

Galileo Instructional Data System Pilot Project Evaluation

Final Evaluation



A Measurement Incorporated Company

Galileo Instructional Data System Pilot Project Evaluation

Final Evaluation

February, 2009

MAGI Services
A Measurement Incorporated Company
7-11 South Broadway, Suite 402
White Plains, NY 10601
(914) 682-1969 Fax: (914) 682-1760

Table of Contents

Introduction	7
Methodology	9
Key Findings	11
Evaluation Goal 1: To assess changes in teachers’ use of benchmark and formative assessment data in the classroom.....	11
Evaluation Goal 2: To assess changes in schools’ use of student intervention remedial and services and to determine the impact of services on assessment benchmark scores.....	13
Evaluation Goal 3: To examine the relationship between teachers’ use of benchmark and formative assessment data on performance on the assessments and on the MCAS.....	16
Evaluation Goal 4: To examine sustainability efforts for the use of benchmark assessments by districts and schools.....	20
Follow up to Case Study of Fitchburg District including highlights from B.F. Brown Visual Arts School.....	22
Conclusions and Recommendations	27
Appendix	31

Introduction

In 2005, the Massachusetts Department of Education (MADOE) contracted with Assessment Technology, Inc. to supply and implement an instructional data system named Galileo Online. This database system integrates curriculum mapping, assessment, reporting and analysis tools, and a standards-based grade book, which allows district and school staff, including teachers, an easy-to-use system for identifying trends in student learning and making improvements in classroom instruction. Key features of the data system are listed below.

- A bank of assessment items aligned with Massachusetts learning standards in Mathematics and English/Language Arts for grades 3-10
- Tools for constructing district benchmarking assessments and on-demand classroom assessments
- The ability to analyze, report, and longitudinally track student achievement data.
- Curriculum sequencing and pacing functions
- Standards-based electronic report card
- Parent and student web access to student assignments and performance data

The three-year pilot began in the 2005-06 academic year with 25 schools from 8 districts. The focus during the first year was on the development and use of benchmark assessments. ATI custom designed benchmark assessments based on each district's specification and the state's learning standards. Districts used the technology to administer the benchmark assessments quarterly throughout the school year. Immediate results on student performance were available for school staff to analyze and use for instruction, curriculum and student intervention/remedial services decision making.

Phase II began in 2006-07 and extended through the third year of the pilot. During the second phase, districts and schools continued to work with ATI to develop and refine quarterly benchmark assessments and teachers were encouraged to develop their own formative assessments for classroom use in between the quarterly school wide benchmark assessments.

The evaluation of the Galileo Instructional Data System Pilot Project (know hereafter as Galileo) was conducted by an independent research and evaluation company, MAGI Services, a Measurement Incorporated Company. The evaluation began in the second year (i.e., the 2006-2007 academic year) and extended through the third and final year of the state funding for the pilot project (i.e., the 2007-2008 academic year). The first year of the two year evaluation was designed to investigate 1) the quality and implementation of Galileo and 2) changes in student benchmark outcomes as a result of data use in the classroom and student intervention services. Following is a summary of the key findings from the interim report.

2006-07 Interim Report Findings

The 2006-2007 interim evaluation report showed a promising start to the pilot project. Districts and schools embraced and supported the use of benchmark assessments and the use of Galileo. District staff, principals/school leadership teams, and teachers indicated in surveys that they *agreed* that the Galileo assessment system was high quality; aligned with learning standards; useful to staff for informing school decision making and teachers' instruction; and was easy of use. Also important, district staff and principals/school leadership teams *strongly agreed* and teachers *agreed* that Galileo addressed an important need in the school. Finally, district staff *agreed* that ATI training supported their efforts to provide teachers with the necessary skills to use the system, generate reports, and analyze the data and, to a lesser extent, develop test items and access items for the formative assessment.

Teachers varied in their use of benchmark assessment data to make adjustments in the curriculum, differentiate instruction, evaluate student progress, and identify struggling students. Teachers were less likely to use formative assessments for the same purposes, with the exception of using them to evaluate the progress of students and identify struggling students. For those teachers who showed higher use of the benchmark data to inform instruction than their peers, students' third quarter benchmark assessment scores¹ were higher as compared to third quarter scores in classrooms where teachers reported lower use.

The interim report also showed that students performed higher on the third quarter benchmark assessment schools where a higher percentage of students (e.g., 51% or more) were enrolled in intervention and remediation services as compared to schools with lower percentages of students enrolled in these programs. While this finding underscores the value of intervention and remediation, the study found that more students were in need of intervention or remediation than those who actually received these services.

Goals of the Final Evaluation

The final year evaluation addressed the following goals: 1) to assess changes in teachers' use of benchmark and formative assessment data; 2) to assess changes in schools' use of student intervention and remedial services and the impact of services on benchmark assessments; 3) to examine the impact of teachers' use of benchmark and formative data on benchmark outcomes and to examine the relationship between benchmark and MCAS assessments; and 4) to examine sustainability efforts for the use of benchmark assessments by districts and schools.

The report begins with an explanation of the methodology used for the study. The key findings are organized by the four goals presented above. Following is a follow-up to a case study conducted on the district of Fitchburg and one of its middle schools which highlights some of the strengths of their benchmark assessment system. The report ends with conclusions and recommendations.

¹ Third quarter benchmark assessment data was used due to the low number of schools that administered the fourth benchmark assessments

Methodology

The final evaluation included the participation of district and school staff from all 7 districts and 24 schools² in the project. District staff (including Math Support Specialists and district leadership teams) and school administrators/school leadership teams completed an online survey that included items related to the quality of Galileo and use of the system at all levels. Math teachers from each of the 25 participating schools completed a paper survey with similar items.

Surveys were returned from 294 teachers from all 24 schools. The 2008 data was matched with the 2007 in order to make one-year comparisons, which resulted in 119 matched teacher responses. District surveys were returned from all 7 districts and school surveys were returned from 23 out of 24 schools.

The primary outcome variable was students' performance on the 4th benchmark assessment. Students in grades 5 through 8 were administered district-based, tailor made benchmark assessments quarterly throughout the school year; therefore, each student had up to 4 benchmark assessment scores. This student-level benchmark assessment data was electronically transferred from ATI, which included 2006, 2007 and 2008 data. Student-level MCAS assessment data was transferred from the MADOE and included 2006 and 2007, with the expectation that 2008 data would be transferred upon availability.

