Educational Evaluation and Policy Analysis
http://eepa.aera.net

A Multistate District-Level Cluster Randomized Trial of the Impact of Data-Driven Reform on Reading and Mathematics Achievement

Deven Carlson, Geoffrey D. Borman and Michelle Robinson

EDUCATIONAL EVALUATION AND POLICY ANALYSIS 2011 33: 378
DOI: 10.3102/0162373711412765

The online version of this article can be found at:
http://epa.sagepub.com/content/33/3/378

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

Educational Evaluation and Policy Analysis
September 2011, Vol. 33, No. 3, pp. 378–398
DOI: 10.3102/0162373711412765

A Multistate District-Level Cluster Randomized Trial of the Impact of Data-Driven Reform on Reading and Mathematics Achievement

Deven Carlson
Geoffrey D. Borman
Michelle Robinson
University of Wisconsin–Madison

Analyzing mathematics and reading achievement outcomes from a district-level random assignment study fielded in over 500 schools within 59 school districts and seven states, the authors estimate the 1-year impacts of a data-driven reform initiative implemented by the Johns Hopkins Center for Data-Driven Reform in Education (CDDRE). CDDRE consultants work with districts to implement quarterly student benchmark assessments and provide district and school leaders with extensive training on interpreting and using the data to guide reform. Relative to a control condition, in which districts operated as usual without CDDRE services, the data-driven reform initiative caused statistically significant districtwide improvements in student mathematics achievement. The CDDRE intervention also had a positive effect on reading achievement, but the estimates fell short of conventional levels of statistical significance.

Keywords: *data-driven reform; experimental design; benchmark assessments; student achievement*

The development of student assessments, accountability programs, and the use of associated data systems have recently emerged as central strategies for improving the nation's public schools. Much of the impetus behind these efforts began with the No Child Left Behind Act of 2001 (NCLB), which ushered in test-based accountability as the predominant model of educational reform promulgated by the federal government. Initially, NCLB compelled states to design comprehensive school accountability systems on the basis of annual student assessments. Using these assessments to measure student progress, the law holds schools and districts accountable for students' academic performance and provides a lever for national reform of American public education. Under the law, schools must ensure that an increasing percentage of students meet state-specified proficiency standards for the schools to be rated as making adequate yearly progress (AYP). The longer a school fails to make AYP, the more severe are the corrective actions it must undertake.

Building on the momentum of the national accountability movement, and exploiting the data warehouses that are accumulating across the country, a growing number of schools and school systems have been implementing policies that go beyond NCLB in their capitalization on information

This research was supported by a grant from the Institute for Education Science, U.S. Department of Education (R-305A040082), as well as by predoctoral fellowships for the first and third authors from the Institute of Education Sciences through Award R305C050055 to the University of Wisconsin–Madison. However, any opinions expressed are those of the authors and do not represent the positions or policies of the Department of Education.

provided by student assessments. For example, some school districts have recently begun to invest in systems to create enhanced access to student performance data. Other districts are implementing quarterly benchmark assessments, coupled with professional development initiatives, to encourage teachers, principals, and district leaders to use data to inform their policies and practices. Such policies are being guided by a stream of new research, which we turn to later, that describes how practitioners can best access and use data to inform, and potentially change, their practices.

Of course, the ultimate goal of any educational reform is not to change policy and practice but to improve student outcomes. On this score, only a small body of research has evaluated the impacts of adopting more proactive uses of data and data systems on student and school achievement. One of the more prominent undertakings, which serves as the basis for the evaluation presented in this article, is the initiative fielded by developers from the Center for Data-Driven Reform in Education (CDDRE) at Johns Hopkins University. In this initiative, personnel at CDDRE developed a replicable approach to whole-district change on the basis of the concepts of data-driven reform. The goal of CDDRE is to solve the problem of scale in educational reform by working with entire school districts to help district and school leaders understand existing data on student performance, generate additional data to help guide school improvement efforts, identify root causes underlying important problems, and then select and effectively implement evidence-based programs directed toward solving those problems.

The CDDRE intervention consisted of several distinct components that were implemented at different points over a 3-year period. In the 1st year, participating treatment districts implemented benchmark assessments and received extensive consulting services on interpreting the data from the benchmark assessments, and on data-driven reform more generally. In the 2nd year of the CDDRE intervention, school and district leaders were expected to seek out evidence-based reforms that would address the needs and problems identified by the data. In the 3rd and final year, schools were expected to adopt and implement either proven programs or other solutions that were backed by solid evidence, particularly in reading.

The original research design for the evaluation of the CDDRE intervention involved random assignment of nearly 60 school districts, containing more than 500 participating schools, across seven states in three school-year cohorts. In each cohort, treatment districts began receiving the CDDRE intervention in Year 1 while control districts continued with "business as usual" but began receiving the treatment in Year 2. Because the control districts received no intervention during Year 1, the 1st-year outcomes provide pure experimental evaluations of the impacts of the benchmark assessment and data interpretation aspects of CDDRE. These are the results that we focus on within this article. Specifically, we address the following question: Does the implementation and administration of benchmark assessments, coupled with the provision of consulting services to assist in the interpretation of the resulting data, bring about districtwide changes in student performance relative to a control condition in which districts operated as usual without benchmark assessments and associated services?

Background on Data-Driven Reform and Benchmark Assessments

Data-driven reform involves collecting, interpreting, and disseminating data in a manner that is intended to inform and guide district and school improvement efforts. Bernhardt (2003) identified four categories of data that practitioners may analyze to inform reform efforts: student learning and assessment, demographics, school process, and teacher perceptions. Analyzing these four types of data can provide school leaders with a great deal of valuable information. For instance, principals and superintendents may use data to detect specific problems faced by students and teachers, to identify individual schools and demographic groups in need of particular help, or to determine the underlying causes of achievement gaps (Kennedy, 2003; Schmoker, 2003). Of the four types of data that can be used to inform reform efforts, the most common one, and the one that serves as the topic of this article, is student learning and assessment data. Perie, Marion, and Gong (2009) discussed three possible uses of assessment results: (a) instructional: to help teachers adjust their instruction and curriculum to address student

learning needs; (b) evaluative: to help educators evaluate and improve broader schoolwide or districtwide instructional programs; and (c) predictive: to determine each student's likelihood of achieving particular performance standards on yearly assessments. Data-based decision making usually involves extensive professional development for teachers and school leaders to help them use data to set goals, prioritize resources, and make intervention plans (Conrad & Eller, 2003).

In addition to the potential benefits of data-based decision making, researchers have also highlighted some decidedly negative outcomes associated with data use, especially within the context of high-stakes accountability systems. A sampling of the potential negative aspects of data-driven reform include attempts to game the system (Booher-Jennings, 2005; Heilig & Darling-Hammond, 2008), a narrowing of the curriculum (Crocco & Costigan, 2007; Diamond & Cooper, 2007; Ogawa, Sandholtz, Martinez-Flores, & Scribner, 2003; Wright & Choi, 2006), and short-term, superficial changes in practice (Diamond & Cooper, 2007). Clearly, designers of any data-based reform effort must be aware of such concerns as they move toward implementation.

Data-driven reform and changes in teachers' practices. The empirical literature on benchmark assessments, and data-driven educational reform more generally, can be classified into two main groups of studies. The first set of analyses examines how, if at all, teachers alter classroom instruction in response to data on student performance. This literature is fairly well developed, both theoretically and empirically. The empirical studies within this genre are typically based on the results of surveys administered to teachers in a relatively small number of schools or districts. The analyses present compelling evidence that teachers believe that interim tests assist them in identifying areas in which their students would benefit from additional instruction (Dembosky, Pane, Barney, & Christina, 2005; Marsh, Pane, & Hamilton, 2006; Mason, 2002; Stecher & Hamilton, 2006). Teachers also report that they alter their instruction in response to assessment results (Christman et al., 2009; Clune & White, 2008; Stecher et al., 2008). These studies, though, are based on teacher self-reports and do not rely on objective evaluations of how individual teachers actually analyze and use assessment data to inform their classroom practice or how policy conditions may support teachers' ability to use assessment data to improve instruction.

