Evaluation of Michigan's First Cohort of 1003(g) School Improvement Grantees

Outcomes After the First Year

Submitted to:

Linda Forward

Office of Education Improvement and Innovation

Michigan Department of Education

Submitted by:

Jonathan Nakamoto

John Rice

Juan Carlos Bojorquez

WestEd

January, 2013

WestEd

Table of Contents

Executive Summary i

Introduction 1
School Improvement Grants in Michigan 1
Overview of Michigan's 2010 SIG-I Grantees 2
Overview of the Evaluation of Michigan's SIG-I Schools 2

Method 4
Overview 4
Study Timeline 4
Measures 5
Michigan Educational Assessment Program (MEAP) 5
Michigan Merit Examination (MME) 5
Student Demographic Data 6
Identification of the Comparison Group 6
Identification of the Students Included in the Analyses 8
Data Analyses 8
Sensitivity Analysis 11

Findings 12
SIG-I and Comparison Groups Prior to SIG Implementation 12
Student Achievement on the MEAP and MME 12
Student Demographic Characteristics 13
Impact of SIG at Elementary and Middle Schools 14
Impact of SIG at High Schools 17
Conclusion and Next Steps 19

References 20

Executive Summary

This is the third report from the Evaluation Research Program at WestEd on findings from the evaluation of the Michigan 1003(g) School Improvement Grants (SIGs) that were awarded in fall 2010. The three-year evaluation is using both quantitative and qualitative approaches to determine the effectiveness of SIG and to provide information that can be used to improve and enhance SIG interventions. This report focuses only on academic outcomes and presents findings on the impact of SIG after the first year of the program (i.e., the 2010-11 school year).

WestEd used a quasi-experimental design to examine the impact of SIG on students' mathematics, reading, and science achievement across all 28 SIG-I schools after one year. Achievement was assessed using scale scores from the Michigan Educational Assessment Program (MEAP) and the Michigan Merit Examination (MME). WestEd identified a comparison group of 56 non-SIG-I schools and compared the achievement of the students in these schools with the achievement of the students in the 28 SIG-I schools. WestEd conducted the analyses with student-level outcome data using hierarchical linear modeling (HLM), which accounted for prior scores on these same assessments as well as student demographic characteristics.

There were no statistically significant differences between scale scores on the mathematics, reading, or science assessment between SIG-I and comparison schools after the first year of the SIG. This was the case when scores were examined separately for each assessed grade and when pooled across assessed grades in elementary and middle schools. Subsequent reports will assess the differences on these outcomes between SIG-I and comparison schools after years two and three of SIG.

Introduction

This is the third report from the Evaluation Research Program at WestEd on findings from the evaluation of Michigan's first cohort of 1003(g) School Improvement Grants (SIGs).[1,2] Specifically, this report presents findings from a quasi-experimental design of the impact of SIG after one year using a matched-group comparison design. Previous reports have described implementation findings from the first and second years of the grant. Future reports will discuss implementation findings from the third and final year, case studies of implementation at six schools implementing SIG, and program impacts after the second and third years.

The introduction of this report provides a brief overview of SIGs, background information on the recipients of Michigan's SIGs. The next chapter provides information on the method, including outcome measures, the selection of the comparison schools, and the statistical model used to estimate SIG impact. The final chapter discusses the findings.

SCHOOL IMPROVEMENT GRANTS IN MICHIGAN

In 2010, Michigan received the seventh-largest ARRA-supplemented SIG grant in the nation, $115 million. A portion of these funds, $86.25 million, was available for immediate grant awards (Scott, 2011).[3] There were 228 Tier I, II, and III schools in Michigan eligible to apply for the SIG grants in 2010, 108 of which were Tier I and II schools. These schools were deemed eligible based on their state testing data from the 2007-08 and 2008-09 school years (Hurlburt, Le Floch, Therriault, & Cole, 2011).

Eighty-four of the eligible schools applied for SIG funds. The Michigan Department of Education (MDE) used a variety of factors in reviewing applications. First, MDE reviewed student academic performance on the Michigan Educational Assessment Program (MEAP) and Michigan Merit Examination (MME) in the three preceding academic years. MDE also reviewed each school's results from the Comprehensive Needs Assessment tool, or similar instruments or analysis of school and student needs. Michigan required that applicants submit evidence of engagement on the part of community stakeholders, including parents, teachers' unions, and the local school board. The selected schools were expected to set rigorous, achievable goals to increase academic performance each year, and use interim assessments to provide regular achievement progress reports. For high schools, ACT's PLAN and EXPLORE were required for assessment of college readiness.

---

[1] Although the 1003(a) and 1003(g) SIGs are distinct and separate funding streams, the use of "SIG" in this report refers to the latter only.

[2] The focus of the evaluation and the current report is on the 28 schools in Michigan that received 1003(g) SIG funds beginning in fall 2010. Additional schools have received SIG awards since then; however, they are outside the scope of the current evaluation.

[3] The remainder of the $115 million was rolled into a subsequent round of SIGs that were awarded to additional schools in 2011. These schools are not included in this evaluation.