Matched t-test statistics were used to compare teachers' use of benchmark and formative assessment data to drive instruction from 2007 to 2008. HLM analyses were used to determine if teachers' use of benchmark assessment data would result in higher benchmark assessment scores. Hierarchical Linear Modeling (HLM) is a type of sophisticated regression model that takes into account the hierarchical structure of education data (students within classrooms within schools). It is a more conservative approach compared to multiple regression; however, the estimates are more precise when looking at students who share the same classrooms, schools, etc. Regression analyses were used to determine if benchmark assessment scores would predict students' performance on MCAS.

² One district and school dropped from the pilot in the final year of the project

Key Findings

Evaluation Goal 1: To assess changes in teachers' use of benchmark and formative assessment data in the classroom

Table 1 compares the 2007 and the 2008 mean scores for teachers' reported levels of use of the benchmark assessment data in their classrooms. Teachers were asked in a survey to rate their level of use in each area on a scale from 1 (not at all) to 6 (extensively).

Table 1
One-year Comparison of Teachers' Use of Benchmark Assessments

Benchmark assessment data is used to...	N	2007 Mean (SD)	2008 Mean (SD)	Difference
Adjust curriculum in areas where students encountered problems	115	4.03 (1.5)	4.50 (1.3)	.47*
Differentiate instruction based on student needs	113	3.99 (1.4)	4.30 (1.4)	.31*
Evaluate the progress of students	111	4.41 (1.3)	4.57 (1.4)	.15
Place students in instructional groups	115	3.54 (1.7)	3.87 (1.5)	.33*
Identify struggling students	115	4.50 (1.5)	4.63 (1.4)	.13

*differences were statistically significant, $p < .05$

- As seen in **Table 1**, teachers reported more uses of benchmark assessment data in the classroom in 2008 compared to 2007. Three out of five areas were statistically significant in favor of the 2008 reports. Specifically, teachers in 2008 compared to 2007 reported statistically higher use of benchmark assessment data to...
 - Adjust the curriculum in areas where students encountered problems (4.03 to 4.50),
 - Differentiate instruction based on student needs (3.99 to 4.30), and
 - Place students in instructional groups (3.54 to 3.87).

Table 2 compares the 2007 and the 2008 mean scores on teachers' reported levels of use of the formative assessment data in their classrooms. Similar to data presented above, the scores could range from 1 (not at all) to 6 (extensively).

Table 2
One-year Comparison of Teachers' Use of Formative Assessments

Formative assessment data is used to...	N	2007 Mean (SD)	2008 Mean (SD)	Difference
Adjust curriculum in areas where students encountered problems	115	3.86 (1.7)	4.20 (1.5)	.33
Differentiate instruction based on student needs	115	3.80 (1.7)	4.20 (1.5)	.40*
Evaluate the progress of students	115	4.07 (1.7)	4.43 (1.5)	.35
Place students in instructional groups	115	3.41 (1.8)	3.80 (1.5)	.39*
Identify struggling students	114	4.09 (1.8)	4.52 (1.5)	.43*

*differences were statistically significant, $p < .05$

- As seen in **Table 2**, teachers reported more use of formative assessment data in the classroom in 2008 compared to 2007. Similar to the benchmark assessment use, three out of five areas were statistically significant higher.
- Specifically, teachers reported significantly more use of formative assessment data to...
 - Differentiate instruction based on student needs (3.80 to 4.20),
 - Place students in instructional groups (3.41 to 3.80), and
 - Identify struggling students (4.09 to 4.52).

Summary

One year comparison data of teachers' use of benchmark and formative assessment data showed greater use of the data to drive instruction in 2008 compared to 2007. Specifically, teachers statistically increased their use of benchmark data to adjust curriculum, differentiate instruction and place students in instructional groups. Similarly, teachers were more likely to use formative assessment data to differentiate instruction and place students in instructional groups; however as compared to benchmark use, teachers showed a greater increase in using formative assessment data to identify struggling students. Even still, in all areas, teachers showed slightly greater use of benchmark assessment data as compared to formative assessment data in their instruction.

Evaluation Goal 2: To assess changes in schools' use of student intervention and remedial services and to determine the impact of services on benchmark assessment scores

The evaluation looked at two facets of student intervention and remedial services on the basis of principal surveys including 1) when services were provided and 2) the number of schools that were serving all students in need. **Table 3** outlines the percent of schools that provided supplemental intervention and remedial services in 2007 and 2008.

Table 3
One Year Comparison of the When Supplemental Intervention and Remedial Services were Provided at the School

	2007 Percent (frequency) of schools (N=22)	2008 Percent (frequency) of schools (N=22)
Before School	27% (6)	36% (8)
After School	82% (18)	73% (16)
Pull-out during the day	73% (16)	64% (14)
Saturdays	14% (3)	23% (5)
Summer School	68% (15)	45% (10)

- In 2008, more schools were providing services before school (36%) and on Saturdays (23%).
- Fewer schools were providing services in summer school (45%), after school (73%) and pull-outs during the day (64%) in 2008 compared to 2007.

Table 4 presents one year comparison data on the ratio of the percent of students who *needed* intervention services to the percent of students who *received* services by school. If a school provided intervention services to the same percentage of students who were in need of services (i.e., all students in need), the school received an “equal” label. If the school provided intervention services to fewer students than were in need, the school received a “less” label. The final column indicates if a school improved (↑), declined (↓) the “need to receive” ratio or remained the same with two “equal” (+) or two “less”(-) from 2007 to 2008.

Table 4
On Year Comparison of Changes in “Need to Receive” Status
By School

	2007	2008	One Year Change
Schools that Served all Students in 2007 and 2008			
School 2	Equal	Equal	+
School 7	Equal	Equal	+
School 9	Equal	Equal	+
School 11	Equal	Equal	+
School 16	Equal	Equal	+
School 17	Equal	Equal	+
School 18	Equal	Equal	+
Schools that Improved by Serving all Students by 2008			
School 6	Less	Equal	↑
School 10	Less	Equal	↑
School 20	Less	Equal	↑
School 21	Less	Equal	↑
School 12	Less	Equal	↑
Schools that Did Not Serve all Students in 2007 and 2008			
School 4	Less	Less	-
School 5	Less	Less	-
School 8	Less	Less	-
School 14	Less	Less	-
School 15	Less	Less	-
School 22	Less	Less	-
Schools that Declined by Not Serving all Students by 2008			
School 1	Equal	Less	↓
School 3	Equal	Less	↓
School 13	Equal	Less	↓
School with Missing Data			
School 19	n/a	Equal	

- In 2007, 48% (10/21) of schools provided intervention services to all students in need. One year later, 59% (13/22) of schools provided services to all students in need. This increase was not statistically significant.
- From 2007 to 2008, 5 schools improved their status by providing services to all students in need (i.e., moving from the “less” label to the “equal”).

