Data Analysis and Evaluation of McGraw-Hill's

Everyday Mathematics Program and its Impact on M-Step Scores in

Michigan Schools

David Marshall, Ph.D., Auburn University

Daniel Henry, Ph.D., Director, Auburn Center for Evaluation

A UBURN
C ENTER FOR
E ✓ALUATION

AUBURN
UNIVERSITY ™

Executive Summary

Twelve school districts in Michigan have adopted McGraw-Hill's *Everyday Mathematics*, a core curriculum-based approach to teaching mathematics in elementary school classrooms developed at the University of Chicago. In 2019, the Auburn Center for Evaluation (ACE) was asked to analyze mathematics test data from the state of Michigan's Michigan Student Test of Educational Proficiency for students served by the program to answer this main research question: "*To what extent is the utilization of McGraw-Hill's elementary mathematics curriculum associated with differences in year-end standardized test scores?*"

To answer that question, a number of statistical analyses were employed matching school districts who did not use the *Everyday Mathematics* and those who did. The results suggest that students who participated *Everyday Mathematics* produced better results on state tests than those who did not.

Key Findings

1. Overall, students enrolled in school districts using the McGraw-Hill curriculum had proficiency rates that were greater than demographically matched peers. Fourth-grade McGraw-Hill students had proficiency rates that were 10.9% greater than matched peers; fifth-grade students had rates that were 13.9% greater than matched peers.

2. Non-White McGraw-Hill students performed substantially outperformed non-White matched peers on the math M-STEP, scoring proficient between 17 and 19 percentage points higher than matched peers on the third, fourth, and fifth-grade assessments.

3. Fourth and fifth grade McGraw-Hill students significantly outperformed their matched peers by gender. McGraw-Hill male students earned proficient scores by between 6 and 10 points higher than their matched counterparts. McGraw-Hill female students earned

proficient scores that were approximately 5 points higher than their matched peers in both grades.

4. Fourth and fifth grade economically disadvantaged McGraw-Hill students had proficiency rates that were 20% greater than their matched peers.

Limitations of the Evaluation

1. Data for the evaluation were completely dependent on the quality and quantity of information collected and reported by the Michigan Department of Education and McGraw-Hill.

2. Transience or mobility of the student population in participating schools is a concern. Comparisons made in the evaluation are based on the assumption that children in participating schools received the "treatment" of the *Everyday Mathematics* program, but a more extensive per-pupil analysis of student exposure to *Everyday Mathematics* and student mobility is necessary to fully understand this factor.

3. Intermediate and long-term shifts in knowledge, attitudes, perceptions, and achievements in mathematics may not have resulted from the *Everyday Mathematics* alone. Many schools in Michigan have multiple federal and state initiatives in effect at the same time. Since many of these initiatives are aimed at providing increased academic achievement, this may or may not have had an impact on student test scores.

4. Interpretations based on statistical significance alone should be made with caution.

Data and Method

This evaluation sought to answer the single evaluation question: *To what extent is the utilization of McGraw-Hill's elementary mathematics curriculum associated with differences in year-end standardized test scores?* The state of Michigan had 540 operational local education agencies during the 2017-18 school years (hereafter referred to as school districts), and these districts diverge widely along a number of variables. In light of these differences between districts, it was deemed impractical to compare the 12 school districts which employ McGraw-Hill's curriculum to 528 which did not. Thus, a matching approach was employed to create a comparison group of districts similar to those with which McGraw-Hill partners.

Data Sources and Matching Process

To address the evaluation question, two sets of publicly available data were obtained from the Michigan Department of Education's (MDE) website – one that included enrollment data for each school district, and one that included Michigan Student Test of Educational Proficiency (M-STEP) data. The M-STEP has been administered annually since Spring 2015 as a summative statewide assessment for English Language Arts, Math, Science, and Social Studies for grades three through five. According to the Michigan Department of Education (2019) website:

> The M-STEP is a 21st Century online test given for the first time in the Spring of 2015. It
> is designed to gauge how well students are mastering state standards. These standards,
> developed for educators by educators, broadly outline what students should know and be
> able to do in order to be prepared to enter the workplace, career education training, and
> college. M-STEP results, when combined with classroom work, report cards, local

district assessments and other tools, offer a comprehensive view of student progress and achievement. (n.p.)

