

Galileo and Interim Assessment

By Lynne Sacks, Graduate Student Intern, Office of Strategic Planning, Research, and Evaluation

Assessments—measurements of what students know and are able to do—are a central component of both the 1993 Massachusetts Education Reform Act and the federal No Child Left Behind (NCLB) legislation. While the MCAS, a summative assessment, offers an annual snapshot of the progress students are making, it does not provide an ongoing measure of how well students are mastering the curriculum for teachers to use to guide instruction. As the Educational Testing Service’s *Assessment Manifesto* explains, “to support learning, assessments must evolve from being isolated occasional events attached to the end of teaching to becoming an ongoing series of interrelated events that reveal changes in student learning over time.”¹ For this reason, many districts nationwide are developing a comprehensive approach to interim assessments.

Interim assessments come in several forms, but *benchmark* and *formative* assessments are among the most common. *Benchmark assessments* are structured assessments that are standardized within a district or school and are generally given several times a year. They are designed to provide information that is useful for student progress monitoring and for both programmatic and classroom-level decision making. While benchmark assessments vary, some are designed to align with a district or state mandated summative assessment. *Formative assessments* are less formal, given more frequently, and are designed to help teachers assess student understanding at the classroom or individual student-level. The data from formative assessments is not intended to be aggregated.

Both types of assessments can provide teachers with information about students’ performance on selected content standards that can then be used to modify instruction or provide students with additional support, if needed. The ultimate goal of interim assessments is to improve student achievement. High quality assessments are a necessary, but not sufficient tool for helping students meet standards. The effective interaction between assessments and instruction is ultimately what leads to improved student achievement.

The Massachusetts Department of Elementary and Secondary Education (ESE) engaged in a three year pilot program with districts across Massachusetts to implement Galileo Online, a system of interim assessments developed by Assessment Technologies, Inc. (ATI). This brief looks at recent evaluation findings from the Galileo program and the lessons that they might provide for the future use of interim assessments. Among the major findings from the evaluations:

¹ Stiggins, R. (2008). *Assessment manifesto: A call for the development of balanced assessment systems*. Portland, OR: Educational Testing Service (ETS) Assessment Training Institute.

- The design of the Galileo system and the technical characteristics of the assessments seem to be well established, at least in mathematics where it has been most heavily implemented.
- Overall, Galileo has received a positive response from teachers according to anonymous surveys.
- An external evaluation indicated that the use of student performance data from Galileo is linked to improved student-level outcomes measured by Galileo benchmarks.
- Research has not yet established a connection between the implementation of Galileo to improved school-level outcomes as measured by MCAS.
- Almost all of the original pilot districts have continued using the system beyond the grant funding period. Some of these districts have implemented systematic approaches to improving teaching and learning in which Galileo is an important component.

The Galileo pilot program

Galileo is a customized system of benchmark and formative assessment created by Assessment Technologies, Inc. (ATI), an Arizona-based assessment developer. In 2005, the Department initiated a pilot “to evaluate the capacity of an instructional data system to support the systematic improvement of teaching and learning.”² It selected Galileo through a competitive process. Twenty-five schools in eight districts (Chelsea, Chicopee, Fitchburg, Leominster, Lowell, New Bedford, Pittsfield and West Springfield) participated in Phase I of the project during the 2005—2006 school year. Approximately 15,000 students in these districts took part in the initial year of Galileo assessments. Nine districts³, including approximately 28,000 students in 67 schools, are currently participating in the pilot (Springfield alone accounts for 11,400 students and 38 schools).

Phase I focused on implementation of a comprehensive instructional data system to identify trends in student learning, improve classroom instruction, and ultimately raise student achievement. The state’s requirements for the assessment system included alignment between the assessment items and the Massachusetts standards and the ability to analyze and track student achievement over time.

In Phase II, during the 2006—2007 and 2007—2008 school years, the program’s two goals were to develop formal systems for student intervention and support and to engage teachers in classroom formative assessment. ESE provided professional development and ongoing assistance to district leadership teams in support of project goals. Districts were responsible for managing the implementation of Galileo, including the training of school administrators and teachers.

How does Galileo work?

With district input, ATI creates customized benchmark assessments from a secure item bank. Benchmark assessments are generally given district-wide three or four times a year

² See Galileo Pilot Project description at <http://www.doe.mass.edu/omste/galileo/default.html>.