OVERVIEW OF MICHIGAN'S 2010 SIG-I GRANTEES

Of the 84 applicants in 2010, MDE awarded funds to 28 schools. Individual grants ranged in size from approximately $605,000 to $4.9 million to be expended over three years, with an average grant per school of $2,955,678. Nineteen SIG-I schools selected the transformation model, while nine opted for the turnaround model. Traditional public schools comprised 27 of the schools (in 17 districts) and one school was a charter school under a charter management company.[4] Six SIG-I schools were classified as urban fringe, five rural, and 17 as central city. Thirteen of the central city schools were in three districts. Exhibit 1 displays the distribution of SIG model by school grade level in Year 1, and shows that the largest number (13) of SIG-I schools were high schools that adopted the transformation model. Sixteen of the SIG-I schools were high schools (grades 9-12), seven were middle schools (grades 5-8, 6-8, or 7-8), and another five were elementary schools (grades K-6 or K-8).[5]

Exhibit 1. Distribution of SIG Models and Grades Served in SIG-I Schools in 2010-11

	SIG Model		
Grades Served	Transformation	Turnaround	Total
K-6; K-8	3	2	**5**
5-8; 6-8; 7-8	3	4	**7**
9-12	13	3	**16**
Total	**19**	**9**	**28**

OVERVIEW OF THE EVALUATION OF MICHIGAN'S SIG-I SCHOOLS

In July 2011, MDE contracted with WestEd to conduct an independent evaluation of the 28 SIG-I grants awarded in fall 2010. The evaluation, conducted over three years, is both formative, to provide information that can be used to improve and enhance the SIG interventions, and summative to determine the effectiveness of the transformation and turnaround models. Across the three years, the evaluation combines quantitative and qualitative approaches integrating on-site observations, interviews, surveys, and relevant student- and school-level outcomes for all districts and schools receiving SIG funding, as well as in-depth case studies of six SIG schools. In addition, the evaluation is using a matched-comparison group of schools that did not receive SIG funds in order to estimate the impact of the SIG. By the end of the three-year evaluation, the effectiveness of SIG will be documented, as well as how various components of the turnaround and transformation models are likely contributing to the overall success of the schools.

---

[4] After the first year of the grant, one additional public school converted to a charter school.

[5] Before the first and second year of the grant, the middle school that served 5-8 added K-4 students and was re-classified as an elementary school for the evaluation after year one. In addition, between the first and second year of the grant, one K-6 elementary school added grades 7-8, and was re-classified as a middle school for the evaluation. However, because the current report analyzes outcomes after year one, the school classifications based on grades served in year one were used for matching comparison schools.

The three primary research questions across the three-year evaluation are:

(1) How are the SIGs implemented at the school and district levels?

(2) Does receipt of SIG funding have an impact on outcomes for low-performing schools?

(3) How is implementation of the two SIG intervention models (including specific strategies within those models) related to improvement in outcomes for SIG-I schools?

This report addresses the second research question: *Does receipt of SIG funding have an impact on outcomes for low-performing schools?*

Method

OVERVIEW

WestEd used a quasi-experimental design to examine the impact of SIG after one year on students' mathematics, reading, and science achievement as assessed by the MEAP and MME. We identified a comparison group of non-SIG-I schools and compared the achievement of the students in these schools with the achievement of the students in the 28 SIG-I schools. WestEd conducted the analyses using hierarchical linear modeling (HLM) and accounted for measures of prior achievement and student demographic characteristics.

STUDY TIMELINE

The study timeline is outlined in Exhibit 2. The 28 SIG-I schools received their awards in fall 2010 and had their first year of implementation during the 2010-11 school year. The baseline or pretest administration of the MME, which measured the high school students' achievement prior the start of SIG implementation, occurred in March 2010. The MME results from March 2011 measured the high school students' achievement after nearly one year of SIG implementation. In addition, the baseline administration of the MEAP assessments took place in October 2010 and covered the content taught during the 2009-10 school year. The October 2011 MEAP assessments are based on the content covered during the first year of the SIG grant, during which there was variability across SIG-I schools in the level of implementation of their SIG plans.

Exhibit 2. Timeline of SIG-I Funding and Evaluation

Date	Event/Activity
June 2010	Michigan Department of Education (MDE) awarded SIG funds by the U.S. Department of Education
March 2010	Administration of the Michigan Merit Examination (MME; baseline measure for high schools)
Summer – fall 2010	28 schools received SIG-I awards
October 2010	Administration of the Michigan Educational Assessment Program (MEAP) assessments that covered content from the 2009-10 school year (baseline measure for elementary and middle schools)
2010-11 school year	Year one of SIG implementation in the 28 SIG-I schools
March 2011	Administration of the MME (posttest measure for high schools)
October 2011	Administration of the MEAP assessments that covered content from the 2010-11 school year (posttest measure for elementary and middle schools)

MEASURES

MDE provided WestEd with student-level MEAP, MME, and demographic data from the 2010-11 and 2011-12 school years. Additionally, WestEd downloaded school-level MME results for the 2009-10 school year that were publically available on the MDE website (www.michigan.gov/mde) and the MI School Data website (www.mischooldata.org).

MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM (MEAP)

The MEAP assessments are administered to students in grades 3-9 in the fall of each school year. There are five MEAP assessments: (1) mathematics, (2) reading, (3) science, (4) writing, and (5) social studies. For the current evaluation, WestEd utilized the mathematics and reading assessments completed by students in grades 3-8 and the science assessments completed by students in grades 5 and 8. The mathematics and reading assessments are based on the content covered in the previous school year. In contrast, the science assessments covered content from the prior two to three school years (Michigan Department of Education, n.d.).

All students receive a scale score and a performance level for each MEAP assessment. The scale scores are conversions of the assessments' raw scores that allow the results of the different test administrations (e.g., 2010-11 and 2011-12) to be compared even though new test items are used in the different administrations. The scale scores are divided into ranges that correspond to four performance levels: (1) *Advanced*, (2) *Proficient*, (3) *Partially Proficient*, and (4) *Not Proficient*. Students who score in the *Proficient* or *Advanced* levels are identified as "proficient" on each assessment (Michigan Department of Education, n.d.).