All data analyses were conducted using Stata version 15. District-level demographic data were obtained, and districts were matched using two approaches. Matching approaches are used as part of an effort to control for the influence of other variables on the variables of interest. In the case of this evaluation, it was of interest to learn if standardized test scores were different in districts that used McGraw-Hill's curriculum and those that did not. Districts were matched on two such variables – percentage of students receiving special education services and percentage of economically disadvantaged students, both of which have been demonstrated to adversely impact test scores (Koretz, 2008). Two coarsened exact matching approaches were employed. Coarsened exact matching was appropriate for this task as it created a balance between the treatment and matched groups along relevant variables (Blackwell, Iacus, King, & Porro, 2009). The first approach involved creating a pool of matched districts. In all, 67 districts were matched with 11 of the 12 treatment districts, or districts using the McGraw-Hill curriculum. This approach uses weighting to account for variations between the districts. After the matching process was complete, two one-way analysis of variance (ANOVA) tests were conducted to compare the treatment districts with the pool of matched districts on the two variables of interest – percent of student population that were economically disadvantaged and percent of student population receiving special education services. The assumption of homogeneity of variance was violated for both tests. In other words, the two data distributions were statistically significantly different from one another.

To address these issues, a second coarsened exact matching approach was used which creates a one-to-one match for each treatment district instead of creating a pool of matched

districts. This approach also successfully matched 11 of the 12 districts. Two one-way ANOVAs were conducted to compare the two variables of interest. The tests found the groups to be statistically similar. This time when the same assumptions were checked, no violations were found. District enrollment was a third variable that was initially included in both matching approaches; however, fewer than half of the treatment districts were matched when it was included. Enrollment was accounted for later in the analysis by simulating individual-level data. Each of the treatment districts has employed the McGraw-Hill curriculum for a minimum of four years, and most for greater than five years. As such, length of utilization was not considered in these analyses. Taylor School District was the lone district that was not matched in the process. See Table 1 for a description of this district. Table 2 provides a list of districts in each group and Table 3 compares the treatment and matched districts along the two matched variables.

Table 1
Demographics for Taylor School District

Total Enrollment	6320
Non-White	46.6%
Special Education	20.6%
Economic Disadvantage	78.8%

Note: The Non-White, Special Education, and Economic Disadvantage variables reflect the proportion of the student population for which these labels apply.

Table 2
List of Treatment and Matched Districts

Treatment	Berkley School District, Davison Community Schools, Flat Rock Community Schools, Livonia Public Schools School District, Madison District Public Schools, Novi Community School District, Plymouth-Canton Community Schools, Pontiac City School District, Rochester Community School District, Southfield Public School District, Walled Lake Consolidated Schools

| Matched | Britton Deerfield Schools, Brown City Community Schools, Centreville Public Schools, Chelsea School District, East Grand Rapids Public Schools, Grosse Ile Township Schools, Hamilton Community Schools, Hart Public School District, Manchester Community Schools, Marquette Area Public Schools, Van Dyke Public Schools |

Table 3
Comparison of Treatment and Matched Districts

	Treatment	Matched
Economic Disadvantage	41.87%	38.71%
Special Education	12.8%	12.1%

For both variables of interest, the matching process created a comparison group that had fewer economically disadvantaged students and students receiving special education services than the treatment group. Though these differences are not statistically significant, they possibly minimally underestimate the impact of the McGraw-Hill curriculum in the analyses presented on the following pages.

Third-grade curriculum

According to McGraw-Hill's *Everyday Mathematics* website (2019), the third-grade curriculum focuses on procedures, concepts, and applications in four critical areas:

- Understanding of multiplication and division and strategies within 100.

- Understanding of fractions, especially unit fractions.

- Understanding of the structure of rectangular arrays and of area.

- Describing and analyzing two-dimensional shapes.

Third-grade student performance on the M-STEP was analyzed two different ways. First, the number of students earning proficient scores on the M-STEP in treatment districts was compared to the number of students earning proficient scores in matched districts. All third-grade students

in the treatment and matched districts were compared by conducting a one-way ANOVA. One-way ANOVAs were also conducted for each of five sub-groups of third-grade students: (1) students receiving special education services; (2) economically disadvantaged; (3) non-White students; (4) male students; and (5) female students. See Figure 1 for a graph comparing each of the six total groups that were analyzed for third-grade students.

First, all third-grade students across the two groups were compared in terms of whether they earned proficient scores on the M-STEP. No statistically significant differences were found between the scores of students from the treatment districts (M = 58.2%, SD = .49) and the matched districts (M = 57.2%, SD = .50). Two of the subgroup analyses yielded significant findings. Because districts in the treatment and matched groups were already matched on the proportion of economically disadvantaged students and students receiving special education services, two of the subgroup analyses sought to learn if economically disadvantaged and special education students perform differently on the M-STEP in districts that are similar as a whole in terms of these two subgroup populations. However, no statistically significant differences existed for students receiving special education services. It should be noted that students in the treatment group (M = 33.7%, SD = .47) performed statistically worse on the M-STEP than did students in the matched group (M = 38.9%, SD = .49). In contrast, non-White students utilizing the McGraw-Hill curriculum (M = 50.8%, SD = .50) performed substantially better on the M-STEP than students in the matched group (M = 33.5%, SD = .47). No statistically significant gender differences existed between groups. See Table 4 for means, standard deviations, *p*-values, and effect sizes where applicable for each of the analyses.

