been noted in past studies and several explanations have been posited. First, in general, ELA achievement is more difficult to measure than math achievement, so ELA tests are generally less reliable (Bill & Melinda Gates Foundation, 2012). Second, the tests may have been more closely aligned to the regular English curricula that were in use in both treatment and control schools than to the supplemental Literacy Matters curriculum. The Literacy Matters curriculum was identical in all schools, but the ELA outcome varied by state. Thus, there may be inconsistent overlap between the Literacy Matters curriculum and the tests. Third, standardized ELA tests tend to rely heavily on multiple-choice, reading items (Bill & Melinda Gates Foundation, 2012), whereas Literacy Matters focuses heavily on writing and critical thinking skills. According to IRRE, Literacy Matters was designed to align with national standards similar to those in the new Common Core State Standards and its assessments were more performance-based than typical standardized assessments. As schools move toward use of assessments specifically linked to the Common Core State Standards, Literacy Matters might demonstrate stronger impacts. Last, ELA skills may be more heavily influenced than math skills by factors other than ELA instruction, such as the home environment, other coursework, and extracurricular activities (Early et al., 2014).

Implications for Future Large-Scale Intervention Studies

Schools are under pressure to find feasible, applicable ways of improving instruction and increasing test scores. This project worked with 10 schools over two years to address this need while simultaneously involving them in a rigorous, large-scale, school-randomized

trial. It was not possible to conduct a highly controlled efficacy trial with researcher-administered outcome measures and optimal conditions, while also meeting the needs of participating schools to implement reforms that complemented their ongoing efforts. Thus, conditions required implementation of a hybrid approach in which these promising instructional supports were integrated into low-income, high-needs districts and schools with a full range of real-world challenges. Because the costs and time required for an effectiveness trial such as this one are high, investigators sometimes advance directly from an efficacy trial to implementation. That approach leaves open the question of whether the intervention is feasible under imperfect and complicated real-world circumstances. The field might be well served by adopting the current approach in which researchers conduct effectiveness testing on reasonably well-developed programs that have some empirical support but lack a formal efficacy trial.

Future Analyses

The analyses presented here addressed only the main research question posed by this study: what is the impact of ECED on students' math and ELA achievement scores? Many additional questions could be explored. For instance, the study also included EAR Protocol classroom observations made by independent observers and student and teacher questionnaires. In the future, those data could be used both to look for changes in instruction as a result of the intervention and conditions that lead to greater improvement in instruction.

Treatment-on-the-treated analyses would aid in further understanding the potential of the ECED approach. For instance, analyses that only include students who participated in all the ECED courses and teachers who received all the ECED supports would provide an estimate of the intervention's potential. The current analyses included students at treatment schools who did not enroll in the courses targeted by ECED, were enrolled only a few days, or had very low attendance. Additionally, some of the ECED teachers started working at a treatment school late in the project and received little of the support or implemented few of the strategies.

Conclusions

ECED appears to be a valuable path to increasing students' math success. In schools from five districts across four states, fraught with the types of problems that often arise when serving a high proportion of students from low-income homes, the use of ECED Math resulted in significant improvements in math achievement. The improvements were about two thirds greater than the national average for students in these grades. Further, given the stringent data collection and analytic procedures, difficulties encountered with retention and implementation, low power, and challenges in combining test scores, the effects sizes seen here are likely an underestimate of the true effects of ECED when fully implemented.

ECED is, however, a time-intensive intervention that requires significant changes at the classroom and school levels and significant district- and school-level buy-in and resources. Moreover, the demonstrated improvements were far from sufficient to address the gaps typically seen between schools serving higher and lower income students. Thus, an enhanced

intervention and further research are warranted. Next steps toward strengthening the intervention and evaluation would include taking into account lessons learned from the design and implementation challenges. We expect such steps would provide further evidence that ECED increases high school students' engagement and achievement.

Acknowledgment

The authors wish to thank the Institute for Research and Reform in Education for their support of this work and the participating students, teachers, schools, and districts for their support of the data collection efforts.

Funding

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305R070025 to the University of Rochester. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

ARTICLE HISTORY

Received 27 June 2014


Revised 14 April 2015


Accepted 16 May 2015


EDITORS


This article was reviewed and accepted under the editorship of Carol McDonald Connor and Spyros Konstantopoulos.