The reliability of the MEAP mathematics, reading, and science assessments was satisfactory. The Cronbach's alphas, which measured the assessments' internal consistency, ranged from 0.87 to 0.91 for mathematics, 0.83 to 0.87 for reading, and 0.85 to 0.89 for science. The content validity of the assessments (i.e., whether the assessments cover the standards they are intended to measure) has also been investigated. Panels of experts that reviewed the mathematics and reading assessments determined that the correspondence between the content of the assessments and the grade-level standards was acceptable. The panel of experts determined that the science assessment did not sufficiently address one of the five standards it was designed to assess (Michigan Department of Education, n.d.).

MICHIGAN MERIT EXAMINATION (MME)

The MME is administered in the spring of each school year to students in grade 11 and students in grade 12 who did not take the MME in grade 11 or previously received an invalid score on the MME. The MME assesses students' college- and career-readiness. The MME assesses the following five subjects: (1) mathematics, (2) reading, (3) science, (4) writing, and (5) social studies. WestEd used the mathematics, reading, and science scores of the students in grade 11 in the analyses for the current evaluation.

Consistent with the MEAP assessments, each student receives a scale score and one of the four performance levels for the MME mathematics, reading, and science. In 2010 and 2011, the MME mathematics, reading, and science showed good internal consistency reliability. The Cronbach's alphas ranged from 0.91 to 0.92 for mathematics, 0.88 to 0.89 for reading, and 0.87 to 0.88 for science. The MME showed high levels of convergent validity. In 2010 and 2011, the correlations between the MME mathematics and ACT mathematics was $r = .75$, the correlations between the MME reading and the ACT reading were $r = .86$ and $r = .87$, and the correlations between the MME science and the ACT science were $r = .77$ and $r = .78$ (Michigan Department of Education, 2011a; Michigan Department of Education, 2011b).

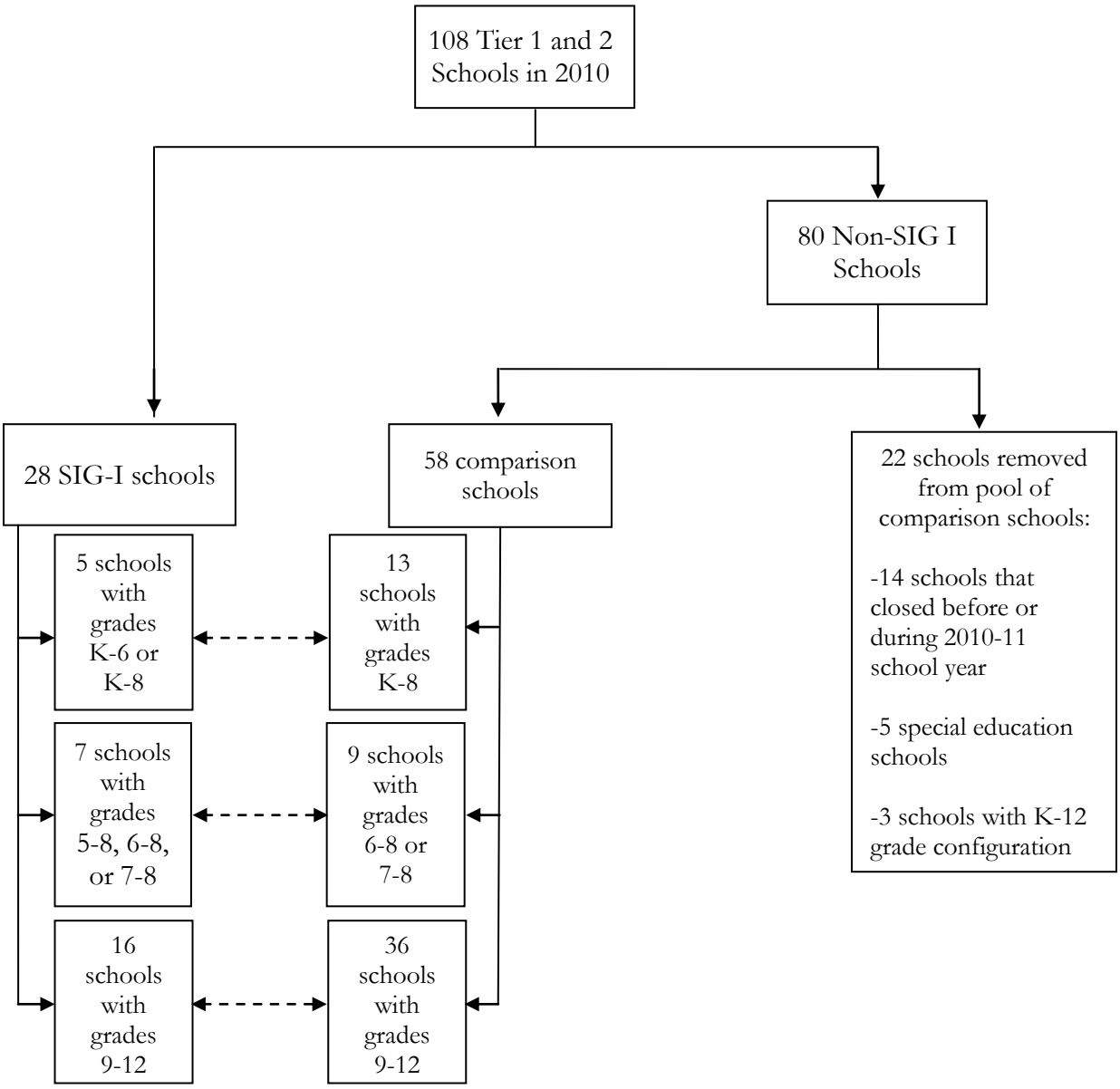
STUDENT DEMOGRAPHIC DATA

MDE provided data on the students' economic status, special education status, English proficiency status, gender, and race/ethnicity. In order to include these variables in our analyses, we created dummy codes for economic status (economically disadvantaged = 1; not economically disadvantaged = 0), special education status (special education = 1; not special education = 0), English proficiency status (limited English proficient = 1; not limited English proficient = 0), and gender (male = 1; female = 0). The race/ethnicity variable had codes for Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, Black/African American, Hispanic/Latino, White, Two or More Races, and Asian. Because Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, Two or More Races, and Asian constituted a total of less than eight percent of the SIG-I and comparison schools, we created an "Other" category by merging these racial/ethnic groups. We created a set of dummy codes that contrasted Black/African American (i.e., the largest racial/ethnic group) with Whites, Hispanics/Latinos, and the Other category.