3rd Grade M-STEP Performance
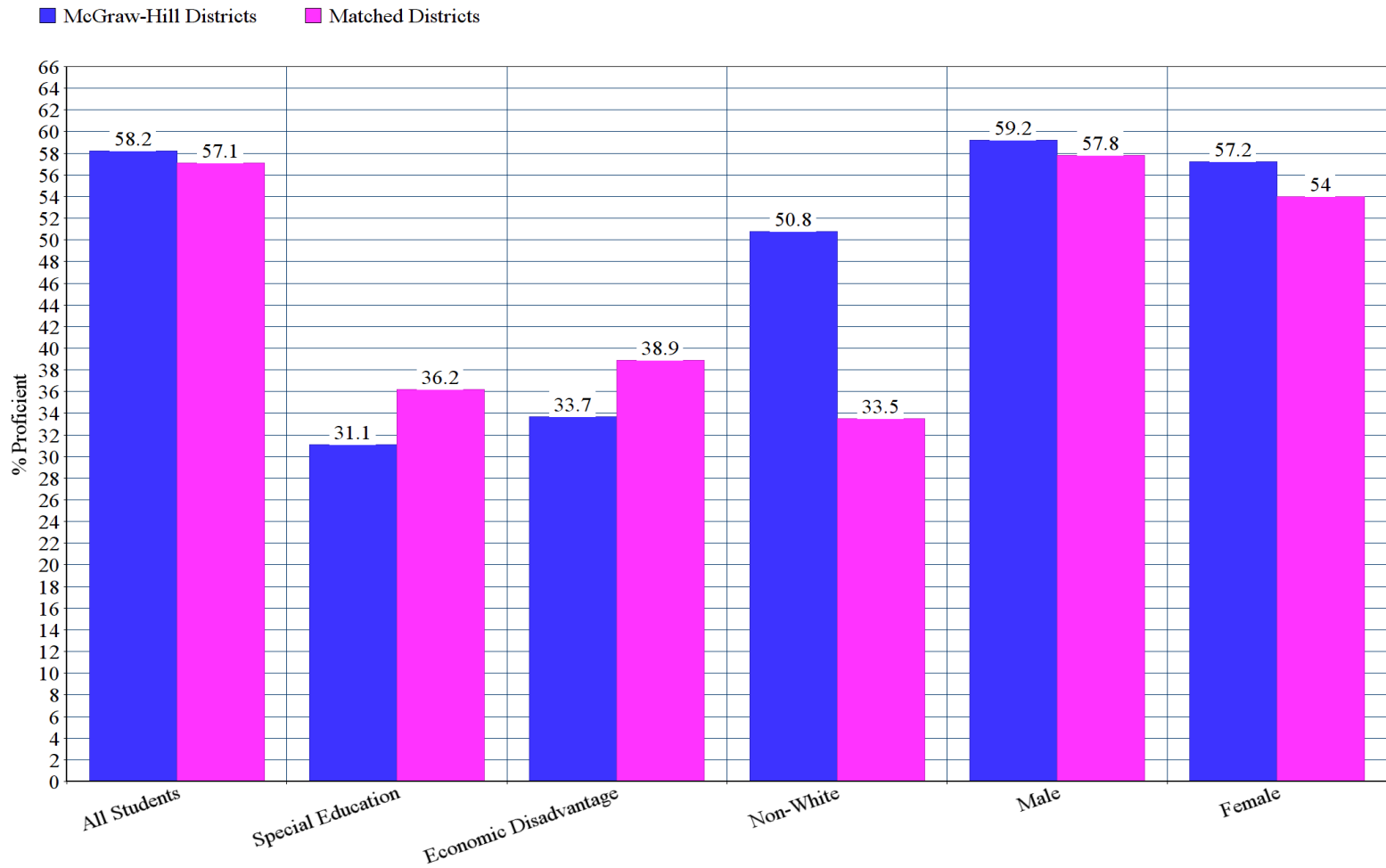
Figure 1 – Third-Grade M-STEP Proficiency

Table 4.
Means, Standard Deviations, and p-values for 3rd Grade M-STEP Proficiency

	M	SD	*p*-value	*d*
All Students				
Treatment	58.20%	.493	0.47	n/a
Matched	57.15%	.495		
Special Education				
Treatment	31.13%	.463	0.25	n/a
Matched	36.17%	.483		
Economic Disadvantage				
Treatment	33.72%	.473	0.03*	0.11
Matched	38.92%	.488		
Non-White				
Treatment	50.77%	.500	0.00**	0.36
Matched	33.54%	.473		
Male				
Treatment	59.22%	.491	0.43	n/a
Matched	57.66%	.494		
Female				
Treatment	57.21%	.494	0.13	n/a
Matched	54.02%	.499		

Note: $** p < .01;$ $* p < .05;$ *Alpha significance level set a priori at .05.*

All third-grade students were also compared in terms of their mean M-STEP scale scores, as were the same five subgroups. Five of the six analyses yielded significant findings; however, only two of the analyses produced findings that were practically significant. Students who received special education services in matched districts (M = 1286.92, SD = 17.77) scored significantly higher than students in treatment districts (M = 1283.23, SD = 7.24). A Cohen's *d* of 0.22 was calculated indicating a medium effect size. Non-White students also in treatment districts (M = 1300.86, SD = 14.72) outperformed peers in matched districts (M = 1292.79, SD = 12.41) in terms of their scaled scores on the M-STEP. A Cohen's d of 0.59 represents a large treatment effect for the McGraw-Hill curriculum with third-grade non-White students. See Table 5 for means, standard deviations, *p*-values, and effect sizes where applicable.