A recent study by Goertz, Olah, and Riggan (2009) provides an objective appraisal of teachers' use of assessment data to inform instruction. The study design involved observing 45 elementary school teachers in nine schools across two districts to examine how teachers used assessments to inform their instructional practices in mathematics during the 2006–2007 school year. Goertz et al. found that there was substantial diversity in the effectiveness with which teachers were able to use the data, and they identified several factors that were correlated with their effectiveness. In general, Goertz et al. found that well-supported districtwide efforts to provide access to the data, along with training on how to use them, were successful in encouraging teachers' use of data. However, although teachers accessed and analyzed the data, the authors found that this information did not substantially change their instructional and assessment practices in the classroom. Teachers used the data largely to decide what content to reteach and to whom but not to make fundamental changes in the content or in the way students were taught. Although school leadership and a culture of data use, along with districtwide supports, were critical for supporting teachers' use of data, Goertz et al. suggested that district and school leaders should consider that teachers need more professional development and support on interpreting data and on connecting this evidence to specific instructional approaches and strategies.

Data-driven reform and student achievement. A second group of studies examines the effects of data-driven decision making, including benchmark assessments, on student outcomes. This literature is less developed, with only three large-scale empirical analyses. The first study evaluated a data-driven instruction program that was implemented in 21 public schools in Boston during the 2005–2006 and 2006–2007 school years (Quint, Sepanik, & Smith, 2008). The intervention consisted of administering several short assessments of reading comprehension to third and fourth grade students throughout the school year. The assessments were designed to mirror the Massachusetts

Comprehensive Assessment System, which is the state's high-stakes test used for NCLB purposes. The results of the assessments were compiled into reports that contained information about each student, and instructional data coaches met with each teacher to review the reports and suggest instructional responses on the basis of the results. The effectiveness of this intervention was assessed using a comparative interrupted time-series design. The results of the evaluation generally failed to reveal statistically significant differences in student achievement between treatment and comparison schools.

In a separate evaluation, Henderson, Petrosino, Guckenburg, and Hamilton (2007) analyzed the effect of benchmark assessments on student achievement in eight Massachusetts school districts. In this intervention, a data management system was used to develop and administer quarterly benchmark assessments in eighth grade mathematics during the 2005–2006 school year. This evaluation, which also used an interrupted time-series design, failed to find statistically significant effects of the intervention. The point estimates were positive, but the study was relatively underpowered, with only 22 treatment schools and 44 matched comparison schools.

Another recent study, conducted by May and Robinson (2007), evaluated Ohio's Personalized Assessment Reporting System (PARS) for the Ohio Graduation Tests. The PARS generates several test score reports for teachers, students, administrators, and parents. The primary goal of PARS is to promote greater passage rates of the Ohio Graduation Tests and greater rates of graduation from high school. The effects of PARS were evaluated using a design in which 60 districts, containing 100 high schools, were randomly assigned to participate in the PARS pilot during the 2006–2007 school year. The results of the evaluation indicated little effect on Ohio Graduation Tests performance for 10th grade students who were taking the test for the first time. However, statistically significant, positive effects of the PARS were found for students who initially failed the Ohio Graduation Tests. Such students were up to 4 times more likely to retake the tests if they were in districts that had been assigned to PARS, and they also scored higher than their peers in districts that had not been assigned to PARS.

Beyond these large-scale studies, there are a number of case studies that attribute achievement improvements to data-driven decision making (e.g., Datnow, Park, & Wohlstetter, 2007; Petrides & Nodine, 2005; Snipes, Doolittle, & Herlihy, 2002). Each of these studies examined multiple districts using in-depth case studies and, in each case, concluded that data use was instrumental in generating the improvements that districts had seen. Although these studies are suggestive of the ability of data use to improve student outcomes, the methods do not allow strong causal conclusions to be drawn. In general, these studies provide after-the-fact explanations for positive results, and it is unclear, for example, whether schools and districts that did not make impressive gains may also have been attempting to use the same data-driven strategies (see Herman et al., 2008).

Related to the literature on the potential achievement impacts of benchmark assessments are studies on the effects of formative assessments. This literature dates back to the 1970s, and literally hundreds of studies have examined the effects of these assessments on student outcomes. The results of these analyses are summarized in an article by Black and Wiliam (1998). These authors reviewed approximately 250 studies that were published on formative assessments between 1988 and 1998 (reviews of studies on the topic published prior to 1988 can be found in Natriello, 1987, and Crooks, 1988). Black and Wiliam's review concluded that formative assessments can have substantial effects on student achievement. They estimated that formative assessments can improve student performance by 20% to 40%; these estimates are consistent with effect sizes on the order of 0.4 to 0.7 (see Henderson et al., 2007).

Advancing the Data-Driven Reform Literature

Although the empirical work that examines the effects of data-driven decision making on student outcomes continues to grow, the effectiveness of data-driven reform remains equivocal and far from conclusive. Teachers seem to believe that accountability systems that offer them access to assessment data can be helpful, but these systems appear to have had mixed effects on actually changing instructional practices. Similarly, the work of Henderson et al. (2007) and May and Robinson

(2007) provides some suggestive evidence of positive impacts, but the results are not fully definitive. A recent report by a group of RAND Corporation researchers noted the relative paucity of research on the relationship between data-driven decision making and student achievement and summed up the most promising way forward by stating that

> experimental studies are needed to more rigorously measure the effects of enhanced provision of data and supports to use it. Standardized interventions can be developed and tested in randomized trials. For example, studies might examine whether the provision of interim progress test data or value-added measures, combined with ongoing professional development for teachers on how to use the information, leads to better instruction and higher achievement than do classrooms without such data and training. (Marsh et al., 2006, p. 12)

This article answers the call for experimental evaluation of the effects of enhanced data provision and training. We use a rigorous randomized design to estimate the impact of benchmark assessments, and support for interpreting the resulting data, on student and school achievement. Our results offer valuable new information on the effects of the growing movement toward data-driven reform on achievement outcomes.

Method

Sample Recruitment and Randomization Procedures

As noted above, the districts and schools that constitute the sample for this analysis are part of a larger research and development project conducted by the CDDRE at Johns Hopkins University.[1] The sample recruitment process, which took place over the course of 3 school years, began with CDDRE officials contacting state departments of education in seven states—Alabama, Arizona, Indiana, Mississippi, Ohio, Pennsylvania, and Tennessee—and asking them to nominate districts with large numbers of low-performing schools for participation in their study. After receiving the nominations from the state departments of education, CDDRE officials approached the first cohort of districts, located in Pennsylvania, Alabama, Arizona, and Ohio, during the 2004–2005 school year and inquired whether there was interest in participating in the study. If district officials agreed to participate, they were told that they would be assigned to begin receiving the CDDRE intervention, which included assistance in the implementation and administration of benchmark assessments, during the 2005–2006 school year or during the 2006–2007 school year. The second cohort of districts, located entirely in Pennsylvania, was approached the following school year and told that receipt of CDDRE services would begin during the 2006–2007 school year or the 2007–2008 school year. A similar process occurred in Wave 3, when participating districts in Pennsylvania, Indiana, Mississippi, and Tennessee were informed that they would begin receiving CDDRE services in either the 2007–2008 school year or the 2008–2009 school year. Districts were offered no direct financial incentives for participating in the study, but all aspects of the CDDRE intervention, including the benchmark assessments and all related consulting services, were provided free of charge.

Across the three waves, the recruitment process resulted in 59 districts agreeing to participate in the reading portion of the CDDRE research project. Of these 59 districts, 57 also agreed to participate in the mathematics portion of the CDDRE study; the two Ohio districts, Kenton City and East Cleveland, agreed to receive CDDRE services in reading but not in mathematics.[2] This decision was conveyed to CDDRE staff members prior to randomization. Within each district, the leadership decided which schools would be targeted to receive the CDDRE intervention. In general, officials identified a subset of the lowest performing schools within the district to receive the benchmark assessments and associated consulting services on interpreting the data. Across the 59 districts participating in the reading portion of the study, district leaders selected 549 schools to receive the CDDRE intervention. In the 57 districts participating in the mathematics portion of the study, district leaders identified 538 schools that would implement benchmark assessments and associated consulting services.

The majority of districts and schools in our sample are low performing, but they are diverse in many other respects. First, the schools and districts in our sample are spread across seven states that represent nearly every region of the country; the sample is geographically diverse. Second, our sample contains both large, urban districts such as Pittsburgh, Pennsylvania, and Phoenix, Arizona, and smaller, more rural districts, such as Duquesne,

Pennsylvania. Third, our sample possesses significant racial and socioeconomic diversity; some districts enroll large proportions of White students, whereas others have significant shares of African American and Hispanic students. The proportion of students eligible for free or reduced-price lunch is also quite variable across districts. Average baseline achievement and demographic characteristics for participating schools in each treatment and control district are presented in Table 1.