IDENTIFICATION OF THE COMPARISON GROUP

Exhibit 3 contains a flow diagram summarizing the number of SIG-I ($n = 28$) and comparison ($n = 58$) schools that resulted from the pool of 108 Tier 1 and 2 schools identified in 2010. All Tier 1 and 2 schools were eligible to apply for the SIG program in 2010 because they were identified as the persistently lowest-achieving schools in the state. The comparison schools came from the pool of 80 schools that did not have their SIG application accepted or did not submit a SIG application. We excluded Tier 1 and 2 schools from the matching process if they closed before the beginning of or during the 2010-11 school year ($n = 14$), were special education schools ($n = 5$), or were K-12 schools/other grade configuration schools ($n = 3$). The five SIG-I elementary schools (i.e., K-6 or K-8 schools) had 13 comparison elementary schools (i.e., K-8 schools), the seven SIG-I middle schools (i.e., grades 5-8, 6-8, or 7-8) had nine comparison middle schools (i.e., 6-8 or 7-8), and the 16 SIG-I high schools (grades 9-12) had 36 comparison high schools (i.e., grades 9-12).

Exhibit 3. Flow Diagram Outlining the Identification of Comparison Schools for the 28 SIG-I Schools



Note: Dashed lines indicate where matching occurred.

Causal estimates about the impact of an educational program resulting from a quasi-experimental design will have the least amount of bias when a closely matched group of comparison schools is identified based on prior achievement and student demographics (Cook, Shadish, & Wong, 2008). Instead of matching schools one-to-one, WestEd used all of the remaining Tier 1 and 2 non-SIG-I as comparison schools. Although it would have been optimal to select the most comparable group of schools from the pool of 80 non-SIG-I schools, we did not do this for two reasons. First, even though all of the Tier 1 and 2 schools had low achievement, there were more SIG-I high schools

with relatively higher achievement than non-SIG-I high schools. As a consequence, there were not enough non-SIG-I high schools with prior achievement similar to the SIG-I high schools to use one-to-one matching. Second, biased results are likely to occur when the number of schools (i.e., level-2 units) included in HLM models is small (Maas & Hox, 2005). This would have been a particular problem with the small number of SIG-I schools that were elementary and middle schools. Including all of the non-SIG-I schools in the comparison group allowed our analyses to include the greatest number of schools.

IDENTIFICATION OF THE STUDENTS INCLUDED IN THE ANALYSES

The analyses with the MEAP assessment data included students in grades 4-8 that attended elementary SIG-I schools or comparison schools during the 2010-11 school year. Because not all of the SIG-I and comparison middle schools had students in grade 6, only students who attended grade 7 in the middle schools (and where therefore tested at the start of grade 8) were included in the MEAP analyses. A total of 5,708 students were included in the MEAP analyses. Students were removed from the analytic sample if they did not attend the same school for the full academic year (n = 1,191), did not have complete data on all of the variables included in the analyses (n = 371), or did not advance one grade level from 2010-11 to 2011-12 (n = 194).

Even though the students were tested on the MEAP in the fall of 2011 for the posttest, the student-level database used for the quasi-experimental design included a variable indicating which school they attended during 2010-11. This variable allowed us to attribute the students' achievement to the schools where they received their instruction in the prior year and is consistent with how Michigan does the calculations for adequate yearly progress (AYP).

The analyses with the MME assessment data included students in grade 11 who attended a SIG-I or comparison school during the 2010-11 school year. The MME analysis included a total of 5,431 students. We removed students from the analytic sample if they did not attend the same school for the full academic year (n = 1,642) or did not have complete data on all of the variables included in the analyses (n = 413).

DATA ANALYSES

We used HLM (Raudenbush & Bryk, 2002) to conduct the analyses for the quasi-experimental design with the MEAP and MME assessment data. The use of HLM is an improvement over other methods, such as multiple regression, because HLM appropriately accounts for the clustering of students within schools. The use of regression analysis with clustered data can produce inflated Type I error rates (i.e., identifying a statistically significant difference when one does not actually exist). In addition, HLM allows control variables at multiple levels (e.g., student demographic characteristics and prior school-level achievement) to be included properly.

The intraclass correlation coefficients (ICCs) based on the MEAP and MME assessments varied across the different subjects and grade levels. The ICCs ranged from .10 to .18 for the mathematics

assessments, .02 to .15 for the reading assessments, and .05 to .13 for the science assessments. With ICCs in these ranges, HLM is needed because the results based regression analyses are not likely to be accurate (Singer, 1998).

The HLM analyses were conducted using Stata's xtmixed command. Due to the small number of schools included in each analysis, we used restricted maximum likelihood estimation. With the exception of the invention status variable, we grand mean centered all of the predictor variables in the analyses. The centering allowed us to calculate the adjusted means for the SIG-I and comparison groups (Raudenbush & Bryk, 2002). We utilized two-level HLM models (i.e., level-1 = students; level-2 = schools). In these models, only the intercepts were specified as random effects. In order to reduce bias, we included pretests (e.g., prior mathematics scores) and demographic characteristics as covariates in the models (Shadish, Cook, & Campbell, 2002).