Table 5.

Means, Standard Deviations, and p-values for 3rd Grade M-STEP Scaled Scores

	M	SD	*p*-value	*d*
All Students				
Treatment	1304.22	10.72	0.03*	0.06
Matched	1303.51	12.64		
Special Education				
Treatment	1283.92	7.24	0.00**	0.22
Matched	1286.23	17.77		
Economic Disadvantage				
Treatment	1288.20	7.87	0.00**	0.15
Matched	1289.46	7.97		
Non-White				
Treatment	1300.86	14.72	0.00**	0.59
Matched	1292.79	12.41		
Male				
Treatment	1305.12	11.15	0.75	n/a
Matched	1304.98	12.50		
Female				
Treatment	1303.27	10.38	0.00**	0.14
Matched	1301.59	13.17		

*Note: ** p < .01; * p < .05; Alpha significance level set a priori at .05.*

Fourth-grade curriculum

According to McGraw-Hill's *Everyday Mathematics* website (2019), the fourth-grade

curriculum focuses on procedures, concepts, and applications in three critical areas:

- Understanding and fluency with multi-digit multiplication and understanding of dividing

  to find quotients with multi-digit dividends.

- Understanding of fraction equivalence, addition and subtraction of fractions with like

  denominators, and multiplication of fractions by whole numbers.

- Understanding that geometric figures can be analyzed and classified based on their

  properties.

The same analyses were conducted for fourth-grade students comparing students in the

treatment district group and matched district group. One-way ANOVAs were conducted for each

of the analyses to explore whether differences existed between the two groups. Students who

used the McGraw-Hill curriculum (M = 60.85%, SD = .488) outperformed their peers in the

matched group (M = 54.89, SD = .498) in terms of scoring proficient on the M-STEP. All but

one of the five subgroup analyses yielded significant findings. See Figure 2 for a graph

comparing each of the six groups that were analyzed for fourth-grade students. No differences

existed between treatment and matched groups for students who received special education

services. For each of the other four subgroups, students utilizing the McGraw-Hill curriculum

outperformed their peers in the matched group. Both male and female students separately had

higher proficiency rates on the M-STEP than students in the matched districts, as did

economically disadvantaged students. The difference was most substantial for non-White

students; those in the treatment group (M = 49.5%, SD = .500) scored proficient on the test more

often than their matched group peers (M = 31.0%, SD = .463), and this yielded a medium effect

size (d = 0.38). Effect sizes are a standardized measure of practical significance, expressed in

terms of standard deviations, or average distance from the mean. Whereas fewer than one-third

of fourth-grade matched district non-White students earned a proficient score on the M-STEP,

almost half of similar students did in the treatment group. See Table 6 for means, standard