To determine the impact of intervention and remedial services on benchmark assessment scores, schools were divided into two categories, those schools that did not provide services to all students in need compared to schools that did provide services to all students in need. ANCOVA analyses were performed with the first benchmark assessment as a covariate. **Table 5** outlines the estimated marginal means and standard error for both groups.

Table 5
Estimated Marginal Means (and standard error) for 4th Benchmark Assessments
for schools who did and did not provide intervention/remedial
services to all student in need

	N	Estimated Mean	Standard Error
Schools <i>not</i> providing services to all in need	7	1165.40	10.73
Schools providing services to all in need	7	1180.34	10.73

- As seen in **Table 5**, schools that provided services to all students in need performed higher on the 4th quarter benchmark assessment compared to schools that did not provide services to all in need. The mean score for schools that serviced all in need was 1180.34, which was 14.94 points higher than the schools that did not service all in need (1165.40). The difference between the groups was not statistically significant, which is likely due to the small sample size.

Summary

The data presented on student intervention and remedial services showed little change in services that were provided. There was a slight increase in the percent of schools offering services before school and on Saturdays. Nearly all schools provided services as part of the regular school day. More significant were the changes in the number of schools servicing students in need of assistance. In 2008, five schools increased their reach to meet more students in need and another seven schools continued to provide services to all students in need. All told, 59% of schools provided intervention and remedial services to all students in need by 2008.

Evaluation Goal 3: To examine the relationship between teachers' use of benchmark and formative assessment data on performance on the assessments and on the MCAS

This study tested the assumptions that 1) teachers' use of benchmark assessment data and formative assessment to drive instruction would increase benchmark assessment scores and 2) benchmark assessment scores would be related to and predict student performance on MCAS. To provide a more complete model, the evaluation also sought to identify factors that were related to teachers' use of the data. Armed with this information, districts and schools could make informed decisions about how to better support teachers' use of benchmark assessment data in their classrooms.

Use of Benchmark Assessment Data

The first step in the analyses was to identify factors that were related to teachers' use of benchmark assessment data. To this end, multiple regression analysis was performed and included the following variables: teacher background characteristics (number of years teaching and highest education degree earned), teachers' perception of the value of the Galileo system (i.e., the extent to which it addresses an important need in their school and their desire to continue use of assessments), the amount of training received to enable full use of Galileo technology, and teacher participation in benchmark development and review of data.

The analysis identified two variables that were significant predictors of teachers' use of benchmark assessment data (**Appendix A, Table 1** contains the technical results). They included:

- participation in the development and review of benchmark assessments—teachers who were more likely to participate in the identification and review of the target standards and the analysis of the benchmark data were also more likely to use benchmark assessment data to drive instruction,
- teachers' perception of the value of Galileo—teachers who strongly agreed that they Galileo assessment system addressed an important need in their school were also more likely to use the assessment data to drive instruction.

The second part of the analyses was to examine the relationship between teachers' use of benchmark assessment data and students' performance on benchmark assessments. The statistical procedure HLM was used for this analysis. HLM, or multi-level analysis, takes into account the nested structure of data, i.e., students within classrooms, by estimating the amount of variation at both of these levels (student and classroom levels). Due to the large sample size requirements for the use of HLM, all students' benchmark assessment scores from grades 5 through 8 were aggregated into one analysis. The benchmark assessment scores were equated across all grades prior to the analyses to allow for greater comparability across grades. Typically, the mean scores from grades 5th through 8th range from 1000 to 1300; therefore, the equating procedure involved subtracting each grade level down to the 5th grade mean score of 1000. For instance, 100 was subtracted from all 6th grade students, 200 from all 7th grade students and 300 from all 8th grade student scores.

While the primary interest was in teachers' use of benchmark assessment data to drive instruction, the analyses also included other variables that could impact student achievement. Together, the variables included in the analyses were students' first benchmark assessment score, number of teachers' years teaching, teachers' highest education degree, and teachers' use of benchmark assessment data to drive instruction.

Results revealed that teachers' use of benchmark assessment data, as reported in the teacher survey was significantly related to students' fourth quarter benchmark assessment scores. Students' first benchmark assessment scores were also related to their fourth benchmark assessment scores (the technical results of the analyses are provided in **Appendix A, Table 2**).

To give an example, teachers' scores on the use of benchmark data ranged from the lowest possible score of 10, indicating no or low use of data to the highest possible score of 60, meaning extensive use of data to drive instruction. If fifth grade teacher A reported the lowest level of benchmark assessment data use (i.e., 10), then his/her students average fourth benchmark assessment score would be expected to be approximately 1068.38. If fifth grade teacher B reported the highest level of data use (i.e., 60), then his/her students average fourth benchmark assessment score would be expected to be approximately 1164.88. The difference between scores on the low and high use would be almost 97 points, which is nearly equivalent to a one grade level difference.

Another way to look at this data is through the use of effect size. To this end, high and low groups were formed with teachers who scored at the 33rd percentile or lower on use of benchmark assessment data designated as the low group and teachers who scored at the 66th percentile or higher on use of benchmark assessments as high group. As seen in Table 6, the effect size for the high group was .385, which translated into a percentile gain of 15 points. This means that students from classrooms where teachers made higher use of the benchmark assessment data scored 15 percentile points higher than students from classrooms where teachers made lower use of benchmark assessment data to inform instruction

Table 6
Effect Size and Percentile Gain

	Effect size	Percentile Gain
1 st benchmark score	0.703	25%
Teacher Use of Benchmark Data	0.385	15%
Teacher Education*	-0.142	-5%
Year Teaching*	0.012	.5%

* non significant

The final step in the analyses was to link students' benchmark assessment scores to students' performance on MCAS via correlation and regression analyses. **Table 7** presents the correlations between the benchmark assessment data and the MCAS scaled scores for 2007; all correlations were significant.