We conducted two sets of analyses with the MEAP assessment data. The first set of analyses was conducted separately with grades 4-8. We did not conduct analyses with the grade 3 data because no pretest mathematics and reading scores were available for grade 2. The analyses with the grades 4-8 data used the scale scores on the MEAP assessments. The second set of analyses combined all of the grade levels into a single analysis for the mathematics, reading, and science assessment data. In order to include the data from grades 4-8 in one analysis, we normalized the pretest and posttest MEAP scale scores within each grade by subtracting the state mean from each student's score and dividing by the state standard deviation. This method of standardization is equivalent to creating z-scores. After normalizing the scores, a score of -1.00 at each grade level was equivalent to a scale score that was one standard deviation below the state mean. The normalized test scores represent the relative performance rankings of students in the sample, rather than their absolute level of performance.

The two-level models that we employed for the MEAP analyses are outlined by the equations below using Raudenbush and Bryk's (2002) terminology and notations. The additional covariates used in the analyses that included multiple grade levels are noted below.

Level-1 model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{Prior Mathematics Score})_{ij} + \beta_{2j}(\text{Prior Reading Score})_{ij} + \beta_{3j}(\text{Demographic Covariate 1})_{ij} + \ldots + \beta_{Qj}(\text{Demographic Covariate Q})_{ij} + r_{ij}$$

Where

Y_{ij} was the average posttest mathematics, reading, or science score for student *i* in school *j*;

β_{0j} was the average posttest score for students within school *j*;

β_{1j} and β_{2j} were level-1 coefficients that described the strength and direction of the associations between the prior mathematics and reading scores and the posttest scores. β_{3j} to β_{Qj} were additional level-1 coefficients that described the strength and direction of the associations between student demographic characteristics and the posttest scores. The student-level control variables included

students' economic status, special education status, English proficiency status, gender, and race/ethnicity. The student demographic characteristics were dummy coded. The analyses that included multiple grade levels also included dummy codes for the students' grade levels. r_{ij} was the residual (i.e., a level-1 random effect) associated with student *i*'s posttest score in school *j* using the level-1 model.

Level-2 model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Intervention Status})_j + u_{0j}$$

Where

γ_{00} was the average posttest score for the comparison group after accounting for the covariates;

Intervention status (SIG-I = 1; Comparison = 0) was a dummy coded variable that contrasted the schools that participated in the intervention with the comparison schools. γ_{01} was the level-2 coefficient that described the strength and direction of the association between the intervention status and the posttest scores. The analyses with the multiple grade levels also included a dummy coded variable contrasting elementary and middle schools. u_{0j} was the school random effect that corresponded to the deviation of school *j*'s level-1 intercept, β_{0j}, from its predicted value using the school-level model.

The HLM analyses with the MME data were similar to the grade level analyses with the MEAP data. The one exception was that student-level pretest achievement data was not available for the high school students. In place of the student-level pretest achievement data, we used school-level achievement data on the MME from the 2009-10 school year as covariates. Although school-level pretest achievement measures are not as effective at reducing bias as student-level measures, research has shown that school-level achievement measures substantially improve the precision of the impact estimates in cluster-randomized studies (Deke, Dragoset, & Moore, 2010). The HLM equations for the MME analyses were nearly the same as the HLM equations outlined above for the MEAP analyses. The prior mathematics and reading scores from the level-1 model were removed and prior mathematics, reading, and science scores were added to the level-2 model.

To assess whether the achievement of the SIG-I and comparison students in grades 3-8 differed significantly on the MEAP assessments at baseline, we used two-level HLM models. The HLM models included the students' prior MEAP scores as the dependent variables and a dummy coded variable contrasting the SIG-I and comparison schools as the level-2 predictor. In addition, we used *t* tests to compare the achievement of the SIG-I and comparison high schools at baseline. For all students, we used multi-level logistic regression models to determine whether the percentages of students in each demographic category differed significantly across the SIG-I and comparison groups.

WestEd examined multiple outcomes in each baseline and outcome analysis. As the number of outcome comparisons increases, the likelihood of committing a Type I error (i.e., identifying a

statistically significant difference when one does not actually exist) increases as well. To address this issue, WestEd used the Benjamini-Hochberg (B-H) correction for each group of baseline and outcome comparisons (Benjamini & Hochberg, 1995; Schochet, 2008; Thissen, Steinberg, & Kuang, 2002).[6]

SENSITIVITY ANALYSIS

WestEd conducted a sensitivity analysis using a regression discontinuity (RD) design to determine how sensitive the findings were to the analytical technique used. An RD design allows researchers to evaluate causal effects of an intervention when the assignment to the treatment condition is based a fixed cutoff on a continuous variable. An RD design permits one to draw stronger causal inferences than any other research design with the exception of a random assignment study. The notion behind the design is that the schools that are just above and below the cutoff are statistically indistinguishable and an unbiased estimate of the program effect results from difference in the scores at the cutoff (Shadish et al., 2002). The SIG-I application scores for the funded and non-funded schools were a viable assignment variable for the high schools. With the exception of three high schools, all of the applications that received funding had scores above the cutoff of 280. There were not sufficient numbers of elementary and middle schools with application scores to include in an RD design.