After each cohort of districts was recruited and district officials identified the schools that they wanted to target for intervention, the randomization process occurred. Randomization, which took place at the district level, was achieved using a randomized block design. More specifically, for each recruitment wave, districts were grouped by state and then, within each state block, were randomly assigned to either the proximate treatment condition or the delayed treatment condition, with a selection probability of approximately 50% in all cases. For the purposes of this analysis, we consider districts assigned to the proximate treatment condition as being in the treatment group and districts assigned to the delayed treatment condition as being in the control group.

Randomized block designs such as the one used in this study have several advantages. First, they ensure that the intervention will be distributed in a fair and equitable manner. In this case, blocking by state ensures that each state will have approximately 50% of its participating districts assigned to the proximate treatment condition. Second, it ensures that the district and school samples in the treatment and control groups will be drawn at equal rates from each state context. Failure to achieve balance on this visible and salient policy and geographic dimension could threaten the face validity of the study. Finally, block randomization designs have the potential to increase the statistical power of an analysis (Raudenbush, Martinez, & Spybrook, 2007). In this case, if the within-state district correlation on our outcome measure is larger than the cross-state correlation for the outcome measure, our randomized block design will result in increased statistical power to detect treatment effects. Table 1 presents the results of the randomization process. For each treatment and control district, it presents the average values for all participating schools on a wide variety of pretreatment covariates, including baseline test scores, enrollment, and several demographic characteristics. The table indicates that the randomization process succeeded in balancing the treatment and control groups on a wide variety of baseline measures. Indeed, *t* tests indicated that there were no statistically significant differences between the treatment and control groups on any of the pretreatment measures.[3]

Design and Implementation of the Treatment Condition

As we described in the introduction, the CDDRE intervention is made up of several components that are implemented across a 3-year period. Our analyses test the impacts of the 1st-year components of the CDDRE treatment, which include the implementation of benchmark assessments and supporting consulting services focused on interpreting the data from the benchmark assessments and on data-driven reform more generally. The counterfactual was "business as usual" during the 1-year treatment delay. The overall CDDRE intervention design is based on the Data-Driven District (3D) model, which was created by CDDRE. The 3D model bridges two major approaches to reform of low-achieving districts and schools: data-based district reform and comprehensive school and classroom reforms. It is designed to align the efforts of state, district, and school-based educators around the goal of accelerating achievement in low-performing schools. The 3D intervention is based on a goal-focused implementation process, which is summarized in Figure 1.

The overall 3D model elements include quarterly benchmark assessments, data reviews, training in leadership and data interpretation, provision of reviews of research on effective programs and practices, and assistance in selecting and implementing proven programs. Of these five elements, described in greater detail below, only the first three apply to Year 1 of the study, which is the focus of this article.

1. Quarterly benchmark assessments tied to state standards and assessments in reading, writing, and mathematics, which are capable of predicting performance on state assessments. These are used both to determine needs for specific interventions and

TABLE 1

Average Baseline Achievement and Demographic Characteristics for Participating Schools in Each Treatment and Control District

District	Cohort	State	Baseline Test: Reading	Baseline Test: Math	Enrollment	% White	% Black	% Hispanic	% FRPL	FTE Teachers	Pupil/ Teacher Ratio
Treatment districts											
Laurel Highlands	1	PA	–0.15	–0.06	606	92.93	6.21	0.38	41.29	35.08	17.08
Philipsburg-Osceola	1	PA	0.06	0.00	351	98.48	0.56	0.62	38.38	27.87	12.35
Reading	1	PA	–0.55	–0.58	735	14.85	14.02	70.39	75.89	40.62	18.40
Upper Darby	1	PA	–0.18	–0.28	991	54.11	28.38	1.80	28.77	61.28	15.73
Woodland Hills	1	PA	–0.33	–0.33	676	39.45	59.21	0.57	59.92	47.08	14.08
Franklin County	1	AL	–0.02	0.08	498	93.18	0.30	6.33	61.49	36.70	13.13
Huntsville City	1	AL	–0.47	–0.41	365	17.57	75.87	5.41	86.26	29.51	12.32
Alhambra	1	AZ	–0.05	–0.01	995	14.05	7.64	72.41	94.31	50.42	19.70
East Cleveland	1	OH	–0.53	NA	574	0.26	99.20	0.00	83.79	31.65	18.18
Allentown	2	PA	–0.46	–0.25	866	23.35	17.48	57.38	74.16	44.23	19.25
Aliquippa	2	PA	–0.51	–0.47	454	23.53	75.82	0.53	77.76	38.33	12.40
Bristol	2	PA	–0.31	–0.28	501	69.87	19.11	7.83	43.12	34.85	15.59
Clairton	2	PA	–0.84	–0.85	301	33.01	66.99	0.00	66.05	26.30	11.27
Harrisburg	2	PA	–0.95	–0.99	597	5.62	75.49	16.74	63.18	42.94	13.34
SE Greene	2	PA	–0.46	–0.61	344	99.14	0.86	0.00	57.40	25.00	13.80
Wilkinsburg	2	PA	–0.75	–0.71	309	2.35	97.16	0.39	68.65	26.80	11.50
Bethlehem	3	PA	–0.04	–0.03	691	56.13	9.50	31.11	40.18	42.07	16.34
Erie	3	PA	–0.43	–0.45	663	30.11	54.69	13.18	90.07	45.23	14.50
Everett	3	PA	–0.12	0.00	311	98.57	0.53	0.08	49.48	19.20	15.12
Farrell	3	PA	–0.57	–0.66	539	23.07	75.75	0.50	55.22	36.40	14.80
Lebanon	3	PA	–0.32	–0.35	633	50.64	6.94	41.14	62.22	37.54	17.46
Michigan City	3	IN	–0.32	–0.26	521	53.50	32.77	4.23	67.20	NA	NA
Richmond	3	IN	–0.25	–0.15	442	76.51	8.49	3.51	66.44	NA	NA
South Delta	3	MS	–0.35	–0.23	423	3.67	95.62	0.71	99.25	24.40	17.07
Yazoo City	3	MS	–0.54	–0.53	706	0.73	98.96	0.10	94.22	36.10	19.45
Fayette	3	TN	–0.37	–0.41	373	30.74	66.53	2.28	78.54	23.29	15.19
Greene	3	TN	–0.07	0.13	430	95.73	1.47	2.35	57.80	26.62	15.98
Macon	3	TN	–0.20	–0.13	480	94.68	0.49	4.18	53.06	25.23	19.23
Maury	3	TN	–0.11	–0.09	585	73.32	20.39	5.42	52.37	36.47	16.01
Wayne	3	TN	–0.13	–0.05	348	97.03	1.84	0.79	70.52	29.92	12.56
Treatment district mean			–0.34	–0.31	544	48.87	37.28	11.68	65.23	35.04	15.42

(continued)

TABLE 1 (continued)