ORCID


Diane M. Early  <http://orcid.org/0000-0002-9071-700X>

Juliette K. Berg  <http://orcid.org/0000-0001-8150-5999>

Stacey Alicea  <http://orcid.org/0000-0003-4801-0139>

Yajuan Si  <http://orcid.org/0000-0001-8707-7374>

Richard M. Ryan  <http://orcid.org/0000-0002-2355-6154>

Edward L. Deci  <http://orcid.org/0000-0001-8246-8536>

References

- Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching. Combining high-quality observations with student surveys and achievement gains*. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf
- Black, A. E., & Deci, E. L. (2000). The effects of student self-regulation and instructor autonomy support on learning in a college-level natural science course: A self-determination theory perspective. *Science Education, 84*, 740–756. doi:10.1002/1098-237X
- Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M.W. (2008). *Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions* (MDRC Working Paper). Retrieved from <http://files.eric.ed.gov/fulltext/ED503202.pdf>
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

- Connell, J. P. (2010, June). The Institute for Research and Reform in Education and First Things First. In *Preparing college and career ready students: Elements of successful programs*. Webinar presented at the American Youth Policy Forum. Retrieved from <http://irre.org/publications/preparing-college-and-career-ready-students-elements-successful-programs>
- Connell, J. P., & Broom, J. (2004). *The toughest nut to crack: First Things First's (FTF) approach to improving teaching and learning*. Retrieved from http://www.irre.org/sites/default/files/publication_pdfs/The%20Toughest%20Nut%20to%20Crack.pdf
- Connell, J. P., Early, D. M., & Deci, E. L. (2014, March). *Every Classroom, Every Day intervention and implementation background*. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.
- Connell, J. P., Klem, A., Lacher, T., Leiderman, S., & Moore, W. (2009). *First Things First: Theory, research, and practice*. Retrieved from http://www.irre.org/sites/default/files/publication_pdfs/First_Things_First_Theory%2C_Research_and_Practice.pdf
- Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety*. San Francisco, CA: Jossey-Bass.
- Danner, F. W., & Lonky, E. (1981). A cognitive-developmental approach to the effects of rewards on intrinsic motivation. *Child Development*, 52, 1043–1052.
- Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. Dallas, TX: NSDC. Retrieved from <http://www.learningforward.org/docs/pdf/nsdcstudy2009.pdf>
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum Press.
- Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research*, 72, 433–479. doi:10.3102/00346543072003433
- Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Research*, 38, 181–199. doi:10.3102/0013189x08331140
- Desimone, L. M., Porter, A. C., Garet, M. S., Suk Yoon, K., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, 24, 81–112. doi:10.3102/01623737024002081
- Early, D. M., Rogge, R., Deci, E. L. (2014). Engagement, alignment, and rigor as vital signs of high-quality instruction: A classroom visit protocol for instructional improvement and research. *High School Journal*, 97(4), 219–239. doi:10.1353/hsj.2014.0008
- Elmore, R. F. (2002). *Bridging the gap between standards and achievement: The imperative for professional development in education*. Retrieved from http://www.shankerinstitute.org/Downloads/Bridging_Gap.pdf
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12(2), 121–138. doi:10.1037/1082-989X.12.2.121
- Gambone, M. A., Klem, A.M., Summers, J. A., Akey, T. A., & Sipe, C. L. (2004). *Turning the tide: The achievements of the First Things First education reform in the Kansas City, Kansas Public School District*. Retrieved from <http://www.ydsi.org/ydsi/pdf/turningthetidefullreport.pdf>
- Garet, M. S., Porter, A. C., Desimone, L. M., Birman, B., & Yoon, K.S. (2001). What makes professional development effective? Analysis of a national sample of teachers. *American Educational Research Journal*, 38(3), 915–945. doi:10.3102/00028312038004915
- Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology*, 52, 890–898. doi:10.1037/0022-3514.52.5.890
- Grolnick, W. S., & Ryan, R. M. (1989). Parent styles associated with children's self-regulation and competence in school. *Journal of Educational Psychology*, 81, 143–154. doi:10.1037/0022-0663.81.2.143
- Harter, S. (1978). Pleasure derived from optimal challenge and the effects of extrinsic rewards on children's difficulty level choices. *Child Development*, 49, 788–799.

- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60–87. doi:10.3102/0162373707299706
- Hill, C. J., Bloom, H. S., Black, A.R., Lipsey, M.W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2, 172–177. doi:10.1111/j.1750-8606.2008.00061.x
- Hulleman, C. S., Rimm-Kaufman, S. E., & Abry, T. (2013). Innovative methodologies to explore implementation: Whole-part-whole—Construct validity, measurement, and analytical issues for intervention fidelity assessment in education research. In T. Halle, A. Metz, & I. Martinez-Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 65–93). Baltimore, MD: Paul H. Brookes.
- Lee, V. L., & Burkam, D. T. (2003). Dropping out of high school: The role of school organization and structure. *American Educational Research Journal*, 40(2), 353–393. doi:10.3102/00028312040002353
- Lee, V. L., Croninger, R. G., & Smith, J.B. (1997). Course-taking, equity, and mathematics learning: Testing the constrained curriculum hypothesis in U.S. secondary school. *Education Evaluation and Policy Analysis*, 19(2), 99–121.
- Ludtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment—A reanalysis of TIMSS data. *Learning Environments Research*, 9, 215–230. doi:10.1007/s10984-006-9014-8
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376–390. doi:10.1037/0033-2909.114.2.376
- National Research Council and the Institute of Medicine. (2004). *Engaging schools: Fostering high school students' motivation to learn*. Washington, DC: The National Academies Press.
- Polikoff, M. S. (2012). Instructional alignment under No Child Left Behind. *American Journal of Education*, 118(3), 341–368. doi:10.1086/664773.
- Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, 31, 3–14. doi:10.3102/0013189X031007003
- Quint, J., Bloom, H. S., Black, A. R., Stephens, L., & Akey, T. M. (2005). *The challenge of scaling up educational reform: Findings and lessons from First Things First, Final report*. Retrieved from http://www.mdrc.org/sites/default/files/full_531.pdf
- Rakes, C. R., Valentine, J. C., McGatha, M.B., & Ronau, R.N. (2010). Methods of instructional improvement in algebra: A systematic review and meta-analysis. *Review of Educational Research*, 80, 372–400. doi:10.3102/0034654310374880
- Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley & Sons.
- Ryan, R.M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78. doi:10.1037/0003-066X.55.1.68
- Serdyukov, P., & Hill, R. (2008). E-learning: What works, what doesn't, what now? In C. Bonk, M. Lee, & T. Reynolds (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 1248–1253). Chesapeake, VA: AACE.
- Si, Y., & Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38, 499–521. doi:10.3102/1076998613480394
- Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best evidence synthesis. *Reading Research Quarterly*, 43(3), 290–322. doi:10.1598/RRQ.43.3.4
- Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, 79(2), 839–911. doi:10.3102/0034654308330968
- Stotsky, S. (2005). *The state of state English standards*. Washington, DC: Thomas Fordham Foundation.

Su, Y., Gelman, A., Hill, J., & Yajima, M. (n.d.). *mi* [Computer software]. Retrieved from <http://www.stat.columbia.edu/~gelman/software/>

U.S. Department of Education. (n.d.). *Common Core of Data (CCD)*. Retrieved from <http://nces.ed.gov/ccd/>

What Works Clearinghouse (WWC). (2014). *Procedures and standards handbook: Version 3.0*. Retrieved from ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

Appendix

Table A1. Predicting math and ELA achievement using unimputed data.

Condition	Math achievement						ELA achievement					
	Year 1 (n = 4,928)			Year 2 (n = 4,438)			Year 1 (n = 4,832)			Year 2 (n = 4,398)		
	Estimate	SE	p	Estimate	SE	p	Estimate	SE	p	Estimate	SE	p
Condition	0.176	0.076	0.036	0.123	0.057	0.050	0.033	0.032	0.319	0.022	0.032	0.513
Covariates												
District 2	0.065	0.117	0.587	-0.088	0.087	0.331	-0.011	0.044	0.802	-0.009	0.046	0.853
District 3	0.151	0.119	0.228	0.178	0.091	0.070	-0.010	0.048	0.836	-0.059	0.052	0.272
District 4	0.282	0.117	0.030	0.510	0.087	0.000	-0.134	0.044	0.012	-0.027	0.047	0.573
District 5	0.197	0.123	0.133	0.276	0.096	0.013	—	—	—	-0.203	0.065	0.008
Baseline	0.614	0.011	0.000	0.504	0.012	0.000	0.817	0.010	0.000	0.790	0.011	0.000
Gender	-0.002	0.021	0.917	-0.056	0.022	0.010	-0.001	0.017	0.948	-0.031	0.018	0.086
Hispanic	-0.135	0.038	0.001	-0.137	0.040	0.001	-0.062	0.031	0.046	-0.060	0.034	0.075
Black	-0.130	0.039	0.001	-0.180	0.042	0.000	-0.097	0.032	0.003	-0.020	0.035	0.572
Asian/Pac. Isl.	0.013	0.050	0.792	0.032	0.052	0.532	-0.003	0.041	0.937	0.050	0.044	0.256
Other ethnicity	-0.198	0.076	0.010	-0.058	0.081	0.473	-0.052	0.060	0.391	-0.077	0.065	0.238
FRPL	0.018	0.025	0.468	0.036	0.026	0.166	-0.019	0.020	0.347	-0.032	0.021	0.136
Spec. ed.	-0.275	0.051	0.000	-0.299	0.056	0.000	-0.150	0.049	0.003	-0.107	0.043	0.012
ELL	-0.008	0.029	0.789	0.504	0.012	0.000	-0.034	0.024	0.162	0.017	0.026	0.516

Note. Unstandardized estimates shown. Students were in 9th grade in Year 1. Most were in 10th grade in Year 2, but some were still in 9th grade. Condition: 0 = control, 1 = treatment; Districts 2–5 are compared to District 1; Gender: 0 = male, 1 = female; each race/ethnicity group is compared to White students; Free/reduced-price lunch: 0 = received neither year of the study, 1 = received one or both years of the study; special education and English language learner: 0 = service not received during Year 1, 1 = service received during Year 1. Effect sizes: Math Year 1: .18; Math Year 2: .13; ELA Year 1: .03; ELA Year 2: .02. All Year 1 ELA data for District 5 were missing.