We conducted the analyses for the RD design using HLM with the students in the 31 high schools that received an application score. The model was consistent with the HLM model outlined in the data analysis section and included the assignment variable in the level-2 model. We conducted a number of different exploratory analyses that included and excluded the schools that received SIG-I funding even though they were below the cutoff of 280. We conducted other exploratory analyses that excluded the schools with application scores that were lower than the bulk of the application scores that were close to the cutoff. The conclusions drawn from the RD design were consistent with the conclusions drawn from the quasi-experimental design. Given the results of the analyses using the RD design were not substantively different than those using the quasi-experimental design, we opted to present only the results from the latter.

[6] The Benjamini-Hochberg correction was chosen because the power losses are smaller compared to other multiple comparison correction procedures (Schochet, 2008).

Findings

SIG-I AND COMPARISON GROUPS PRIOR TO SIG IMPLEMENTATION

STUDENT ACHIEVEMENT ON THE MEAP AND MME

The SIG-I and comparison students were generally well matched on their MEAP mathematics and reading scores prior to the first year of SIG implementation (see Exhibit 4). Generally, the differences on the MEAP mathematics assessment scale scores were not statistically significant. In grade 4, however, the students at SIG-I schools scored significantly higher on the MEAP mathematics assessment than the comparison students. None of the differences between the SIG-I and comparison students on the MEAP reading assessment were statistically significant. The effect sizes based on the differences on the MEAP reading assessment ranged from -0.14 to 0.19.

Exhibit 4. Scale Score Means on the MEAP Mathematics and Reading Assessments for SIG-I and Comparison Students Prior to SIG Implementation

	SIG-I Students		Comparison Students				
	Adjusted Mean	SD	Adjusted Mean	SD	Difference	p value	Effect Size
Mathematics							
Grade 3	316.76	21.43	313.12	14.53	3.64	.50	0.21
Grade 4	414.91	15.53	405.75	13.57	9.16*	< .001	0.65
Grade 5	500.77	19.72	498.71	18.77	2.07	.59	0.11
Grade 6	607.58	17.25	606.60	19.79	0.98	.89	0.05
Grade 7	704.90	18.58	706.49	20.47	-1.59	.56	-0.08
Reading							
Grade 3	317.31	26.32	312.98	21.94	4.34	.50	0.19
Grade 4	411.22	23.30	408.59	26.95	2.62	.67	0.10
Grade 5	505.34	27.73	509.10	27.02	-3.76	.46	-0.14
Grade 6	605.23	23.32	607.59	25.14	-2.36	.71	-0.10
Grade 7	703.19	25.58	702.85	27.52	0.34	.91	0.01

Note: The means are based on the fall 2010 administration of the MEAP and have been adjusted to account for the multi-level structure of the data. Positive differences indicate higher scores for SIG-I students. Each effect size was calculated by dividing the difference by the pooled standard deviation. Grade 3: SIG-I n = 212; Comparison n = 452. Grade 4: SIG-I n = 174; Comparison n = 451. Grade 5: SIG-I n = 213; Comparison n = 503. Grade 6: SIG-I n = 238; Comparison n = 520. Grade 7: SIG-I n = 1,026; Comparison n = 1,919.
*Statistically significant after using the Benjamini-Hochberg correction for multiple comparisons.

As shown Exhibit 5, the SIG-I high schools had higher scores on the MME than the comparison high schools prior to SIG implementation. The differences on the MME mathematics, reading, and

science were all statistically significant. The effect sizes ranged from 0.35 to 0.54 and would be considered moderate in size. These differences make the inclusion of the prior MME mathematics, reading, and science scores in the HLM models critical to accurately assess the impact of the intervention.

Exhibit 5. Scale Score Means on the MME Mathematics, Reading, and Science for SIG-I and Comparison Schools Prior to SIG Implementation

	SIG-I Schools		Comparison Schools				
	Mean	SD	Mean	SD	Difference	p value	Effect Size
Mathematics	1,073.6	15.9	1,053.9	15.5	19.7	< .001	0.54
Reading	1,093.6	10.0	1,082.2	10.1	11.4	< .001	0.35
Science	1,078.6	17.7	1,061.5	13.7	17.1	< .001	0.44

Note: The means are based on the spring 2010 administration of the MME. Positive differences indicate higher scores for SIG-I schools. Each effect size was calculated by dividing the difference by the state-level standard deviation based on student-level data. SIG-I n = 16; Comparison n = 36.

STUDENT DEMOGRAPHIC CHARACTERISTICS

As shown in Exhibit 6, the SIG-I and comparison students included in the MEAP analyses had similar demographic characteristics. The majority of the students were economically disadvantaged and African American/Black. Smaller percentages were special education and limited English proficient students. None of the differences between the two groups reached statistical significance and all of the differences were less than 10 percentage points.

Exhibit 6. Demographic Characteristics of the SIG-I and Comparison Students Included in the MEAP Analyses

	SIG-I Students	Comparison Students	Difference	p value
Economically Disadvantaged	87.8%	83.8%	3.9%	.23
Special Education	12.5%	12.6%	-0.1%	.38
Limited English Proficient	13.3%	3.7%	9.5%	.14
Female	49.7%	49.4%	0.3%	.84
Race/Ethnicity				
African American/Black	70.9%	79.4%	-8.5%	.09
Hispanic/Latino	13.4%	5.0%	8.4%	.12
White	10.8%	13.9%	-3.1%	.06
Other	4.9%	1.7%	3.2%	.50

Note: The Other category for Race/Ethnicity includes Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, Two or More Races, and Asian. The p values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data. SIG-I n = 1,863; Comparison n = 3,845.

The demographic characteristics of the SIG-I and comparison students included in the MME analyses were not as well matched as the two groups in the MEAP analyses (see Exhibit 7). The students in the SIG-I schools were less economically disadvantaged than the comparison students and the SIG-I schools had a slightly greater percentage of limited English proficient students. The SIG-I schools had a much lower percentage of African American/Black students and had higher proportions of White, Hispanic/Latino, and Other students. The differences for economic status, English proficiency status, and all of the racial/ethnic groups were statistically significant.