deviations, p-values, and effect sizes where applicable.
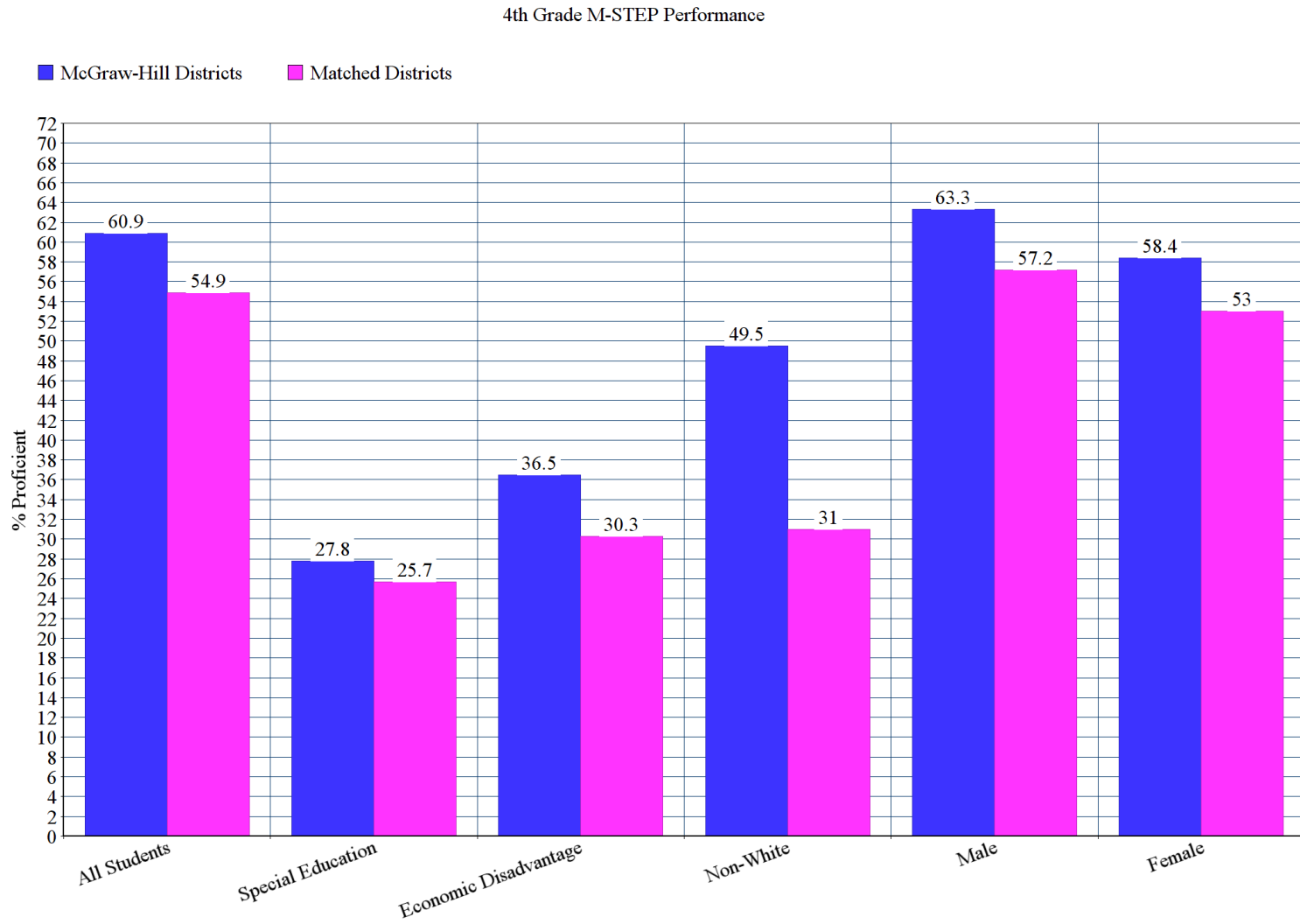
Figure 2 – Fourth-Grade M-STEP Proficiency

Table 6.
Means, Standard Deviations, and p-values for 4th Grade M-STEP Proficiency

	M	SD	*p*-value	*d*
All Students				
Treatment	60.85%	.488	0.00**	0.12
Matched	54.89%	.498		
Special Education				
Treatment	27.84%	.449	0.61	n/a
Matched	25.71%	.439		
Economic Disadvantage				
Treatment	36.47%	.481	0.01**	0.13
Matched	30.34%	.460		
Non-White				
Treatment	49.50%	.500	0.00**	0.38
Matched	31.01%	.463		
Male				
Treatment	63.26%	.482	0.00**	0.12
Matched	57.18%	.495		
Female				
Treatment	58.44%	.492	0.01*	0.11
Matched	53.05%	.499		

*Note: ** p < .01; * p < .05; Alpha significance level set a priori at .05.*

Fourth-grade student data were also analyzed in terms of mean scale scores on the M-STEP. Data were analyzed for all fourth-grade students in the treatment and matched districts, as well as for each of the five sub-groups of interest. Among all fourth-grade students, those enrolled in districts that used the McGraw-Hill curriculum (M = 1404.13, SD = 10.59) scored higher on average on the M-STEP than did their peers in matched districts (M = 1401.44, SD = 10.62). Four of the five subgroup analyses also yielded positive findings for treatment districts. Although no statistically significant differences were found for students receiving special education services, medium to large effect sizes were found for the McGraw-Hill curriculum groups among fourth-grade students who were economically disadvantaged (*d* = 0.21), males (*d* = 0.23), and females (d = 0.32). The largest impact was found for students who were not White; they outscored their matched district peers by nine points on the M-STEP, yielding a large effect

size (d = 0.52). See Table 7 for means, standard deviations, p-values, and effect sizes where applicable for scaled scores.