Table 7
Correlations between Quarterly Benchmark Assessment Data and MCAS Scaled Scores

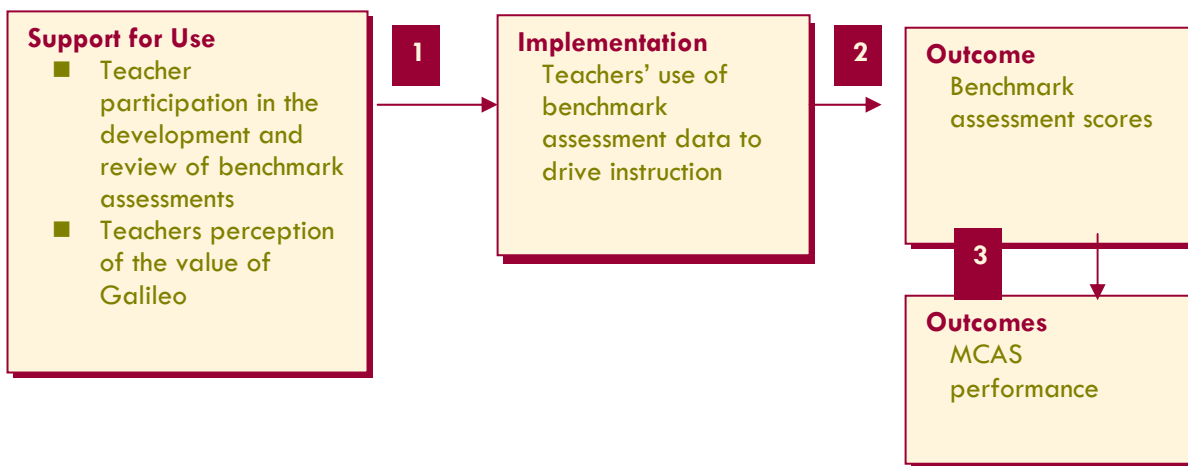
	Correlation with MCAS*
Benchmark 1	.541
Benchmark 2	.536
Benchmark 3	.549
Benchmark 4	.609

*correlations were statistically significant

The regression analyses included each of the benchmark assessments along with students’ demographic data (gender, ethnicity, poverty status and special education status). The analyses were significant; all benchmark assessment scores were significant predictors of MCAS scaled scores. This means that students who performed well on the benchmark assessments would also be expected to perform well on the MCAS. **Appendix A, Table 3** contains the technical results of the regression analyses.

Together, these findings help to outline a conceptual model for understanding how benchmark assessments operate within a school. This model is presented in **Figure 1**. Arrow 1 shows the relationships between variables that are labeled “support for use” which included teachers’ participation in the development and review of benchmark assessments *and* teachers’ perception of the value of Galileo and teachers’ use of benchmark assessment data to drive instruction.

Figure 1
Model for Use of Benchmark Assessment Data



Arrow 2 shows the relationship between teachers’ use of benchmark assessment data and the benchmark assessment scores. Arrow 3 shows the relationship between benchmark assessment scores and MCAS performance.

Use of Formative Assessment Data

The evaluation also examined the impact of teachers’ use of formative assessment data on benchmark assessment scores. These analyses yielded unexpected yet noteworthy findings. Teachers were designated into high and low implementing groups based on the first (low) and fourth (high) quartiles of their total aggregate score on the use of formative assessment data to drive instruction. ANCOVA analyses were performed on the fourth benchmark assessment scores (aggregating across all grades) while holding the first benchmark assessment scores constant to control for initial differences between students.³

As seen in **Table 8**, the low group outperformed the high group. The estimated marginal mean for the low group was 1162.93 whereas the mean for the high group 1144.98. The 17.95 difference was statistically significant. It appears that teachers who made less use of formative assessments had students perform higher on the benchmarks. There are various plausible explanations: 1) teachers may have spent too much time assessing math rather than teaching it, 2) the formative assessments did not

³ HLM analyses were not conducted on this data due to the preliminary findings that indicated a reverse effect on the benchmark assessment scores. Use of HLM would only serve to confirm this finding.

measure the same standards as the benchmark assessments and therefore did little to inform instruction related to benchmarks, 3) teachers who were more adept at integrating formative assessment into daily instruction were less likely to develop formative assessment tests, or 4) too much variability between different formative assessment tests developed by individual teachers to allow for fair comparisons.

Table 8
Comparison of the Mean 4th Benchmark Score
in High and Low Implementing Groups

	N	Mean*	Standard Error
Low Implementing Group	304	1162.93	5.35
High Implementing Group	219	1144.98	6.30

*estimated marginal mean after controlling for the math 1 mean

Even still, these findings prompted a deeper look into the number of re-teaching/enrichment hours that teachers reportedly spent in the classroom via teacher surveys and its impact on benchmark assessment scores. **Table 9** lists the estimated marginal means for each group after controlling for first benchmark assessment scores. Students who participated in the 4+ hours of weekly enrichment had the lowest mean score on the fourth benchmark assessment whereas students in the one hour or less and the 2-3 hour groups had comparable means. This data mirrors the formative assessment data in that more does not translate into higher performance. In this case, the data suggests that more than 3 hours of enrichment weekly may be counterproductive to improving achievement. Again it may be that more time was taken away from instruction on the target standards.

Table 9
Comparison of the Mean 4th Benchmark Scores
in Groups with Varying Hours of Weekly Re-teaching/enrichment

	N	Mean	Standard Error
One hour per week or less	507	1183.40	3.75
2-3 hours per week	472	1183.27	3.86
4 hours or more per week	115	1170.19	7.87

Summary

A model for understanding how benchmark assessments operate within a school was presented and supported. Teachers who valued Galileo and who participated in the development and review of benchmark assessments were likely to use the data to adjust their instruction in their classrooms, which in turn led to higher benchmark assessment scores by year end. Furthermore, the benchmark assessment scores were related to and significantly predicted achievement on the statewide assessment, MCAS.

On the flip side, students did not perform as well on the assessments when teachers made greater use of formative assessment data to inform instruction. Furthermore, students did not benefit from more than 3 hours of re-teaching weekly. These findings suggest that too much assessment and re-teaching may interfere with time that should be spent on teaching new standards.

Evaluation Goal 4: To examine sustainability efforts for the use of benchmark assessments by districts and schools.