Exhibit 7. Demographic Characteristics of the SIG-I and Comparison Students Included in the MME Analyses

	SIG-I Students	Comparison Students	Difference	p value
Economically Disadvantaged	62.4%	74.6%	-12.1%*	.001
Special Education	10.4%	11.4%	-1.0%	.76
Limited English Proficient	7.6%	5.3%	2.3%*	.03
Female	53.1%	54.8%	-1.7%	.34
Race/Ethnicity				
African American/Black	44.6%	82.3%	-37.8%*	< .001
Hispanic/Latino	11.7%	6.5%	5.2%*	.01
White	36.2%	9.7%	26.5%*	< .001
Other	7.5%	1.5%	6.1%*	< .001

Note: The Other category for Race/Ethnicity includes Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, Two or More Races, and Asian. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data. SIG-I n = 1,937; Comparison n = 3,494.
*Statistically significant after using the Benjamini-Hochberg correction for multiple comparisons.

IMPACT OF SIG AT ELEMENTARY AND MIDDLE SCHOOLS

The results of the quasi-experimental design analyses for the MEAP mathematics, reading, and science assessments are shown in Exhibit 8. The exhibit contains the SIG-I and comparison students' mean scale scores (i.e., the unadjusted means) based on the fall 2012 administration of the MEAP, which measured the students' performance after the first year of SIG. The table also includes the adjusted means that account for the students' prior achievement, demographic characteristics, and multi-level structure of the data. The effect sizes included in Exhibit 9 are based on the adjusted means and allow the differences across subject areas and grades to be easily compared.

On the MEAP mathematics assessment, the adjusted means showed that the SIG-I students scored higher than the comparison students in grades 4-8. However, the differences across the individual grades and the difference pooled across all grades were not statistically significant. The largest difference occurred with the students in grade 7 and showed a 4.67 scale score point advantage for

the SIG-I students. The 95 percent confidence interval for this difference ranged from -3.02 to 12.35.

There was a mixed pattern of findings based on the adjusted means for the MEAP reading assessment. The SIG-I students had slightly higher scores than the comparison students in grades 6 and 7. In contrast, the comparison students held slight advantages over the SIG-I students in grades 4, 5, and 8. Consistent with the mathematics results, none of the differences on the MEAP reading assessment were statistically significant. In addition, the reading results pooled across grades 4-8 showed the SIG-I and comparison students had nearly identical performance on the MEAP reading assessment.

The comparison students had higher adjusted mean scores than the SIG-I students in grades 5 and 8 on the MEAP science assessment. However, the effect sizes showed that the differences in both grade levels were small and neither difference was statistically significant. The results pooled across grades 5 and 8 showed a small and non-significant difference favoring the comparison students.

Exhibit 8. Impact of SIG after One Year on Elementary and Middle School Students' MEAP Mathematics, Reading, and Science Scores

	SIG-I Students			Comparison Students			Adjusted Mean Difference	95% CI for the Adjusted Mean Difference	p value	Effect Size
	Unadjusted Mean	Adjusted Mean	SD	Unadjusted Mean	Adjusted Mean	SD				
Mathematics										
All Grades	-0.69	-0.71	0.67	-0.75	-0.73	0.71	0.02	[-0.05, 0.10]	.59	0.03
Grade 4	412.70	410.49	15.26	407.93	409.89	14.52	0.60	[-4.93, 6.13]	.83	0.04
Grade 5	505.97	499.11	22.92	495.76	498.37	16.06	0.74	[-3.66, 5.14]	.74	0.04
Grade 6	606.85	606.94	17.93	604.18	604.31	15.83	2.63	[-1.43, 6.69]	.20	0.16
Grade 7	708.42	707.72	17.14	702.45	703.05	14.97	4.67	[-3.02, 12.35]	.23	0.30
Grade 8	801.18	803.05	16.41	803.89	802.55	19.41	0.50	[-1.57, 2.56]	.64	0.03
Reading										
All Grades	-0.70	-0.71	0.89	-0.72	-0.71	0.92	0.00	[-0.08, 0.08]	.98	0.00
Grade 4	412.92	409.18	27.55	408.66	411.47	23.92	-2.29	[-9.59, 5.01]	.54	-0.09
Grade 5	510.30	506.46	26.14	508.20	509.71	25.75	-3.25	[-11.03, 4.53]	.41	-0.13
Grade 6	606.27	607.38	24.33	606.96	606.35	23.76	1.03	[-4.21, 6.28]	.70	0.04
Grade 7	705.31	704.81	28.90	699.92	699.37	24.71	5.44	[-4.62, 15.49]	.29	0.21
Grade 8	808.69	809.87	20.76	810.60	810.37	23.54	-0.51	[-2.32, 1.31]	.59	-0.02
Science										
All Grades	-0.82	-0.79	0.75	-0.72	-0.72	0.84	-0.07	[-0.16, 0.01]	.08	-0.09
Grade 5	496.37	491.15	23.43	492.72	494.40	19.12	-3.26	[-9.07, 2.56]	.27	-0.16
Grade 8	801.17	802.89	17.57	804.82	804.07	20.51	-1.17	[-3.40, 1.06]	.30	-0.06

Note: Adjusted means are based on MEAP scale scores from the fall 2011 administration. The "All Grades" MEAP scores were normalized within grade level by subtracting the state mean from each student's scale score and dividing by the state standard deviation. A score of -0.75 is equivalent to a scale score that is three-fourths of a standard deviation below the state mean. Positive differences indicate higher scores for SIG-I students. The adjusted means have been adjusted to account for the multi-level structure of the data and the student-level covariates. Each effect size was calculated by dividing the adjusted mean difference by the pooled standard deviation. Grade 4: SIG-I n = 212; Comparison n = 452. Grade 5: SIG-I n = 174; Comparison n = 451. Grade 6: SIG-I n = 213; Comparison n = 503. Grade 7: SIG-I n = 238; Comparison n = 520. Grade 8: SIG-I n = 1,026; Comparison n = 1,919.