³ In the 2007—2008 school year, Springfield and Gill-Montague joined the pilot and West Springfield dropped out.

and are based on the pacing guides of each district. The tests typically include eight standards with five items each and some districts also include open response items that are scored by teachers. In addition, teachers can use a separate, open-access item bank to develop less formal formative assessments. Some important features of Galileo are ease and speed of scoring—using a plain paper scanner—and flexible, comprehensive score analysis by student, class, test item, or standard.

Technical features of Galileo

For the student performance data produced by Galileo to be used effectively for data-driven improvement, the Galileo assessments must be reliable, or consistent, and valid. ATI calculates and reports reliability data for its benchmark assessments. Their analysis shows reliability coefficients between 0.86 and 0.95 for its benchmark assessments, indicating high levels of reliability.

One way to establish the validity of Galileo assessments, or the extent to which they are testing what they are intended to test, is to determine the correlation between performance on Galileo benchmark assessments and subsequent MCAS tests. ATI conducted a correlation study in five Massachusetts school districts during the 2005—2006 school year. The study used equipercetile equating to set cutpoints on the benchmark assessments that corresponded with cutpoints on the MCAS mathematics exam for each of the grade levels included in the study. ATI found that meeting the standard on the Galileo benchmark assessments generally predicted meeting the MCAS standard (i.e., scoring Proficient or Advanced) with 80 to 90 percent accuracy, as shown in Table 1. The benchmark assessments are most reliable in predicting whether a student will fail the MCAS for students who consistently meet or fail to meet the standard on the benchmarks.

Table 1: Percentage of students whose standards mastery was accurately forecasted for mathematics, by grade

Grade	Range of accuracy by district (low to high)	Mean
5 th	78% – 89%	83%
6 th	81% – 90%	86%
7 th	86% – 91%	88%
8 th	89% – 93%	91%

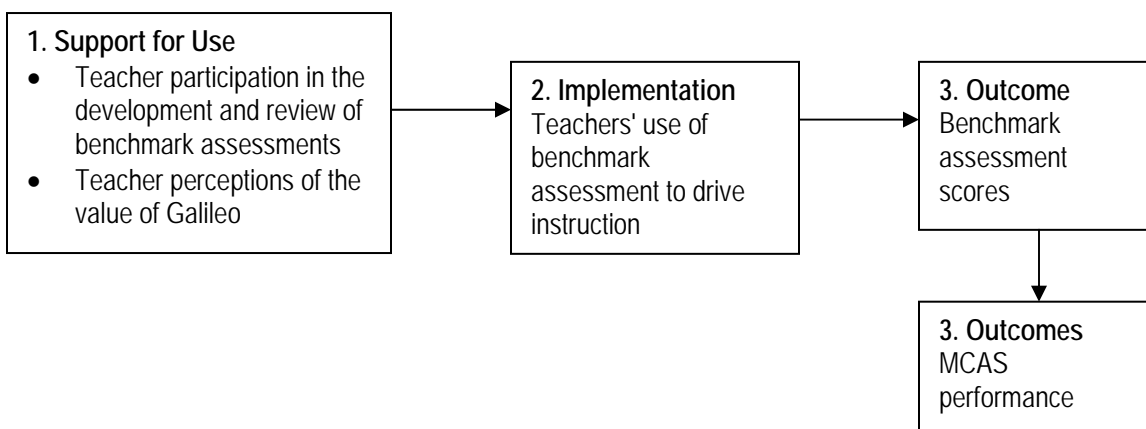
From Assessment Technology, Inc., “Assessing student risk of not meeting Massachusetts state standards,” January, 2007

Based on these annual analyses by ATI in mathematics and ELA, and a similar analysis in mathematics by the external evaluator MAGI Services, it appears that Galileo benchmark assessments are effective at predicting which students will pass MCAS tests and which students will not. These findings imply predictive validity for the Galileo benchmark assessments and suggest that Galileo can help schools identify those students who are most at risk of failing the MCAS early enough in the year for interventions to be implemented.

Program evaluation

MAGI Services conducted evaluations of the Galileo pilot during 2006—2008. Since districts are largely using Galileo to assess students' progress in mathematics, the evaluations focus on mathematics performance. Survey data from administrators and teachers participating in the Galileo pilot program provide information on program quality, support for its use, participation levels, and student interventions. Based on their study, MAGI developed a logic model representing Galileo's implementation:

Figure 1: Model for use of benchmark assessment data



Adapted from MAGI Services, "Galileo instructional data system pilot project evaluation, final evaluation," February, 2009.