Districts

District administrators were surveyed anonymously about the district's plans to sustain use of benchmark assessments. All districts plan to continue the use of Galileo beyond the final year of the pilot though under several varying conditions. Three districts will continue all aspects of the program regardless of funding, whereas another three will continue all aspects if funding could be secured. One district plans to continue some aspects of the program. To continue the use of Galileo four districts were in pursuit of other grant opportunities, three were leveraging other local, state, and federal funds, and another three were integrating key components of Galileo into the district improvement planning/budget process.

Table 10 lists the components/activities that are slated for continued use beyond the pilot year.

Table 10
Components of Galileo Districts would like to Continue Next Year

	Number of Districts (N=7)
District will continue to develop and administer benchmark assessments	7
District will continue to support teachers to develop and administer formative assessments	7
District will continue to analyze, interpret, and incorporate the benchmark assessment data into building-level decision-making	7
District will continue to analyze and interpret the benchmark assessment data to inform classroom instruction	6
District will continue professional development centered on the use of Galileo assessments.	6
District will continue regularly scheduled meetings on benchmark assessment data.	6

Schools

School administrators were also anonymously surveyed about the school's plans to sustain the use of benchmark assessments. Similar to districts, all schools would like to continue use of Galileo though under varying conditions. As seen in **Table 11**, four schools (17%) would continue some aspects of the program and another 3 (13%) would continue *all* aspects of the program *if* additional funding could be secured. Eleven schools (48%) intend on continuing all aspects of the program and two schools are continuing some aspects of the program.

Table 11
Schools Level of Commitment to Continued Use of Galileo

	Number (%) of Schools
School will continue <i>some</i> aspects of the program if additional funding could be secured.	4 (17%)
School will continue <i>all</i> aspects of the program if additional funding could be secured.	3 (13%)
School is continuing <i>some</i> aspects of the program.	2 (9%)
School is continuing <i>all</i> aspects of the program.	11 (48%)

Table 12 lists the components/activities that are slated for continued use beyond the pilot year. Nearly all schools (91%) would like to continue to develop and administer benchmark assessments. The vast majority (87%) would like to continue to support teachers' use of formative assessments, use of benchmark assessment data to inform building-level decision-making and classroom instruction and use of regularly scheduled meetings on benchmark assessment data.

Table 12
Components of Galileo Schools would like to Continue Next Year

	Number (%) of Schools
School will continue to develop and administer benchmark assessments	21 (91%)
School will continue to support teachers to develop and administer formative assessments	20 (87%)
School will continue to analyze, interpret, and incorporate the benchmark assessment data into building-level decision-making	20 (87%)
School will continue to analyze and interpret the benchmark assessment data to inform classroom instruction	20 (87%)
School will continue professional development centered on the use of Galileo assessments.	17 (74%)
School will continue regularly scheduled meetings on benchmark assessment data.	20 (87%)

Summary

Districts and schools were interested in continuing efforts to implement Galileo and various components of the program. All districts and the vast majority of schools would like to continue to develop and administer benchmark assessments, support teachers to develop and administer formative assessments, and analyze, interpret, and incorporate the benchmark assessment data into building-level decision-making.

Follow up to Case Study of Fitchburg District including highlights from B.F. Brown Visual Arts School

The City of Fitchburg with its 41,000 inhabitants is located close to the Massachusetts/ New Hampshire border and is 50 miles northwest of Boston. It is situated in a hilly topography on the Nashua River which once helped the city thrive in paper industry. With the recent exodus of the industry, the less prosperous city has diversified into other manufacturing and non-manufacturing industries.

The City School District of Fitchburg served approximately 5, 331 students of various backgrounds in 2008. B.F. Brown Arts Vision School is one of four middle schools in the district that served 486 students in grades 5-8. **Tables 12** and **13** present demographic information comparing the students in Brown, the district of Fitchburg and the state. The tables demonstrate that students who attended Brown and the district of Fitchburg were multi-ethnic and a larger percentage of them were Hispanic and Asian compared to the state. Over half of students came from low-income households and many were bilingual and limited English proficient, which was nearly double that of the statewide averages.

Table 12
2007-08 Ethnicity of Students at Brown, Fitchburg and the State

	% of Brown	% of Fitchburg	% of State
African American	8.4	7.0	8.1
Asian	10.5	6.7	4.9
Hispanic	34.2	37.8	13.9
Native American	0.0	0.1	0.1
White	44.7	46.7	70.8
Native Hawaiian, Pacific Islander	0.0	0.2	0.3
Multi-Race, Non-Hispanic	2.3	1.6	1.9

Table 13
2007-08 Special Populations at Brown, Fitchburg and the State

	% of Brown	% of Fitchburg	% of State
First Language not English	29.0	30.0	15.1
Limited English Proficient	11.7	12.8	5.8
Low-income	63.0	59.7	29.5
Special Education	21.6	19.4	16.9

Much of the history behind Fitchburg and Brown's involvement in the Galileo pilot project has been documented in a case study that was conducted by University of Massachusetts Center for Education Policy in 2007.⁴ To summarize from this report, Fitchburg's goals for participation in the Galileo pilot were to

- assess students' standards-based learning,
- generate conversations and actions around student learning and instruction,
- provide the district with just-in-time data and to

⁴ See Militello, Sireci, and Schweid (2007). Readiness, Fit and Coherence: The implementation of formative assessment products in three Massachusetts school districts. University of Massachusetts Center for Education Policy.

- provide the district with a data warehouse in order to make multiple sources of school data accessible.

In doing so, the district put into place three benchmark assessments using a “mock-MCAS” schedule. The assessments included 35-40 multiple choice in addition to short answer or open ended items. The items were matched to standards based on the district’s pacing guide, so that each standard was assessed by five items. Fitchburg partnered with its neighboring district of Leominster to score the open ended and short answer items. The districts also shared resources to bring in specialists to provide additional training to teachers on use of formative assessment. The district developed a formalized debriefing process that followed each administration of the benchmarks. The debriefing allowed teachers opportunities to analyze and engage in discussions about the assessments with the support of a Math Support Specialist.

The goal of the follow up study was to document the progress made during the final year of the pilot. To this end, interviews were conducted with the Assistant Superintendent, the Math Director and the Math Support Specialist at the district. The principal and several math teachers who taught 7th and 8th grades were interviewed at Brown. This report summarizes common themes that were generated through the analyses of the interviews. The common themes highlight the evolving progress made in the development of a formalized benchmarking system and the particular strengths of the system in Fitchburg and Brown.