District	Cohort	State	Baseline Test: Reading	Baseline Test: Math	Enrollment	% White	% Black	% Hispanic	% FRPL	FTE Teachers	Pupil/ Teacher Ratio
Control districts											
Tuscaloosa City	1	AL	–0.58	–0.68	424	1.60	96.44	1.78	91.26	32.25	13.20
Fairfield City	1	AL	–0.12	–0.03	401	0.38	99.00	0.62	80.89	27.50	14.48
Flagstaff	1	AZ	0.19	0.12	467	47.11	2.34	20.40	55.89	29.43	16.19
Phoenix	1	AZ	–0.25	–0.44	464	6.14	5.24	85.27	75.22	24.58	18.71
Kenton City	1	OH	0.00	NA	253	94.74	0.50	1.57	45.76	15.22	17.80
Blairsville	1	PA	0.11	0.19	443	97.18	2.50	0.09	39.48	27.83	15.93
Highlands	1	PA	0.22	0.20	490	91.68	6.87	0.46	40.83	33.90	14.87
Sharon City	1	PA	–0.14	–0.15	583	70.20	27.67	0.96	66.43	38.40	14.65
Wilkes-Barre	1	PA	0.22	0.24	774	74.80	15.49	7.93	35.85	50.30	15.50
William Penn	1	PA	–0.54	–0.65	510	13.69	83.19	0.95	48.05	30.85	17.35
York	1	PA	–0.61	–0.50	735	24.05	42.06	32.82	78.85	45.67	15.71
Duquesne	2	PA	–0.78	–0.42	353	7.42	92.00	0.58	76.96	29.25	11.55
Lancaster	2	PA	–0.49	–0.38	465	22.32	23.30	51.95	76.83	27.55	17.37
Pittsburgh	2	PA	–0.24	–0.21	346	40.01	56.75	1.02	60.73	23.71	14.90
SE Delco	2	PA	–0.41	–0.51	826	44.79	52.10	1.65	50.65	53.44	15.32
Steelton	2	PA	–0.61	–0.58	661	34.12	52.63	12.59	47.55	51.50	12.85
Sto-Rox	2	PA	–0.67	–0.67	438	58.81	39.77	1.19	65.66	33.00	13.17
Towanda	2	PA	–0.15	–0.11	582	98.08	0.81	0.58	42.02	34.83	16.40
Big Beaver Falls	3	PA	0.01	0.19	446	68.39	30.74	0.63	61.18	36.00	12.98
Central Dauphin	3	PA	–0.30	–0.29	734	44.18	39.23	11.54	41.19	52.32	14.06
Roosevelt	3	AZ	–0.54	–0.61	604	3.44	17.60	77.22	84.99	33.00	18.15
Anderson CS	3	IN	–0.40	–0.40	673	69.10	21.21	3.93	55.47	NA	NA
Humphreys	3	MS	–0.28	–0.29	460	1.77	96.95	1.24	97.92	24.90	18.73
West Bolivar	3	MS	–0.14	–0.11	338	3.34	95.87	0.79	96.92	21.53	16.03
Roane	3	TN	0.04	0.14	452	94.00	4.51	0.79	49.11	26.45	16.97
Robertson	3	TN	–0.02	–0.04	601	79.21	12.01	8.15	42.58	36.61	16.81
Carter	3	TN	–0.12	–0.10	398	97.70	0.66	1.25	67.69	28.77	13.70
Hancock	3	TN	–0.45	–0.28	517	98.82	1.00	0.09	84.32	38.50	13.40
Hawkins	3	TN	–0.11	0.01	342	97.75	1.17	0.74	65.27	24.22	13.49
Control district means			–0.25	–0.23	510	51.20	35.16	11.34	62.95	33.27	15.37

Note. FRPL = free or reduced-price lunch; NA = not available.

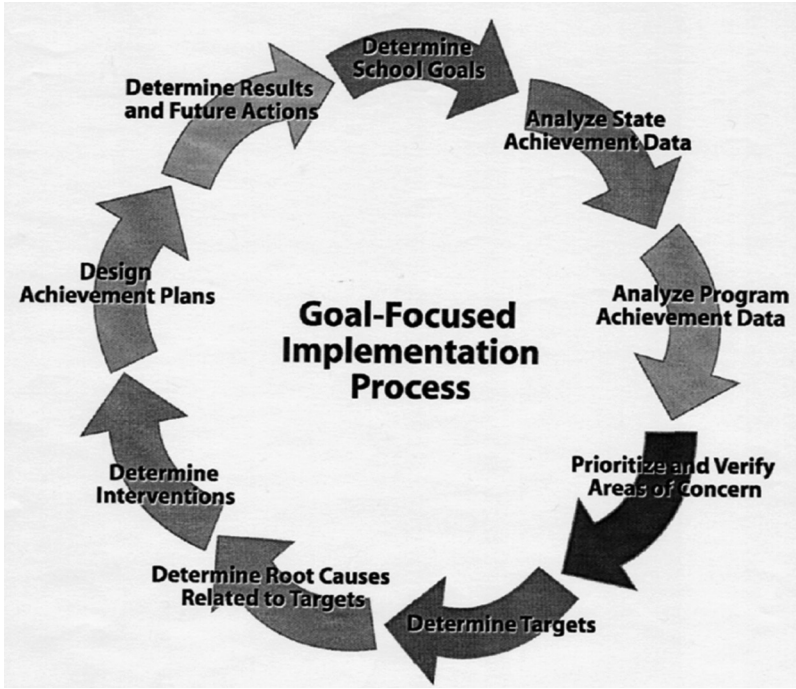

FIGURE 1. *The Center for Data-Driven Reform in Education goal-focused implementation process.*

to evaluate students' progress toward state goals after interventions are implemented.
2. Detailed reviews of state test data, benchmark assessments, questionnaires from educators at all levels, and other indicators to identify areas of need for schools that are not meeting AYP goals or that are at risk for AYP failure.
3. Training for state, district, and building leaders in interpreting and using data, managing resources to focus on areas of need, and leading a data-driven reform process.
4. Provision of clear, actionable reviews of research on interventions for the types of problems likely to be identified in the data review process.
5. Assistance in selecting and then implementing in low-performing schools interventions drawn from many sources and providers, designed to help schools meet specific goals. Interventions favorably reviewed by the What Works Clearinghouse, or reviews using similar standards, are emphasized.

The benchmark assessments are designed to monitor the progress of children in Grades 3 to 8 (in Pennsylvania, Grades 3–11) in reading and mathematics and to guide data-driven reform efforts. These benchmark assessments, called 4Sight, were created from the same assessment blueprints as those used to construct the state assessments and were written to mirror each state assessment's content, coverage, difficulty, item types, proportions of open-ended items, and use of illustrations and other supports. Student scores on the 4Sight benchmarks correlated with scores on the state test in the range of .80 to .85. These assessments were designed to be administered four or five times per year to predict what students, student subgroups, classes, and schools would have scored on the state assessments. Although 4Sight tests were available for each state involved in the research, comparable benchmark assessments that were administered within the district prior to the study were accepted as part of the treatment (in lieu of 4Sight), and all other services were provided as planned.[4] Information collected from these assessments was used to compile data review reports, described in greater detail below, that were discussed in the monthly data training with school staff members.

To best assist treatment schools with their data-driven reform efforts, CDDRE consultants worked with district and school leaders to review a broad range of data to identify problem areas at each school. The CDDRE consultants were experienced senior educators who had been superintendents, central office administrators, or successful principals. They were drawn from the staffs of CDDRE school reform partners: the Success for All Foundation, the National Center for Education and the Economy, Modern Red Schoolhouse, Co-Nect, the National Institute for Direct Instruction, Howard University's Talent Quest, and the University of Memphis. These experts looked at data on state tests, broken down by subskills, grades, and student subgroups, to identify school strengths as well as areas in need of intervention.[5] Consultants were also charged with reviewing data on quarterly benchmark assessments, as well as data on such indicators as retention rates, special education placements, attendance, and disciplinary actions (e.g., suspensions). Teachers were also surveyed to collect information on their perceptions of school strengths and needs. Data reviews were expected to produce clear, easily interpretable reports that focus on actionable information. In each treatment school, these reports were reviewed by a teacher leadership team as well as by the building principal to provide a broad range of school staff members an opportunity to select or create solutions to the problems identified.

In conjunction with the data review, the consultants also carried out training for practitioners. These training sessions were held on a monthly basis at each treatment school and were attended by district, school, and teacher leaders (e.g., "school action teams"). Consultants led school staff members in the interpretation and use of the data produced by the data review process, showed them how they may use the data to pinpoint areas of need within their school, and informed them of programs available to address the specific problems identified. Special software enabled school leaders and teachers to examine the data by state standard, grade, class, student subgroup, and other relevant features.

This large-scale trial was designed to estimate the effect of assignment to receive CDDRE services on reading and mathematics achievement under typical implementation conditions. As a result, no formal implementation study was conducted. However, general information concerning overall compliance with the key aspects of the treatment—the data reviews, the training sessions for school action teams, and the benchmark assessments—is available and described below.

District- and school-level compliance with the data reviews and training sessions was generally sound. CDDRE staff members indicate that all schools and districts assigned to the treatment group were represented by essential personnel—those individuals necessary for advancing the CDDRE model—at all data reviews and training sessions. The only form of noncompliance with respect to the data reviews and training sessions was isolated absences due to illnesses or other excused reasons.

In contrast to the limited noncompliance with the data reviews and training sessions is some significant school-level noncompliance in the use of benchmark assessments, particularly in the first cohort of treatment schools. As described above, these assessments were designed to be administered four or five times per academic year. However, the reviews of state assessment data and the related training sessions for school and district personnel were much more demanding and time intensive than initially anticipated by CDDRE staff members. This fact, coupled with the inevitable difficulties that accompany the startup of a large-scale effectiveness trial, resulted in approximately 60% to 70% of the first-cohort treatment schools administering only one or two benchmark assessments to their students. In the two subsequent cohorts, however, these problems were largely resolved, and over 90% of treatment schools administered either three or four benchmark assessments. Furthermore, as noted above, benchmark assessments were administered in two of the control districts, Phoenix and Anderson CS, prior to the implementation of the CDDRE intervention.