Table 7.
Means, Standard Deviations, and p-values for 4ᵗʰ Grade M-STEP Scaled Scores

	M	SD	p-value	d
All Students				
Treatment	1404.13	10.59	0.00**	0.25
Matched	1401.44	10.62		
Special Education				
Treatment	1381.84	8.29	0.23	n/a
Matched	1380.82	10.48		
Economic Disadvantage				
Treatment	1389.32	8.93	0.00**	0.21
Matched	1387.73	6.04		
Non-White				
Treatment	1399.33	15.10	0.00**	0.52
Matched	1392.30	11.59		
Male				
Treatment	1405.35	11.37	0.00**	0.23
Matched	1402.70	11.85		
Female				
Treatment	1402.94	9.93	0.00**	0.31
Matched	1399.83	10.05		

*Note: ** $p < .01$; * $p < .05$; Alpha significance level set a priori at .05.*

Fifth-grade curriculum

According to McGraw-Hill's *Everyday Mathematics* website (2019), the fifth-grade

curriculum focuses on procedures, concepts, and applications in three critical areas:

- Developing addition/subtraction fluency with fractions and understanding of

  multiplication/division of fractions in limited cases.

- Developing fluency with decimal operations, extending division to 2-digit divisors,

  integrating decimals into the place-value system, and understanding operations with

  decimals to hundredths.

- Developing an understanding of volume.

Data were analyzed for fifth-grade students in the same manner as third- and fourth-grade

students. One-way ANOVA tests were conducted to explore differences between M-STEP

results for students in treatment districts that use the McGraw-Hill curriculum and matched

districts that do not. Data were first analyzed in terms of whether students earned a proficient

score on the standardized math test. Fifth-grade students in the treatment group (M = 55.14%,

SD = .497) earned proficient scores on the M-STEP more often than did students in the matched

comparison group (M = 48.76%, SD = .500). Four of the five sub-groups of interest also yielded

statistically significant findings for the treatment districts. See Figure 3 for a graph comparing

each of the six groups that were analyzed for fifth-grade students. For students receiving special

education services, no statistically significant differences were found between the treatment and

matched districts. However, significant differences were found for economically disadvantaged

students, non-White students, males, and females. In each of these cases, students in the

treatment districts outperformed their counterparts in the matched districts. Two of the analyses

yielded medium effect sizes for the treatment groups. The largest effect size (d = .41) was for

non-White students in treatment districts (M = 43.10%, SD = .495), who earned proficient M-

STEP scores at almost twice the rate of non-White students in matched comparison districts (M

= 24.22%, SD = .429). See Table 8 for means, standard deviations, *p*-values, and effect sizes

where applicable.
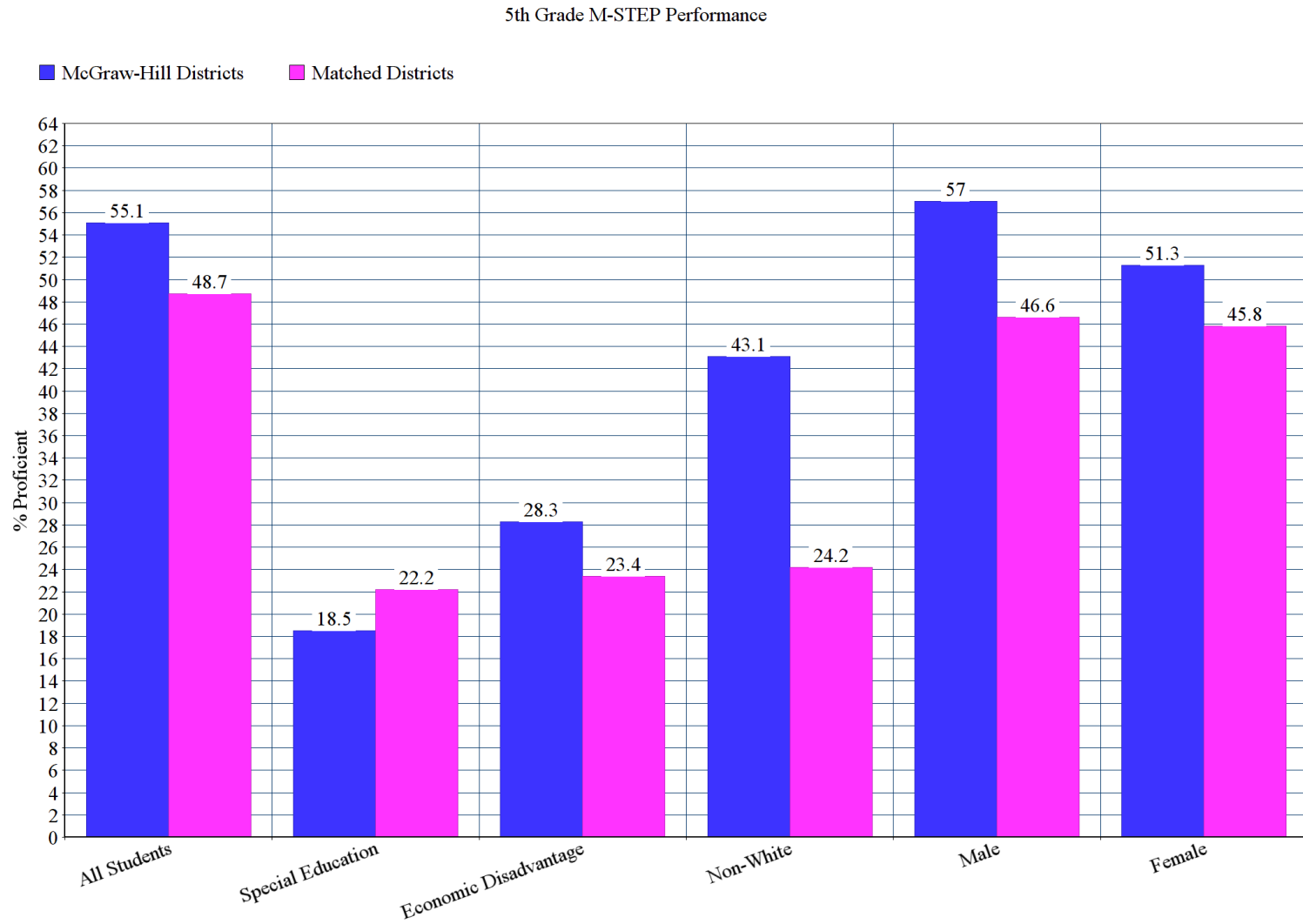
Figure 3 – Fifth-Grade M-STEP Proficiency