MAGI's initial evaluation, using data from the 2006—2007 school year, offers insights into possible gaps between providing the assessments and their impact on changing instruction. Results from a teacher survey that MAGI conducted show strong positive responses by teachers to questions about the quality of the assessments and reports, but less positive responses to questions about implementation. For example, the mean teacher score for the appropriateness of the difficulty and rigor of the benchmark assessments is 4.12 out of 5 and for reflecting the range of cognitive skills covered by state standards is 4.14 out of 5. Mean teacher scores for the usefulness of assessment reports for classroom-level and student-level planning and decision-making are both 4.26 out of 5. However, mean teacher scores on the amount of time available for using Galileo data are much lower: 3.71 for time to review data from the assessments, 3.39 for time to plan instructional activities to address areas of student weakness, and 3.35 for time to collaborate with other teachers to analyze assessment data from the Galileo assessments. This suggests that there are structural barriers to fully leveraging the potential of Galileo and similar systems.⁴ Results from the 2007—2008 school year show significant increases from the previous year in the reported use of Galileo data to inform instructional practices, so it may be that over time these challenges can be alleviated.⁵

⁴ MAGI Services. (September, 2007). "Galileo instructional data system pilot project evaluation, interim report." See <http://www.doe.mass.edu/omste/galileo/0907interim.pdf>.

⁵ MAGI Services. (February, 2009). "Galileo instructional data system pilot project evaluation, final evaluation." See <http://www.doe.mass.edu/omste/galileo/06-08eval.pdf>.

The final evaluation, using data across the 2006—2007 and 2007—2008 school years, also compares scores on the third Galileo benchmark assessment between students based on the level of implementation in their classrooms. High-implementing classrooms are defined as those whose teachers ranked in the 66th percentile or above in their reported use of benchmark data to inform instruction, while low-implementing classrooms are defined as those that ranked in the 33rd percentile or below on the implementation scale. The study uses hierarchical linear modeling (HLM) to statistically control for mitigating factors such as prior achievement and teaching experience. The table below expresses the relationship between certain variables and a student's score on the third benchmark in terms of an effect size that is translated into a percentile gain. These findings suggest that after controlling for other factors, "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."⁶

Table 2: 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%
Number of years of teaching experience*	0.012	0.5%

*Not statistically significant.

These results suggest an important link between the way that teachers utilize an assessment and data system like Galileo and their students' performance.

Determining the effect on school-level MCAS performance

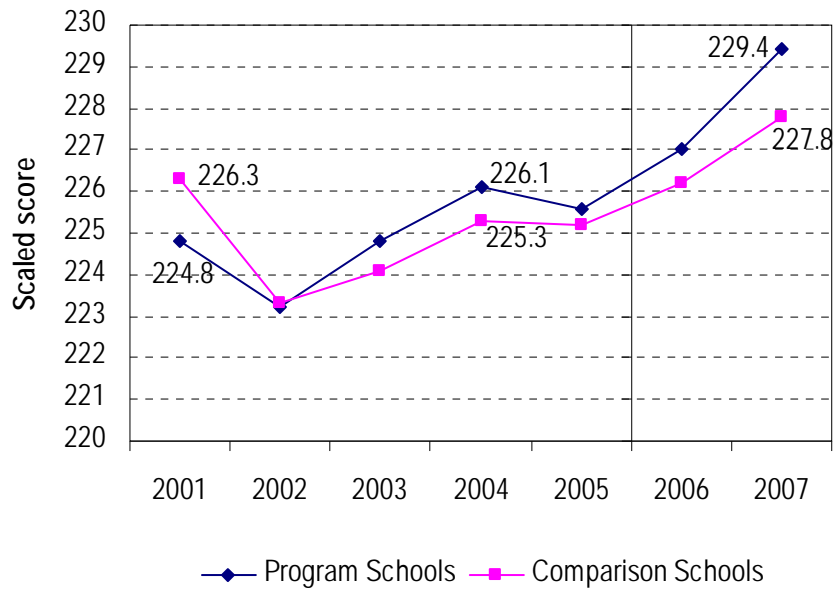
Ultimately, the goal of implementing a program like Galileo is to improve student achievement across a school or district, as demonstrated by increased scores on the MCAS. To assess this, ESE asked the Regional Educational Laboratory, Northeast and Islands (REL-NEI) to do a preliminary analysis of MCAS mathematics results for schools participating in the program.⁷ The evaluation matched each school in the pilot program with two comparison schools and examined MCAS score improvements over time, comparing across the treatment and control groups. The analyses show that the scores of eighth grade students in schools participating in the Galileo program increased over prior years' test scores in both the first and second years of implementation. The score

⁶ Ibid.