Taken as a whole, the imperfect implementation of benchmark assessments, coupled with the administration of benchmark assessments in two control districts, is likely to depress the impact estimates of the CDDRE intervention. However, the noncompliance described above reflects implementation as it would likely occur in the real world. Consequently, the impact estimates presented in this study are likely to be reflective of the achievement effects that would be

experienced by schools and districts when the intervention is adopted and implemented under typical conditions.

Data

Implementation of the benchmark assessments, coupled with assistance in interpreting the resulting data, was ultimately intended to drive improvement in student performance on the state accountability tests. As a result, the outcome measure in this analysis is school-level performance on state-administered achievement tests.[6] Using state-administered achievement tests as an outcome measure has a unique set of advantages and drawbacks (May, Perez-Johnson, Haimson, Sattar, & Gleason, 2009). The main advantage of using state assessments is the fact that nearly every student is tested at state expense, and grade- or school-level data are generally made publicly available; these features serve to limit the cost of conducting a large-scale experimental study such as this one. Indeed, the achievement data used to construct the outcome measure in this analysis were collected from state departments of education by CDDRE personnel.

Any analysis that uses state-administered achievement tests as an outcome measure must assess whether such tests can provide valid and reliable information about the effects of the intervention. Because the 4Sight benchmark assessments were developed in a state-specific manner and were constructed from the same assessment blueprints as the respective state assessments, we believe that state-administered achievement tests represent not only an appropriate outcome measure but the ideal outcome measure.

One potential disadvantage of using state-administered assessments as the outcome measure is the fixed nature of assessment dates; researchers cannot schedule the assessment to be administered directly on completion of the intervention. For six of the seven states in our analysis—Alabama, Arizona, Mississippi, Ohio, Pennsylvania, and Tennessee—this issue presented no meaningful complications. These states administered their achievement tests in the spring, which meant that data reviews and associated training sessions were conducted, and benchmark assessments were administered, after the administration of the baseline assessment but prior to administration of the outcome assessment.[7] Indiana, however, administers its achievement tests in the fall, a fact that introduces some potential complications into the analysis, particularly with respect to the outcome measure.[8] In the study design, the 2007–2008 school year was identified as the outcome year for the three participating Indiana school districts. However, using fall 2007 achievement results as the outcome measure means that full implementation of the intervention had not occurred in the treatment districts prior to administration of the outcome assessment. The most obvious response to this issue, using fall 2008 achievement results as the outcome measure, allows the full implementation of the intervention in treatment districts but introduces the complication that control districts experienced a partial implementation of the intervention prior to administration of the outcome assessment. Given the inability to conduct the ideal analysis, we use achievement results from fall 2007 as the outcome measure in our primary analysis but conduct a sensitivity analysis in which all Indiana schools are excluded from the analysis.[9]

A second potential disadvantage associated with using state-administered assessments as the outcome measure stems from concerns of comparability. The use of state-administered achievement tests can create complications when attempting to analyze and compare achievement data across grades, subjects, and especially states. To facilitate such comparisons in this analysis, we followed the guidance of May et al. (2009) and transformed all achievement data into *z* scores. Because achievement data were obtained in three different formats across the seven states, the process of generating *z* scores was not identical across the states.

Five states—Arizona, Indiana, Mississippi, Ohio, and Pennsylvania—publish, for each school in the state, the average scale score for each grade in both reading and math. After collecting these data for all schools in our sample, we obtained the respective statewide means and student-level standard deviations in both reading and math for each grade from state departments of education or the relevant technical reports. Then, for each school's grade-level mean scale score, we subtracted the appropriate statewide mean and divided by the standard deviation to obtain a school-specific grade-level *z* score. We then created a school-level *z* score by using each school's grade-level enrollment data to compute a weighted

average of all available grade-level *z* scores. Because we analyzed math and reading separately, we created a school-level *z* score for each subject.

For Tennessee and Alabama, the computation of subject-specific school-level *z* scores followed somewhat different procedures. The Tennessee Department of Education does not publish average grade-level scale scores by school, but it does provide school-level achievement expressed as a normal curve equivalent (NCE) score. Using standard methods, we transformed the school-level NCE score into a *z* score.[10] For Alabama, the state department of education publishes, for each school, grade-level percentiles of performance on the SAT-10, which is the commercial test on which the Alabama Reading and Mathematics Test is based.[11] Again, we used standard methods to transform the school-level percentiles into *z* scores.

Our analysis also includes a baseline measure of school-level achievement. We used a baseline achievement measure to improve the precision, and corresponding power, of the estimated impacts (Bloom, 2005; Raudenbush et al., 2007). Like the outcome measure, the school-level baseline achievement measure takes the form of a *z* score. The procedures used to calculate the baseline achievement measure were identical to those used to calculate the outcome measure and are described above. In addition to a baseline achievement measure, our analysis also contains baseline measures of several school- and district-level demographic characteristics that may further improve the precision of our impact estimates. At the school level, we included measures of the percentage of students who are racial minorities and the percentage of students eligible for free- or reduced-price lunch. At the district level we also included measures of the percentage of minority students and the percentage of students eligible for free or reduced-price lunch, and we took into account the percentage of students receiving special education services as a baseline covariate at this level.

Missing Data Procedures and Final Analytic Sample

School-level achievement measures were not available for some of the schools in our sample. In reading, we were not able to collect school-level achievement outcomes for 25 schools, which represent less than 5% of the 549 schools in our sample. These 25 schools were spread across 14 districts; 8 of the schools were located in districts that had been assigned to the treatment group, while 17 of the schools were located in districts that had been assigned to the control group. We excluded the 25 schools missing school-level achievement outcome data from our reading analysis. Given the relatively small number of schools missing data, this decision did not compromise the integrity of our analysis.

In reading, a small number of schools for which outcome data were collected were missing baseline achievement or free-lunch eligibility data. More specifically, there were five schools missing baseline achievement data and three schools missing free-lunch eligibility data. For these schools, we followed the guidance that Puma, Olsen, Bell, and Price (2009) provided on the topic of missing data in cluster randomized trials. In particular, we assigned the schools the average district value for the missing measure, either baseline achievement or free-lunch eligibility, and included a dummy variable in the analytic model.

As noted earlier, both Ohio districts contacted by CDDRE agreed to participate in the reading portion of the research project but not the mathematics portion. As a result, our math analysis contains schools from 57 districts. As was the case with reading, school-level achievement outcomes were not available for a small number of schools in our sample. Specifically, such data were not available for 24 of the 538 schools that had been selected by district officials to administer benchmark assessments in mathematics. Seven of these schools were located in districts that had been assigned to the treatment group, while the remaining 17 schools were in districts that had been assigned to the control group. We excluded the 24 schools missing school-level mathematics achievement outcome data from our math analysis. As was the case in reading, for the small number of schools missing baseline achievement data (3 schools) or free-lunch eligibility data (3 schools), we assigned the average district value for the missing measure and included a dummy variable in the analytic model.

After implementing these missing data procedures, our final estimation sample for the reading analysis was composed of 524 schools located in 59 districts. For math, the final estimation sample consisted of 514 schools located in 57 districts.

Analytic Framework

In this cluster randomized trial, randomization took place at the district level while outcome data were collected at the school level. In such designs, correct estimation of treatment effects is at the level that the cluster was randomized (Bloom, 2005; Raudenbush, 1997). Although we have school-specific achievement measures, analysis of treatment effects at this level of aggregation will produce artificially low standard errors and thus overly precise impact estimates. We perform our impact analyses within a multilevel modeling framework, which prior work has shown to be an appropriate method for analyzing data from cluster randomized trials (Raudenbush, 1997).[12] A main advantage of this approach is the fact that variability in the outcome measured can be partitioned, and modeled, at multiple levels. In the case of this analysis, multilevel modeling allowed us to model variability from both school- and district-level factors in the statistical model we used to estimate the effect of being assigned to administer benchmark assessments on average school-level achievement.

The fully specified linear model for the school level, or Level 1, of the analysis can be written as follows:

$$Y_{ij} = \beta_{0j} + \beta_1(\text{Base Ach.})_{ij} + \beta_2(\text{FRPL})_{ij} + \beta_3(\text{Pct. Min.})_{ij} + \beta_4(\text{Base Ach. Miss.})_{ij} + \beta_5(\text{FRPL Miss.})_{ij} + \varepsilon_{ij},$$

where Y represents achievement expressed as a z score, and i and j are index schools and districts, respectively. In addition, β_0 represents the intercept for mean district achievement, and the model contains school-level measures of baseline achievement (Base Ach.), the percentage of students eligible for free or reduced-price lunch (FRPL), and the percentage of students who are minorities (Pct. Min.). This level of the model also contains dummy variables for schools that were missing (Miss.) measures of baseline achievement or the percentage of students eligible for free or reduced-price lunch. Finally, the model contains a school-level residual, represented by ε_{ij}.