IMPACT OF SIG AT HIGH SCHOOLS

The quasi-experimental design results for the MME mathematics, reading, and science assessments are shown in Exhibit 9. The exhibit includes the SIG-I and comparison students' mean scale scores (i.e., the unadjusted means) on the spring 2011 MME administration, which assessed the performance of students in grade 11 after the first year of SIG implementation. The exhibit also contains the adjusted means that account for the students' demographic characteristics, prior school-level achievement, and the multi-level structure of the data. The effect sizes included in Exhibit 9 were calculated using the adjusted means and allow the differences to be compared across subject areas.

The unadjusted scale score means were higher for the SIG-I students than for the comparison students on the MME mathematics, reading, and science assessments. However, the differences between the two groups on the adjusted means were less than two scale score points for each subject. None of the differences on the mathematics, reading, and science assessments were statistically significant.

Exhibit 9. Impact of SIG after One Year on High School Students' MME Mathematics, Reading, and Science Scale Scores

	SIG-I Students			Comparison Students			Adjusted Mean Difference	95% CI for the Adjusted Mean Difference	p value	Effect Size
	Unadjusted Mean	Adjusted Mean	SD	Unadjusted Mean	Adjusted Mean	SD				
Mathematics	1,074.52	1,062.22	41.66	1,055.98	1,062.72	48.25	-0.50	[-5.89, 4.89]	.86	-0.01
Reading	1,092.91	1,085.39	31.54	1,081.18	1,085.69	31.02	-0.30	[-3.53, 2.94]	.86	-0.01
Science	1,085.52	1,076.05	38.01	1,069.57	1,074.36	41.50	1.69	[-2.70, 6.08]	.45	0.04

Note: The means are based on MME scale scores from the spring 2011 administration. The adjusted means have been adjusted to account for the multi-level structure of the data and the student- and school-level covariates. Each effect size was calculated by dividing the adjusted mean difference by the pooled standard deviation. SIG-I n = 1,937; Comparison n = 3,494.

CONCLUSION AND NEXT STEPS

The impact findings in the current report are based on state assessments of student performance in reading, math, and science administered after one year of SIG or, in the case of high schools, in the spring of the first year of SIG. Progress with regards to SIG in these schools may inform the impact findings. Using qualitative data collected during the first year of SIG, WestEd found that after one year of SIG, many schools had made substantial progress in terms of replacing the principals, providing professional development and coaching, placing teachers in leadership roles, and empowering school staff (Bojorquez, Rice, Hipps, & Li, 2012). However, several aspects of SIG implementation were lagging such as the development of the critical skills and competences used to hire and place school staff and data-driven processes for monitoring progress of SIG implementation and student learning. Also, the extent to which districts supported and facilitated the implementation of SIG was often limited in the first year. For example, many principals felt constrained by districts with regards to budgeting and curricula selection, and bureaucratic hurdles at the district level reportedly hindered full implementation of year-one SIG plans in a few districts. Finally, there appeared to be less progress on SIG implementation in schools that adopted the turnaround model versus those that adopted the transformation model.

Subsequent reports will examine SIG impact after year two and correlate those with quantitative data collected about implementation during the same year. In addition, a subsequent report will examine SIG impact in year three and use case studies at several SIG-I schools to provide a more in-depth examination of SIG implementation over the three years.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, 57,* 289-300.

Bojorquez, J., Rice, J., Hipps, J., & Li, J. (2012). *Evaluation of Michigan's 1003(g) School Improvement Grants: Implementation Trends in the First Year*. Los Alamitos, CA: WestEd.

Cook, D. T., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*, 724-750. doi: 10.1002/pam.20375

Deke, J., Dragoset, L., & Moore, R. (2010). *Precision gains from publically available school proficiency measures compared to study-collected test scores in education cluster-randomized trials* (NCEE 2010-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hurlburt, S., Le Floch, K. C., Therriault, S. B., & Cole, S. (2011). *Baseline Analyses of SIG Applications and SIG-eligible and SIG-awarded Schools* NCEE 2011-4019). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86-92. doi: 10.1027/1614-2241.1.3.86

Michigan Department of Education (n.d.) *Michigan Educational Assessment Program Technical Report 2010-2011*. Retrieved from http://www.michigan.gov/documents/mde/MEAP_2010-2011_Technical_Report_394693_7.pdf

Michigan Department of Education. (2011a). *Michigan Merit Examination technical manual: 2010 testing cycle*. Retrieved from: http://www.michigan.gov/documents/mde/MME_2010_Technical_Manual_final_359989_7.pdf

Michigan Department of Education. (2011b). *Michigan Merit Examination technical manual: 2010 testing cycle*. Retrieved from: http://www.michigan.gov/documents/mde/MME_2011_Tech_Manual_final_122211_372971_7.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Schochet, P. Z. (2008). *Technical methods report: Guidelines for multiple testing in impact evaluations.* Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Scott, C. (2011, February). *Changing tires en route: Michigan rolls out millions in School Improvement Grants.* Washington, DC: Center on Education Policy.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*, 323-355. doi: 10.3102/10769986023004323

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27,* 77-83.