Formalized Debriefing on the Benchmark Data

During the final year of the pilot, the district ramped up the debriefing process by 1) incorporating more professional development on assessment literacy and 2) encouraging teachers to discuss benchmark data with students. As for the additional professional development, the goal was to assist teachers in transferring more of the data discussions into instructional changes. Up until this point, there remained some uncertainty among teachers about how to integrate assessment with instruction rather than treating the two as separate events. During the debriefing meetings, the Math Support Specialists provided teachers with articles on formative assessments, discussed the uses of formative assessment, and then assisted teachers in creating action plans based on the data that they could take back to their classrooms to implement.

From the teachers’ perspective, the debriefing process was a big success on several counts. First, teachers appreciated opportunities to sit with the staff from the school and district to discuss the scores and possible reasons as to why students did or did not score well on benchmarks. For instance, some teachers voiced their concern that time was too limited to allow for full coverage of all the standards in a given quarter or in some cases, teachers felt that they didn’t teach a standard most effectively. In other cases, teachers felt that a particular question was low quality or too wordy. In each of these scenarios, debriefing sessions gave teachers the opportunities to get to the root of why students were having difficulty with some standards.

Secondly, teachers agreed that the debriefing sessions provided them with much appreciated opportunities to learn from each other and develop solutions on how to address students’ needs collectively. In one of their discussions, teachers agreed that some of the tactics used in the early grades did not benefit student in the long run. For example, the mnemonic, Please Excuse My Dear Aunt Sally, was used in the early grades to help students remember operations; however, it also falsely implies a predetermined order among inverse operations (e.g., multiplication and division). Through discussions about instructional strategies as a group within and across grades, teachers identified and agreed upon the most and least effective instructional strategies.

The second addition to the debriefing process involved efforts to encourage teachers to engage in assessment discussions with their students. According to the Asst. Superintendent, “we are shifting to a student focus by incorporating ‘accountability talks’ into instruction.” The district wanted to see the students engaged in self reflection so that they would take more ownership of their learning. They wanted the students to take an active role in self-guided goals for learning based on their assessments and to create student portfolios that include students’ self reflections on their progress in learning the standards.

Teachers admitted that they were still in the very beginning stages of engaging their students in self reflection. Some teachers were further along than others. All agreed that students should be involved in group discussions about their answer choices, both correct and incorrect, as a way of correcting misconceptions and learning different methods of solving math problems. For instance, one teacher gave the correct benchmark assessment answers to the students (after taking the test) and then gave them the opportunity to work in teams to derive the correct answers. Teachers would like to evolve in this area over the course of the next year.

Professional Development Centered on Formative Assessment

Professional development opportunities were aplenty at Fitchburg. The benchmark assessment data, in addition to teacher interest and needs, informed the type and content of professional development. One of the major foci of professional development this past year was on formative assessment; however, the district continued to provide learning opportunities in pedagogy and content.

The district brought in external supports for curriculum topic studies, which included a continuation of the “Got Math” workshop series. This series stemmed from the findings on the math test that teachers voluntarily took two years ago. It was at this time that the district realized that many teachers did not take math related courses during their higher education learning. The 2007-08 Got Math series continued coverage of math content that was offered in 2006-07 for new staff; however, additional topics were included in the series. The series ran for 8 weeks, three times throughout the year.

Teachers also participated in content institutes that were designed to bridge pedagogy and content with a focus on benchmark assessments. For instance, there was an upcoming week long institute that was being offered the week after school ended for summer session. The institute was to cover math content, pedagogy, analysis of student work and discussions about all types of formative assessments and uses. Teachers were also given the time to revisit benchmarking planners and to re-arrange standards for next year, based on the data that they received during the past year.

Professional development also came in the form of fun. Fitchburg partnered up with surrounding districts (Leominster, Gardner, and Lowell) to put on the “Amazing Math Race.” Teacher teams of three competed at various math tasks such as puzzles, logic problems and GPS activities. Fitchburg was the proud winner of the “Race.”

Re-teaching

Re-teaching was another part of Fitchburg’s benchmark assessment efforts. At Brown, a re-teaching block was provided every other day. Students were grouped according to benchmark assessment scores. The 5th grade piloted America’s Choice Navigator intervention program this past year with success and Brown will implement the program at all four grades next year. The program was designed to supplement and augment math instruction with benchmarked instructional modules that focused on specific math concepts. As part of the re-teaching block, math teachers worked with the “specials” teachers to incorporate math into art, music and gym (e.g., Fit Math in gym).

In addition to the re-teaching blocks, the school also conducted 8 weeks of “mini sessions” just prior to MCAS with small groups of students who were only two standards behind in an effort to bring them up to proficiency levels in time for the MCAS.

In 2008, Brown was awarded an Extended Learning Day grant, which would allow the school to extend instruction time from 6 hours a day to 7 hours and 45 minutes in the 2008-2009. As a result, re-teaching blocks will be provided on a daily basis. The extended day grant will also allow Brown to increase teachers’ common planning time from 30 minutes every 12 days to 150 minutes every 7 days. Unlike re-teaching efforts that occur at the expense of new instruction, re-teaching at Brown will occur in addition to regular classroom instruction.

Pretesting and Posttesting

A new feature in the district’s assessment efforts during the 2007-08 school year was a pretest and posttest that was given to students at the beginning and end of the year in addition to the three benchmark assessments. The test included all the learning standards for each grade level, e.g. all standards that they would learn for the upcoming year. At posttest, the students took the same exact test, thereby allowing teachers to gauge how much students learned from the beginning to the end of the year. Furthermore, the data was meant to be a springboard for planning instruction and prioritizing learning standards. As one teacher put it, “Now that we have one year of data (pre and post), we can start the next year with our students and show them how they scored and then set goals for the year.”

One interesting finding from the pretest, according to the Asst. Superintendent, was that there were some standards that the students did particularly well on, suggesting that “our students are not just blank slates. Indeed, the students knew more than the teachers thought that they knew.” This information suggested that teachers did not have to give the same emphasis to every learning standard for the upcoming school year. For instance, one teacher stated that her students did well on graphing data; therefore she assigned graphing data activities as homework or for students’ journals. This way, the students would continue to practice graphing data without monopolizing instruction time for new content. The principal of B.F. Brown also underscored this point. He said, “Time is of the essence because there are many standards to cover at each grade level. We can now use the pretest data to shuffle around some of our standards and reprioritize our teaching efforts.” Furthermore, “we have a clearly defined population of students, which allows us to hone in on what we need to teach.” These efforts underscore the value of using data to drive instruction.