The district-level portion of the model, or Level 2, is slightly different from the school-level portion of the model. At Level 2, the intercept for mean district achievement is modeled as a function of a grand mean, treatment status, baseline district-level demographics, a vector of randomization blocks, and a district-level residual. More formally, the district-level portion of the model can be written as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Treat.})_j + \gamma_{02}(\text{Pct. Min.})_j + \gamma_{03}(\text{FRPL})_j + \gamma_{04}(\text{Pct. Sp. Ed.})_j + \delta(\text{Rand. Block})_j + \tau_j.$$

Variability attributable to the randomization blocks could be addressed using either a random-effects approach or a fixed-effects approach. In a random-effects approach, the randomization blocks would be specified as a third level of the multilevel model. In this conception, schools would be nested within districts, which would in turn be nested within randomization blocks.[13] The fixed-effects approach, which is the strategy we use, accounts for variability attributable to randomization blocks by including dummy variables at the district level, which is the level at which randomization occurred. According to Schochet (2008), the fixed-effects approach is most realistic in the majority of evaluations of educational interventions. The random-effects approach is appropriate if the analytic sample is representative of some broad, well-defined population, but this is rarely the case in evaluations of educational interventions. It is certainly not the case in this analysis, where officials within state departments of education nominated districts for participation in the CDDRE study and district officials identified the specific schools that would be part of the research project. As a result, we believe that including fixed effects for randomization blocks in our analytical model is the proper approach in the context of this evaluation.

Results

Prior to estimating the fully specified model written above, we first estimated an unconditional model to determine how much variation in average school-level achievement was attributable to school-level factors and district-level factors. That is, we estimated the following model separately for reading and math:

$$Y_{ij} = \gamma_{00} + \varepsilon_{ij} + \tau_j.$$

The results of the unconditional model indicated that in reading, approximately 57% of the variation

TABLE 2
Multilevel Models Predicting Average School Math Score

Variable	Model 1	Model 2	Model 3	Model 4
Fixed effect				
Treatment		.002	.059**	.059**
		(.072)	(.026)	(.029)
Baseline score: school			.885***	.803***
			(.024)	(.028)
% minority: school				–.002***
				(.001)
% FRPL: school				–.0008*
				(.0004)
% minority: district				–.0001
				(.001)
% FRPL: district				.001
				(.001)
% special education: district				.001
				(.005)
Baseline missing			.152*	.202*
			(.091)	(.121)
FRPL missing				–.114
				(.120)
Random effect				
District (intercept)	.073	.057	.006	.006
	(.016)	(.015)	(.002)	(.002)
Residual	.080	.080	.023	.022
	(.005)	(.005)	(.002)	(.001)
Model statistics				
n	514	514	514	514
Number of groups	57	57	57	57
Wald χ^2	NA	18.60**	1,573.04***	1,650.53***

Note. FRPL = free or reduced-price lunch; NA = not available. Values in parentheses are standard errors. Fixed effects for randomization blocks included in the analytic model are not shown but are available upon request.
*$p < .10$. **$p < .05$. ***$p < .01$.

in average school-level achievement was attributable to school-level factors; the remaining 43% of variation was attributable to district-level factors. The model returned similar results for math, with approximately 52% of the variation in average school-level achievement attributable to school-level factors and 48% attributable to district-level factors. In addition to allowing for the variance to be partitioned across the two levels of the model, estimation of the unconditional model also provided estimates of the total amount of variability, which are necessary for calculating the proportion of variance explained by the covariates included in the model. Full results of the unconditional model can be found in column 1 of Tables 2 and 3.

After estimating the unconditional models, we turned to estimating a model containing the treatment indicator and fixed effects for the randomization blocks. The results of this specification are presented in the second results column in Tables 2 and 3. In this specification, which contained no baseline covariates, the treatment effects were estimated with little precision and no statistically significant effects were observed.

In an effort to gain precision in the impact estimates, our third specification used a measure of baseline achievement. The inclusion of this pretreatment covariate increased precision substantially; the standard errors for the impact estimates decreased by over 50%. In addition, we detected

TABLE 3
Multilevel Models Predicting Average School Reading Score

Variable	Model 1	Model 2	Model 3	Model 4
Fixed effect				
Treatment		–.044	.027	.033
		(.062)	(.020)	(.020)
Baseline score: school			.893***	.802***
			(.025)	(.032)
% minority: school				–.0013**
				(.0006)
% FRPL: school				–.0009**
				(.0005)
% minority: district				.0001
				(.001)
% FRPL: district				.0001
				(.001)
% special education: district				.004
				(.004)
Baseline missing			.109	.142*
			(.073)	(.085)
FRPL missing				–.135
				(.109)
Random effect				
District (intercept)	.056	.042	.002	.006
	(.013)	(.011)	(.001)	(.002)
Residual	.073	.073	.025	.022
	(.005)	(.005)	(.002)	(.001)
Model statistics				
n	524	524	524	524
Number of groups	59	59	59	59
Wald χ^2	NA	22.36**	1,526.66***	1,621.62***

Note. FRPL = free or reduced-price lunch; NA = not available. Values in parentheses are standard errors. Fixed effects for randomization blocks included in the analytic model are not shown but are available upon request.
*$p < .10$. **$p < .05$. ***$p < .01$.

a positive, statistically significant treatment effect for the model of average school-level mathematics achievement. Assignment to receive CDDRE services was estimated to increase average mathematics achievement by approximately 0.06 student-level standard deviations. No statistically significant treatment effects were observed for reading.

Our fully specified model contained pretreatment measures of school- and district-level demographics in addition to the covariates included in previous specifications. The results of the fully specified model are presented in column 4 of Tables 2 and 3. The inclusion of school- and district-level baseline demographic characteristics did little to change the point estimates of the treatment effects, or the precision of these estimates.[14] The data-driven reform model was again found to have a positive effect on mathematics achievement; assignment to the treatment group was estimated to increase average achievement by approximately 0.06 student-level standard deviations.[15] This estimate is statistically significant at the $p < .05$ level. In reading, the coefficient on the treatment indicator in our fully specified model was positive at 0.033, but it did not reach a conventional level of statistical significance ($p < .10$). Taken together, the results indicate that district-level assignment to implement a data-driven reform initiative can

cause increased achievement, particularly in mathematics.[16]

Discussion

Having established that benchmark assessments have a positive effect on mathematics, and possibly reading, achievement, we next turn our attention to assessing the magnitude of these effects. One standard method for expressing the magnitude of an effect is through the use of effect sizes, a method that transforms the regression coefficients into standard deviation units. In one sense, the coefficients on the treatment variables are already presented as effect sizes; the method we used to create our dependent variable permits the coefficients to be interpreted in student-level standard deviation units. However, interpreting district-level achievement impacts in student-level standard deviation units is not entirely straightforward. Is a 0.06 student-level standard deviation increase in average district-level achievement substantively meaningful? An effect of this magnitude would be considered quite small in the context of an analysis in which student-level achievement served as the outcome measure. However, district-level achievement is substantially less variable than student-level achievement, which means that the coefficient estimate of 0.06 has the potential to be substantively meaningful. The absence of other trials that have used analyses with similar features—district-level randomization with an outcome measure of school-level achievement standardized by student-level standard deviations—has resulted in the lack of a widely recognized benchmark against which our results can be compared.

Hedges (2009) presents additional methods for calculating effect sizes in the context of a cluster randomized trial such as this one. The first approach involves calculating the ratio of the estimated treatment effect and the between-cluster variability. In math, the results of the unconditional model indicate that the standard deviation for the district-level random effect is 0.269. Dividing the point estimate of the treatment effect by this standard deviation reveals that the effect of benchmark assessments in math is equivalent to an effect size of approximately 0.21. In reading, the estimated treatment effect corresponds to an effect size of about 0.14. A second approach involves calculating the ratio of the estimated treatment effect and the estimated within-cluster variability. In the case of this analysis, the estimate of within-cluster variability is very similar in magnitude to the estimate between-cluster variability. As a result, the within-cluster effect sizes are very similar to the between-cluster effect sizes. Specifically, the effect size using this approach is 0.201 in math and 0.12 in reading. When interpreting these effect sizes, it is important to recall that the estimates of between- and within-cluster variability are specific to the sample of schools and districts that serve as the basis of this analysis. As a result, it is unclear whether the effect sizes based on these estimates are generalizable to other samples or populations of interest. However, as noted above, our sample is diverse in many respects, which provides some basis for believing that these effect size estimates may be broadly representative of effect sizes that might be obtained from other samples or populations of interest.