⁷ Henderson, S., Petrosino, A., Guckenbug, S., & Hamilton, S. (April, 2008). "A second follow-up year for *Measuring how benchmark assessments affect student achievement*," (REL Technical Brief, REL Northeast and Islands 2008–No. 002). Washington, DC: U.S. Department of Education. Retrieved from <http://ies.ed.gov/ncee/edlabs>

improvements were statistically significant in both years. However, the schools that were used as a control group also saw statistically significant improvements. While the schools using Galileo saw larger increases in scores, as shown in Figure 2, the differences between the Galileo and non-Galileo scores were not significant.

Figure 2: Scaled eighth-grade MCAS mathematics scores for program and comparison schools, 2001–2007*



Adapted from REL-NEI, "A second follow-up year for *Measuring how benchmark assessments affect student achievement*," April, 2008
 *The vertical line represents when Galileo testing began.

It is possible, however, that limitations of REL-NEI’s study are affecting the results. First, the control schools used in the study are likely implementing alternate reforms also intended to raise student achievement, including the implementation of district-wide assessment systems. The study did not look at what the control or treatment schools were doing other than whether they were part of the Galileo pilot or not. Therefore, we are likely seeing the difference between two different approaches to improvement rather than between no intervention and the use of Galileo. Second, as the study authors acknowledge, the control schools may differ from the implementation schools in ways that affect the results. For example, the study schools and comparison schools had statistically significant differences in scores on the mathematics Composite Performance Index (CPI) and in percentages of African-American students. The comparison schools, taken as a group, had higher initial CPI scores in mathematics than the program schools as well as a higher average percentage of African American students. The differences were statistically significant in both cases. Third, the scores are not disaggregated either by level of program implementation or by subgroup, which makes it difficult to tell whether some districts or groups of students have had greater gains in test scores than others.

The importance of district implementation

The effectiveness of programs like Galileo that focus on interim assessment depends to a large extent on the quality of implementation. A case study of Fitchburg Public Schools in the final external evaluation report provides a narrative account of effective implementation.

In Fitchburg, the use of Galileo goes far beyond collaborating with ATI to develop benchmark assessments aligned with the district's pacing guides. Following each benchmark assessment, district and school personnel meet with mathematics teachers to participate in a formal debriefing process to examine the test data, with a particular focus on how instruction can be modified to address weaknesses in student mastery. Benchmark assessments include open response items and all teachers receive training in scoring. Structures have been created within the school day to provide additional time for students to be regrouped based on assessment results so that they can be provided with targeted intervention, including both reteaching and enrichment. District leaders reported at the ESE Curriculum and Instruction Summit in December 2008 that they have built upon the use of Galileo assessment data by providing extensive professional development on formative assessment instructional techniques to engage students in the assessment and improvement process.

Fitchburg represents one of several pilot districts that have built a balanced assessment and intervention system in which Galileo serves as an important component to support the systematic improvement of teaching and learning.

Implications

Evidence from the evaluations indicates that the use of student performance data from Galileo is linked to improved student level outcomes as measured by Galileo benchmarks. Galileo can also predict how students will perform on the MCAS, though no link has been drawn between the implementation of Galileo to improved school-level outcomes on MCAS. Surveys show teachers believe that Galileo is a rigorous and useful assessment instrument and that they are working to use the information to a greater extent to guide instruction.

The effectiveness of any assessment ultimately depends on how the results are used to influence instruction. While more research is needed to determine the precise effects of interim assessment on student achievement, there is reason to believe that it can be a useful tool. This report has mostly focused on the Galileo assessment system itself, with some self-reported evidence on the use of data by individual teachers, but professional literature and anecdotal evidence point to the importance of district- and school-level systems of intervention for sustained improvement. ESE will need to consider the findings from the Galileo pilot and other research evidence as it determines whether, how, and to what extent the agency will have a role in shaping how interim assessments are used in the Commonwealth in the future. *

Lynne Sacks is a doctoral candidate at the Harvard Graduate School of Education and was an intern in the Office of Strategic Planning, Research, and Evaluation at the Massachusetts Department of Elementary and Secondary Education in 2008.