Sustainability

In the words of the Assistant Superintendent, “everyone agrees that there is real value to formative assessment. The use of benchmark assessments has changed the culture of our schools.” Indeed, sustaining the assessments has been a “grassroots effort.” The district has committed funding to continue all features of the system, including supports for the teachers, e.g., the Math Support Specialist and the building level math coaches.

The principal of B.F. Brown commented that “our school has made leaps and bounds from last year in terms of use of assessments. Testing is less of a distraction and most teachers schedule it themselves.” The district has noticed that some schools have moved beyond the typical protocols of data analysis. “When a school asks for specific types of data, we know that they are trying new things and we asked them to share with other schools.” Teachers also agree that they were “past the mechanics of testing” and have furthered their exploration of the uses of Galileo. There was real buy-in from staff as a result of involving everyone in the efforts to formalize the benchmarking assessment.

Conclusions Recommendations

Learning is driven by what teachers and pupils do in classrooms. Teachers have to manage complicated and demanding situations, channeling the personal, emotional, and social pressures of a group of 30 or more youngsters in order to help them learn immediately and become better learners in the future. Standards can be raised only if teachers can tackle this task more effectively.

Black and William (1998)

Benchmark assessment is a powerful tool to help teachers instruct students more effectively and responsively. It provides teachers with ongoing feedback about students' learning and mastery of subject material which in turn, allows teachers to make responsive instructional adjustments, such as re-teaching, trying alternative instructional approaches, offering more practice, or referring students for intervention services as needed. Past studies have demonstrated that the use of formative assessment results in significant learning gains and is particularly effective with low-achieving students including students with learning disabilities.⁵

This data driven approach is in stark contrast to one shot, end of the year testing, which is the approach of statewide assessments; however, both types of testing can be used to complement each other. When both are aligned to the same learning standards, teachers can use benchmark assessments to steer students toward meeting proficiency or better on the statewide assessments. Much like instructional pacing calendars, benchmark assessments help to keep teaching and learning on track.

Such was the case with a benchmark assessment pilot project in the state of Massachusetts. Over the past 3 years, 7 districts and 24 schools have participated in formalized benchmark assessment testing through the work of ATI and its instructional data system named Galileo Online. This study culminates the outcomes of the pilot and draws the following conclusions about the use of benchmark assessments.

Teachers who participate in the development and review of the assessments and who value the system are likely to use the data to inform instruction in their classrooms. Research shows that when teachers are involved in the development and rollout of an initiative, they are likely to take ownership of it and are committed to its success. The same held true in this current study. Similar to Fitchburg, districts and schools might increase teacher commitment and the use of benchmark data for instruction via formalized debriefing sessions. Teachers in Fitchburg felt that formalized debriefing sessions provided them opportunities to have a voice in the process and to help shape the system. Indeed, the benchmarking system was described as a “grassroots effort.” This resulted from setting the tone for

⁵ Black, P. and William, D. (1998). Assessment and classroom learning. *Assessment in Education*, p 7-74.

using data in a non-threatening manner. Teachers were provided routine opportunities to discuss root causes of student performance, work out solutions together, and inform future assessments. The debriefing sessions were also viewed as a tool for enhancing teaching; the district built in supports to enable teachers to learn more about data driven instruction by folding in professional development opportunities.

Benchmark assessment data that is used to drive instruction results in increased mastery of learning standards by year end. Benchmark assessment data that is used to drive instruction results in increased mastery of learning standards by year end. This same finding was reported in the interim report in 2007; however, this report expanded and strengthened the finding by using conservative statistical analyses. HLM is highly regarded in the education research literature because it simultaneously takes into account the influence of student level and classroom level factors. In this study, the analyses showed that teachers who made full use of the data to drive instruction had the potential of raising student proficiency on the benchmarks one grade level compared to teachers who did not utilize benchmark data at all. Looking at the data from the standpoint of effect size, students from classrooms where teachers made greater use of benchmark data showed a 15 percentile gain over students from classrooms where teachers made less use of the benchmark data. If teachers continue to increase their use of data in the upcoming year, we would expect to see continued improvement in the benchmark assessment scores.

Galileo benchmark assessments are tied to and aligned with the MCAS statewide assessments. Students who perform well on benchmark assessments are expected to perform well on the statewide assessment. This study found significant correlations between the benchmark assessments and MCAS, which averaged .558, and regression analyses showed that the benchmark assessments were significant predictors of MCAS. In a separate study ATI reported larger correlations between the benchmark assessments and MCAS using a larger sample. Interestingly, the predictive validity of benchmark assessments has been addressed in a recent study conducted by MREL which Galileo was not included in. In contrast to the findings of this report, the MREL study found weak or no evidence of predictive validity in three out of four benchmark assessments.⁶

This study found other noteworthy findings. Specifically, more schools were servicing all students who were in need of intervention or remediation and in those schools, students demonstrated higher benchmark assessment scores than schools that serviced fewer students in need. It seems that benchmark assessments help to better identify students in need of intervention and remediation though this study can not verify if this resulted in more students being served without further investigation.

The study also found that teachers who made greater use of formative assessment data and reteaching efforts that exceeded 3 hours a week had students who performed lower on the fourth benchmark assessment after controlling for prior performance on the benchmarks. It is unclear as to why this happened, though there are several plausible explanations, namely that 1) too much time was spent assessing math and re-teaching standards rather than teaching new material or 2) the formative assessments or re-teaching efforts did not correspond with the standards measured by the benchmark assessments. Further investigation into the quality and type of formative assessments and re-teaching efforts are encouraged.

Finally, districts and schools who participated in this pilot planned to continue their use of benchmark assessments, which suggests that they are satisfied with the system and see the real value in the use of benchmarks.

⁶ Brown, R & Coughlin, E. (2007). The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region. REL 2007-No. 017.

All told, the findings in this study demonstrate strong evidence that the benchmark pilot project was a success. The benchmark assessments provided teachers with a tool to inform and shape their instruction throughout the school year with the goal of increased student mastery of learning standards. The benchmarks were reasonably aligned with the statewide assessments and could be used to inform districts and schools on how students will perform on the statewide assessments. Indeed, this benchmark assessment pilot project is well positioned to serve as a model for use in other districts and schools in Massachusetts and we recommend that the Department consider more dissemination.