Given the unique features of this analysis—district-level randomization with an outcome measure of school-level achievement standardized by student-level standard deviations—we felt that it was important to present our estimates in multiple contexts. Each of the three effect size estimates presents the results in a distinct light, and each is distinctly informative. Taken as a whole, we believe the results illustrate that data-driven reform efforts can have not only a statistically significant effect on achievement but a substantively meaningful impact as well.

Indeed, the effects do appear meaningful, but what mechanisms might explain them? First, it could be the case that the observed treatment effects are the result of a practice effect. Dozens of studies have illustrated that repeated exposure to a test of cognitive ability can result in increased scores (see Kulik, Kulik, & Bangert, 1984, and Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007, for meta-analytic reviews of this research). In the context of the intervention analyzed in this study, students were administered one to four benchmark assessments very similar in nature to the state-administered accountability assessments, which constituted the outcome measures. The practice and experience that students gained from the administration of the benchmark assessments could have helped them prepare for, and ultimately achieve, higher scores on the end-of-year state-administered achievement tests.

Alternatively, it is possible that the benchmark assessments could have familiarized teachers with the content of the state assessments and made them more aware of the impending high-stakes tests. In response to this increased awareness, teachers may have altered their instruction to more closely align it with the state assessment. That is, the benchmark assessments could have induced teachers to "teach to the test" to a greater extent than they had previously. Several studies have suggested that teachers and administrators use data to alter their instruction in a manner that is primarily intended to produce improved results on the state accountability tests (Diamond & Cooper, 2007; Diamond & Spillane, 2004). Although such practices have been criticized for several reasons, including their potential for narrowing and fragmenting the curriculum, they could result in improved student performance on state assessments.

Finally, the benchmark assessments may have facilitated awareness among teachers and other school staff members of particular areas of weakness exhibited by specific students, classrooms, or schools. As a result of this knowledge, teachers may have been able to more effectively target instruction and other supplemental educational services to the areas and students in need of greater assistance. These improved educational practices could have increased student cognitive ability and this improved cognitive ability may have then been reflected by better performance on state assessments. Such a mechanism is aligned with the CDDRE program theory, as well as the theory underlying the concept of benchmark assessments more generally (Kennedy, 2003; Perie et al., 2009; Schmoker, 2003).

That being said, any or all of the three mechanisms discussed above could be responsible for the positive effects produced by the 1st year of the CDDRE intervention; the data available for this analysis are not sufficiently rich to adjudicate among the multiple viable explanations. Future work attempting to disentangle this issue would serve as a valuable extension to the present study. Such work could be done using observational methods, surveys, teacher logs, or several other approaches to get inside the "black box" and untangle the various causal mechanisms that may be at play.

Using a cluster-randomized design, this study presents the results of one of the first large-scale efforts to assess the causal effects of a data-driven reform on achievement outcomes. Furthermore, although cluster-randomized designs are becoming increasingly common for evaluating the effects of educational interventions, this is the first known educational evaluation in which school districts served as the unit of randomization. The randomization of entire school districts has several implications, with perhaps the most important one being the mitigation of partial equilibrium concerns; several educational interventions have been found to be effective in small-scale efficacy trials but are later found to produce no positive impacts when they are evaluated on a larger scale. By randomizing nearly 60 school districts, the results presented here are somewhat insulated from such concerns. The external validity of these results is further enhanced by the fact that the study design incorporates districts from seven states. Such a design provides this study with a relatively large sample size, and subsequently greater power to detect effects, but it also allows us to gain a sense of the effect of scaling up this intervention across districts that are diverse in many political, geographic, and socioeconomic contexts. Smaller sample sizes and narrower scope have prevented authors of most prior studies on the topic from being able to make such generalizations.

Given the delayed treatment design used in this study, the results presented above represent pure experimental intention-to-treat impacts. Previous empirical work has provided suggestive evidence that data-driven reform can produce improved achievement outcomes, but these earlier studies were either somewhat underpowered (Henderson et al., 2007) or focused on the evaluation of a pilot program (May & Robinson, 2007). This study provides the best evidence to date that data-driven reform efforts, implemented at scale, can result in substantively and statistically significant improvements in achievement outcomes.

Notes

1. For a description of this larger study, please see Slavin et al. (2010).

2. Notably, all districts nominated by the state departments of education and recruited by CDDRE personnel agreed to participate in at least the reading and language arts portion of the study. The success in securing district participation, which is at least partially attributable to

the substantial control that districts had over many aspects of the intervention, bodes well for the generalizability of the results.

3. The results of these two-tailed *t* tests are not shown but are available upon request. In addition to conducting *t* tests at the district level, we also estimated a multilevel logistic regression (schools nested within districts) in which a school's treatment status was modeled as a function of several pretreatment characteristics, including baseline achievement, percentage White, percentage Black, percentage Hispanic, and percentage eligible to receive free or reduced-price lunch. The results, which are available upon request, show that none of the baseline characteristics included in the model is a statistically significant predictor of treatment status. The results of this regression provide further evidence that the randomization procedure succeeded in producing balanced treatment and control groups.

4. Documentation indicates that benchmark assessments were administered prior to the implementation of the CDDRE intervention in four districts. Two of these districts—Anderson CS, Indiana (15 schools), and Phoenix, Arizona (17 schools)—were in the control group, and two—Michigan City, Indiana (13 schools), and Richmond, Indiana (12 schools)—were members of the treatment group. CDDRE personnel indicate that these districts were the only ones using data-driven reform practices prior to implementation of the intervention.

5. Under the CDDRE model, the review of state test data occurs prior to the implementation of benchmark assessments. This provides teachers, principals, and other personnel with the skills and experience to review the data generated by the benchmark assessments.

6. The specific tests used as the outcome measures are as follows: Alabama: Stanford Achievement Test–10 (SAT-10); Arizona: Arizona Instrument to Measure Standards; Indiana: Indiana Statewide Testing for Educational Progress–Plus; Mississippi: Mississippi Curriculum Test 2; Ohio: Ohio Achievement Test; Pennsylvania: Pennsylvania System of School Assessment; and Tennessee: Tennessee Comprehensive Assessment Program.

7. The majority of treatment districts had their initial meetings with CDDRE personnel prior to the administration of the baseline assessment. These meetings, however, were purely organizational in nature and unlikely to have any effect on baseline achievement. All substantive components of the intervention took place after the administration of the baseline assessment. Similarly, the majority of control districts had their initial meetings with CDDRE personnel prior to the administration of the outcome assessment. Again, though, these meetings were organizational in nature and unlikely to have any effect on achievement. To the extent that these slight inconsistencies between treatment delivery and achievement measures had any effect on this analysis, they would depress the estimated treatment effects.

8. All Indiana districts participating in the study were part of the third cohort, which was designed to have a baseline year of 2006–2007 and an outcome year of 2007–2008. Using achievement results from fall 2006 as our baseline achievement measure presented no problems, because all data reviews and associated training sessions were conducted, and benchmark assessments were administered, after fall 2006.

9. By using an assessment administered prior to full implementation as the outcome measure, treatment effect estimates are likely to be depressed, a conjecture that is tested empirically in the sensitivity analysis. Two main reasons underlie our decision to use achievement results from fall 2007 as the outcome measure in our primary analysis. First, it is consistent with the other states, where the baseline measure of achievement was uncontaminated by implementation in the control group. Second, CDDRE staff members began implementing the intervention in Indiana districts in February 2007, earlier than implementation occurred in other treatment districts. As a result, partial implementation of the intervention had taken place by the time the outcome assessments were administered in fall 2007.

10. The standard method for transforming an NCE into a *z* score involves subtracting 50, which is the definitional mean of an NCE, and dividing by 21.06, which is the definitional standard deviation. In this case, the NCE was not anchored to the most recent school year. As a result, the statewide mean was not 50 but was 58 in math and 57 in reading. As a result, when creating the *z* scores, we subtracted the statewide mean and divided by 21.06. In effect, we treat the school-level NCE as if it is centered on the empirical statewide mean, as opposed to 50.