Table 8.
Means, Standard Deviations, and p-values for 5th Grade M-STEP Proficiency

	M	SD	*p*-value	*d*
All Students				
Treatment	55.14%	.497	0.00**	0.13
Matched	48.67%	.500		
Special Education				
Treatment	18.53%	.389	0.30	n/a
Matched	22.30%	.418		
Economic Disadvantage				
Treatment	28.27%	.450	0.02*	0.12
Matched	23.38%	.424		
Non-White				
Treatment	43.11%	.495	0.00**	0.41
Matched	24.23%	.429		
Male				
Treatment	56.97%	.495	0.00**	0.21
Matched	46.61%	.499		
Female				
Treatment	51.29%	.499	0.01**	0.11
Matched	45.77%	.499		

*Note: ** p < .01; * p < .05; Alpha significance level set a priori at .05.*

Fifth-grade data were also analyzed in terms of whether differences existed between treatment and matched districts on mean M-STEP scaled scores. Overall, students enrolled in treatment districts (M = 1498.52, SD = 11.36) had significantly higher mean scaled scores than their peers in matched comparison districts (M = 1495.66, SD = 10.92), representing a medium effect size (*d* = 0.26). When sub-groups were analyzed, no significant differences were found between treatment and matched districts for students receiving special education services or economically disadvantaged students. However, gender differences were found between the two groups of school districts. Both male and female students in treatment districts outscored their counterparts on the M-STEP, and both analyses yielded medium effect sizes. Similar to the findings from analyses conducted for third-grade and fourth-grade students, the largest difference between treatment and matched districts was found for non-White treatment district students, for

which a large effect size was found ($d = 0.64$). On average, non-White students enrolled in districts using the McGraw Hill curriculum scored nine points higher than their peers in matched comparison districts. See Table 9 for means, standard deviations, p-values, and effect sizes where applicable.

Table 9.
Means, Standard Deviations, and p-values for 5th Grade M-STEP Scaled Scores

	M	SD	p-value	d
All Students				
Treatment	1498.52	11.36	0.00**	0.26
Matched	1495.66	10.92		
Special Education				
Treatment	1472.41	7.58	0.62	n/a
Matched	1472.83	14.33		
Economic Disadvantage				
Treatment	1482.34	8.05	0.33	n/a
Matched	1481.96	8.30		
Non-White				
Treatment	1494.25	15.62	0.00**	0.64
Matched	1485.32	12.29		
Male				
Treatment	1499.55	11.72	0.00**	0.37
Matched	1495.02	12.55		
Female				
Treatment	1497.60	10.89	0.00**	0.26
Matched	1494.87	9.92		

*Note: ** p < .01; * p < .05; Alpha significance level set a priori at .05.*

Data Limitations

Some limitations should be noted for this evaluation. First, while 11 of the 12 treatment districts were matched using the coarsened exact matching approach, there was no match for Taylor School District. This is likely due to the high proportion of students that receive special education services in the district (20.6%). As such, the findings articulated in this report reflect M-STEP performance in the other 11 treatment districts, and do not necessarily hold true for Taylor School District. Also, the statistical analyses conducted for this evaluation can detect whether differences exist between sample means. Two key variables were controlled for in the matching (special education and economic disadvantage); however, schools are complex organizations and it is always possible that other variables or more complex interactions between variables influenced the results of M-STEP tests. These analyses simply detected whether differences existed between the treatment and matched groups. Finally, the sub-group analyses for students receiving special education services excluded, by necessity, a small number of students in a few districts for each analysis. For example, students from five third-grade districts, six fourth-grade districts, and four fifth-grade districts were omitted from these analyses due to the small sample sizes that existed. Finally, because the Michigan Department of Education reports its findings in aggregate, not at the individual-level, some analyses could not be performed. While individual-level data could be simulated given complete data for other sub-groups, districts with fewer than ten students receiving special education services in a given grade-level were omitted here.