Appendix

Appendix A: Statistical Results

Table 1
Coefficients (unstandardized and standardized) and t-values for predictors of teachers' use of benchmark assessment use

Model	Unstandardized Coefficients	Standardized Coefficients	Beta	t.	Sig.
	B	Std. Error			
1 (Constant)	9.528	3.306		2.882	.004
Years of teaching experience	-.043	.060	-.035	-.718	.473
Highest educational degree	-.850	.535	-.077	-1.588	.114
Value of Galileo	1.749	.264	.346	6.633	.000
Extent of training on use of Galileo technology	.224	.413	.029	.541	.589
Teacher participation in the development of BA	2.367	.242	.508	9.778	.000

Table 2
HLM Analyses

Level 1 Solution

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	1046.40	5.7029	63	183.49	<.0001
cMATH1	0.7404	0.02434	813	30.43	<.0001

Level 2 Solution

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	1049.08	5.4941	60	190.95	<.0001
cMATH1	0.7373	0.02618	813	28.16	<.0001
cBAUSE_08	1.9390	0.7165	60	2.71	0.0088

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Education	-9.7545	5.1457	60	-1.90	0.0628
Years Teaching	0.3962	0.6501	60	0.61	0.5445

Table 3
Regression Analyses

Regression: Student Demographics + bench1 07 on MCAS scaled scores

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.617 ^a	.381	13.374	2.882

^a Predictors: (Constant), Bench1, gender_07, ethnicity_07, sped_07, Freelunch_07

ANOVA^b

Model	Sum of Squares	Df	Mean Square	F	Sig.
1 Regression	869045.4	5	173809.075	971.707	.003 ^a
Residual	1410568	7886	178.870		
Total	2279613	7891			

^a Predictors: (Constant), Bench 1, gender_07, ethnicity_07, sped_07, Freelunch_07

^b Dependent Variable: mscalaeds_07

Coefficients^a

Model	Unstandardized Coefficients	Standardized Coefficients	Beta	t.	Sig.
	B	Std. Error			
1 (Constant)	168.550	1.415		119.083	.000
gender_07	.819	.303	.024	2.699	.007
ethnicity_07	1.988	.162	.118	12.275	.000
sped_07	-9.583	.451	-.195	-21.229	.000
freelunch_07	-5.189	.332	-.151	-15.624	.000
Bench1	.052	.001	.445	47.582	.000

^b Dependent Variable: mscalaeds_07

Regression: Student Demographics + bench2 on MCAS scaled scores 2007

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.618 ^a	.382	.382	13.375

^a Predictors: (Constant), Bench2, gender_07, ethnicity_07, sped_07, Freelunch_07

ANOVA^b

Model	Sum of Squares	Df	Mean Square	F	Sig.
1 Regression	866112.5	5	173222.499	968.323	.000 ^a
Residual	1399629	7824	178.889		
Total	2265741	7829			

^a Predictors: (Constant), Bench2, gender_07, ethnicity_07, sped_07, Freelunch_07

^b Dependent Variable: mscalaeds_07

Coefficients^a

Model	Unstandardized Coefficients	Standardized Coefficients	Beta	t.	Sig.
	B	Std. Error			
1 (Constant)	173.650	1.323		131.298	.000
gender_07	.899	.304	.026	2.953	.003
ethnicity_07	1.860	.163	.111	11.398	.000
sped_07	-10.010	.451	-.204	-22.200	.000
freelunch_07	-5.189	.332	-.151	-15.624	.000
Bench2	.048	.001	.443	47.540	.000

^b Dependent Variable: mscalaeds_07

Regression: Student Demographics + bench3 on MCAS 2007

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.621 ^a	.386	.386	13.323

^a Predictors: (Constant), Bench3, gender_07, ethnicity_07, sped_07, Freelunch_07

ANOVA^b

Model	Sum of Squares	Df	Mean Square	F	Sig.
1 Regression	879802.9	5	175960.583	991.295	.000 ^a
Residual	1399810	7886	177.506		
Total	2279613	7891			

^a Predictors: (Constant), Bench3, gender_07, ethnicity_07, sped_07, Freelunch_07

^b Dependent Variable: mscaleads_07

Coefficients^a

Model	Unstandardized Coefficients	Standardized Coefficients	Beta	t.	Sig.
	B	Std. Error			
1 (Constant)	173.157	1.312		132.015	.000
gender_07	.988	.302	.029	3.271	.001
ethnicity_07	1.858	.162	.111	11.491	.000
sped_07	-9.457	.450	-.193	-21.017	.000
freelunch_07	-5.315	.330	-.155	-16.083	.000
Bench3	.048	.001	.452	48.395	.000

^b Dependent Variable: mscaleads_07

Regression: Student Demographics + bench4 on MCAS scaled scores 2007

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.661 ^a	.436	.436	12.360

^a Predictors: (Constant), Bench2, gender_07, ethnicity_07, sped_07, Freelunch_07

ANOVA^b

Model	Sum of Squares	Df	Mean Square	F	Sig.
1 Regression	357877.1	5	71575.423	468.483	.000 ^a
Residual	462010.6	3024	152.781		
Total	819887.7	3029			

^a Predictors: (Constant), Bench2, gender_07, ethnicity_07, sped_07, Freelunch_07

^b Dependent Variable: mscaleds_07

Coefficients^a

Model	Unstandardized Coefficients	Standardized Coefficients	Beta	t.	Sig.
	B	Std. Error			
1 (Constant)	160.531	2.192		73.222	.000
gender_07	.829	.452	.025	1.832	.067
ethnicity_07	1.669	.234	.109	7.118	.000
sped_07	-7.485	.658	-.161	-11,369	.000
freelunch_07	-5.189	.509	-.139	-9.053	.000
Bench1	.048	.001	.443	47.540	.000

^b Dependent Variable: mscaleds_07

The logo for MAGI Services features the word "MAGI" in a stylized, bold, purple font. The letters are interconnected, with the 'A' and 'G' sharing a vertical stroke. To the right of "MAGI", the word "Services" is written in a smaller, purple, sans-serif font. Below the main logo, the text "A Measurement Incorporated Company" is written in a very small, purple, sans-serif font.

MAGI Services
A Measurement Incorporated Company