11. The Alabama Reading and Mathematics Test is the state-administered achievement test that Alabama uses for accountability purposes. It is based on the SAT-10 but includes additional items that allow it to more closely reflect Alabama's curricular standards. The school-level percentiles for the SAT-10 are relative to the national population of SAT-10 test takers, but the Alabama average is very close to the national average for all grades in both reading and math.

12. This modeling approach is often referred to as hierarchical linear modeling in the education literature and mixed-effects modeling in the econometrics literature.

13. Recall that randomization blocks are defined as each cohort-state combination. In the reading analysis, there are 10 randomization blocks: Alabama 1, Arizona 1, Ohio 1, Pennsylvania 1, Pennsylvania 2, Arizona 3, Indiana 3, Mississippi 3, Pennsylvania 3, and Tennessee

3. In the math analysis, there are 9 randomization blocks: all of those listed in the previous sentence except Ohio 1.

14. The baseline covariates in the fully specified model behave largely as expected. Baseline achievement is an extremely strong predictor of the outcome measure. However, even with a baseline achievement measure included in the model, the percentage of minority students and the percentage of students eligible for free or reduced-price lunch are estimated to be statistically significant predictors of the outcome measure.

15. Implicit in this analysis is the assumption that treatment assignment did not spur within-district mobility in a manner that changed the composition of participating schools' student or teacher populations. Because only a subset of schools in most districts participated in the intervention, it is possible that students with certain characteristics would choose to transfer into participating schools, which could result in a biased estimate of the treatment effect. For example, if socioeconomically advantaged, high-achieving students transferred into schools using the intervention, the treatment effect estimates would be upwardly biased. To investigate the plausibility of such concerns, we performed two analyses. First, we replicated Table 1 for the treatment year. That is, we calculated the average demographic characteristics in the treatment year for participating schools in each treatment and control district and performed *t* tests. These *t* tests revealed no statistically significant differences between the treatment and control groups on any of the demographic measures. Second, we estimated multilevel regressions (schools nested within districts) in which each demographic characteristic in Table 1 was modeled as a function of treatment status. The results of these regressions show that treatment status is not a statistically significant predictor of any treatment-year school-level demographic characteristic. The results of these two analyses, which are available upon request, provide strong evidence that treatment assignment is not inducing within-district mobility in a manner that alters the composition of participating schools' student or teacher populations and thus biases the estimated treatment effects.

16. The results of the sensitivity analysis that excludes Indiana schools from the estimation sample are substantively similar to the primary results. As predicted, exclusion of the Indiana schools results in a slight increase in the estimated treatment effect in math. Full results of the sensitivity analysis are available upon request.

References

Bernhardt, V. L. (2003). *Using data to improve student learning in elementary schools*. Larchmont, NY: Eye on Education.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*, 7–74.

Bloom, H. S. (Ed.). (2005). *Learning more from social experiments: Evolving analytic approaches*. New York: Russell Sage.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*, 231–268.

Christman, J., Neild, R., Bulkley, K., Blanc, S., Liu, R., Mitchell, C., & Travers, E. (2009). *Making the most of interim assessment data: Lessons from Philadelphia*. Philadelphia, PA: Research for Action.

Clune, W. H., & White, P. A. (2008). *Policy effectiveness of interim assessments in Providence public schools* (WCER Working Paper No. 2008-10). Madison: University of Wisconsin–Madison, Wisconsin Center for Education Research.

Conrad, W. H., & Eller, B. (2003, April). *District data-informed decision making*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Crocco, M. S., & Costigan, A. T. (2007). The narrowing of curriculum and pedagogy in the age of accountability: Urban educators speak out. *Urban Education*, *42*, 512–535.

Crooks, T. J. (1988) The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*, 438–481.

Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data: How high-performing school systems use data to improve instruction for elementary students*. Los Angeles: Center for Educational Governance, University of Southern California.

Dembosky, J. W., Pane, J. F., Barney, H., & Christina, R. (2005). *Data-driven decision making in southwestern Pennsylvania school districts* (WR-326-HE/GF). Santa Monica, CA: RAND Corporation.

Diamond, J. B., & Cooper, K. (2007). The uses of testing data in urban elementary schools: Some lessons from Chicago. *Yearbook of the National Society for the Study of Education*, *106*, 241–263.

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record*, *106*, 1140–1171.

Goertz, M. E., Olah, L. N., & Riggan, M. (2009). *From testing to teaching: The use of interim assessments in classroom instruction* (Research Report #65). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*, 373–385.

Hedges, L. V. (2009). Effect sizes in nested designs. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 337–355). New York: Russell Sage.

Heilig, J. V., & Darling-Hammond, L. (2008). Accountability Texas-style: The progress and learning of urban minority students in a high-stakes testing context. *Educational Evaluation and Policy Analysis*, *30*, 75–110.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (Issues & Answers Report, REL 2007 No. 039). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands.

Herman, R., Dawson, P., Dee, T., Greene, J., Maynard, R., Redding, S., & Darwin, M. (2008). *Turning around chronically low-performing schools: A practice guide* (NCEE #2008-4020). Washington DC: Institute of Education Sciences, U.S. Department of Education.

Kennedy, E. (2003). *Raising test scores for all students: An administrator's guide to improving standardized test performance*. Thousand Oaks, CA: Corwin.

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*, 435–447.

Marsh, J. A., Pane, J. F., & Hamilton, S. (2006). *Making sense of data-driven decision making in education: Evidence from recent RAND research*. Santa Monica, CA: RAND Corporation.

Mason, S. (2002). *Turning data into knowledge: Lessons from six Milwaukee public schools*. Madison: Wisconsin Center for Education Research.

May, H., & Robinson, M.A. (2007). *A randomized evaluation of Ohio's Personalized Assessment Reporting System (PARS)*. Madison, WI: Consortium for Policy Research in Education.

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using state tests in education experiments: A discussion of the issues* (NCEE 2009-013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist*, *22*, 155–175.

Ogawa, R. T., Sandholtz, J. H., Martinez-Flores, M., & Scribner, S. P. (2003). The substantive and symbolic consequences of a district's standards-based curriculum. *American Educational Research Journal*, *40*, 147–170.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*, 5–13.

Petrides, L., & Nodine, T. (2005). *Anatomy of school system improvement: Performance-driven practices in urban school districts*. San Francisco, CA: Institute for the Study of Knowledge Management in Education and New Schools Venture Fund.

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

Quint, J., Sepanik, S., & Smith, J. (2008). *Using student data to improve teaching and learning: Findings from an evaluation of the Formative Assessments of Students Thinking in Reading (FAST-R) program in Boston elementary schools*. New York: MDRC.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173–185.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*, 5–29.

Schmoker, M. (2003). First things first: Demystifying data analysis. *Educational Leadership*, *60*, 22–24.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*, 62–87.

Slavin, R. E., Holmes, G., Madden, N. A., Chamberlain, A., Cheung, A., & Borman, G. D. (2010). *Effects of a data-driven district-level reform model,* Working Paper. Baltimore, MD: Center for Data-Driven Reform, Johns Hopkins University.

Snipes, J., Doolittle, F., & Herlihy, C. (2002). *Foundations for success: Case studies of how urban school systems improve student achievement*. Washington, DC: Council of the Great City Schools.

Stecher, B., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., McCombs, J. S., . . . Naftel, S. (2008). *Pain and gain: Implementing No Child Left Behind in three states, 2004–2006*. Santa Monica, CA: RAND Corporation.

Stecher, B. M., & Hamilton, L. S. (2006). *Using test-score data in the classroom* (WR-375-EDU). Santa Monica, CA: RAND Corporation.

Wright, W. E., & Choi, D. (2006). The impact of language and high-stakes testing policies on elementary school English language learners in Arizona. *Education Policy Analysis Archives*, *14*, 1–75.

Authors

DEVEN CARLSON is a Ph.D. Candidate in the Department of Political Science and a graduate research fellow at the Wisconsin Center for Education Research at the University of Wisconsin-Madison. His research interests include education policy, social policy, and policy analysis.

GEOFFREY D. BORMAN is Professor of Education and Sociology at the University of Wisconsin-Madison. His areas of research include experimental and quasi-experimental design, educational policy, and educational inequality.

MICHELLE ROBINSON is a graduate student in the Department of Sociology and a graduate research fellow at the Wisconsin Center for Education Research at the University of Wisconsin-Madison. Her research interests include: urban education reform, institutional barriers to educational access by disadvantaged communities, the effects of inequality on family functioning, and family-school cultural mismatch.

Manuscript received July 20, 2010
Revision received December 12, 2010
Accepted May 15, 2011