Summary

This evaluation sought to understand if students who were enrolled in districts using the McGraw-Hill curriculum performed better on the M-STEP assessment than students in similar

districts that employed a different math curriculum. Overall, students enrolled in treatment districts performed better than those who were not. Although no significant differences were found for third-grade students in terms of proficient scores, fourth- and fifth-grade students in treatment districts did outperform their peers in matched districts. When scaled scores were compared, treatment district students outperformed their peers in each of the three grades that were examined. For students receiving special education services, significant differences were only found among third grade students for mean scale scores. This was one of only three analyses conducted for this evaluation that found matched district students to outperform treatment district students (out of 32 total analyses). Fourth- and fifth-grade students receiving special education services performed similarly across both groups. This finding was not surprising given that proportion of students receiving special education services was one of the variables controlled for with the matching approach. The other variable factored into the matching process, economic disadvantage, yielded mixed findings. Third-grade students in districts that were similar in terms of their proportion of economically disadvantaged students performed better in matched districts; however, similar fourth-grade students performed significantly better in treatment districts. Fifth-grade findings were mixed and not practically significant.

Aside from the overall findings, the most important findings were found among the race and gender sub-groups. Although no gender differences were found among third-grade students included in these analyses, significant differences were found for fourth- and fifth-grade students. Treatment district male students outperformed matched district male students in terms of both proficient scores and mean scaled scores in both grades; similar findings were found for fourth- and fifth-grade female students. This is particularly encouraging given the deserved

attention that has been given to gender differences in STEM. A large body of literature (e.g.

Wong & Degol, 2017) has found females less likely to pursue careers in STEM; as such,

curricula that support female learning in STEM subject areas, including math, are worth

pursuing. The most substantial findings, however, were among non-White students. Although

White students still outperform non-White students in Michigan on the M-STEP, these analyses

found that non-White students enrolled in districts using the McGraw-Hill curriculum to

substantially outperform their non-White peers in other districts. See Table 10 for a summary of

non-White student M-STEP performance.

Table 10.
Non-White Student M-STEP Performance and Effect Sizes

	Proficient Score Effect Size	Scaled Score Effect Size
3rd Grade	0.36**	0.59**
4th Grade	0.38**	0.52**
5th Grade	0.41**	0.64**

*Note: ** p < .01; * p < .05; Alpha significance level set a priori at .05.*

Non-White effect sizes for treatment districts were found to increase for each subsequent

grade during the elementary years, and scaled score effect sizes ranged from 0.52 to 0.64,

representing scores that are more than half a standard deviation higher than non-White students

enrolled in matched comparison districts that did not use the McGraw-Hill curriculum. The

achievement gap between White and non-White students, and its persistence, has been well-

explored in the literature (e.g. Ladson-Billings, 2006). These findings should be taken seriously

in light of efforts to close the achievement gap between White students and students of color.

Key Findings:

1. Overall, students enrolled in school districts using the McGraw-Hill curriculum had

   proficiency rates that were greater than demographically matched peers. Fourth-grade

McGraw-Hill students had proficiency rates that were 10.9% greater than matched peers; fifth-grade students had rates that were 13.9% greater than matched peers.

2. Non-White McGraw-Hill students performed substantially outperformed non-White matched peers on the math M-STEP, scoring proficient between 17 and 19 percentage points higher than matched peers on the third, fourth, and fifth-grade assessments.

3. Fourth and fifth grade McGraw-Hill students significantly outperformed their matched peers by gender. McGraw-Hill male students earned proficient scores by between 6 and 10 points higher than their matched counterparts. McGraw-Hill female students earned proficient scores that were approximately 5 points higher than their matched peers in both grades.

4. Fourth and fifth grade economically disadvantaged McGraw-Hill students had proficiency rates that were 20% greater than their matched peers.

References

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). CEM: Coarsened exact matching in Stata. *The Stata Journal, 9*(4), 524-546.

Koretz, D. *Measuring up: What educational testing really tells us.* Cambridge, Massachusetts: Harvard University Press.

Ladson-Billings, G. (2006). From the achievement gap to the educational debt: Understanding achievement in U.S. schools. *Educational Researcher, 35*(7), 3-12.

McGraw-Hill (2019). *Exploring everyday mathematics*. Retrieved 2/28/19 from https://www.mheducation.com/prek-12/program/microsites/MKTSP-TRA15M0/program-materials.html.

State of Michigan, (2019). M-Step, *Michigan student test of educational proficiency*. Retrieved 2/28/19 from https://www.michigan.gov/mde/0,4615,7-140-22709_70117---,00.html.

Wang, M., & Dego, J.L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review, 29*(1), 119-140.