# Which One is “Just Right”? What Every Educator Should Know About Formative Assessment Systems[*]

Matthew Militello
Neil Heffernan

**Abstract**

Recent legislative and local school accountability efforts have placed a premium on the collection, analysis, and use of student assessment data. Beyond state-based summative student achievement data, schools are searching for in-year information about student performance. As a result, school leaders are searching for the formative assessment system that meets their needs; they are looking for the formative assessment system that is just right.

NOTE: This module has been peer-reviewed, accepted, and sanctioned by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration. In addition to publication in the Connexions Content Commons, this module is published in the International Journal of Educational Leadership Preparation, [1] Volume 4, Number 3 (July - September, 2009). Formatted and edited in Connexions by Theodore Creighton, Virginia Tech.

## 1 Introduction

The pressure to improve student performance in schools is currently battering education on two fronts. First, the public demands our education prepare our students to compete in a global economy. International assessment results reminds us of the gap between our students and their international peers (see Friedman, 2007). The second front has been educational law. Assessment and educational accountability were directly

[*] Version 1.2: Aug 20, 2009 3:28 pm +0000

[†] http://creativecommons.org/licenses/by/3.0/

[1] http://ijelp.expressacademic.org

linked in 2001 with the No Child Left Behind (NCLB) legislation. This legislation ushered in a new era of accountability rooted in the collection, analysis, and use of student assessment data for educational improvement.

As a result, school districts and their leaders must raise the stakes on educational assessments. Beyond using state-level data, leaders search for other assessments to use within schools during the school year to determine student achievement growth between state-level assessments. Two things are now clear in schools today: (1) the assessment "tail [is] definitely…wagging the curriculum/ instruction canine" (Popham, 2004, p. 4), and (2) it is up to school leaders to find a formative assessment system that meets the needs of their students, teachers, and parents. More specifically, schools are searching for an assessment that can show within year *growth* on learning objectives, *diagnose* within year learning needs of students, and to *predict* achievement level on the state assessment.

The demand for formative assessment is clear. Schools are employing a variety of Formative Assessment Systems (FAS) ranging from "home-grown" tests created by teachers themselves to over 20 commercially packaged assessment systems costing $12 or more per student (Militello, Sireci, & Schweid, 2008). FAS combine tightly knit assessment instruments, data-warehousing, analysis, and reporting batteries (Sharkley & Murnane, 2006).

Formative Assessment Systems (FAS) are a fast-growing and under-studied phenomenon, with major implications for educator practice. Moreover, school leaders are uniquely positioned to access (e.g., purchase), assess (e.g., monitor and evaluate), and support and resource (e.g., train and develop) assessment data in schools. The purpose of this article is to provide a framework for school leaders who have to make important decisions about formative assessment systems.

## 2 The Goldilocks Dilemma

Fit matters. Finding the "just right" match between what something has to offer and the intended utility should be of the upmost importance to consumers. The story of Goldilocks provides an allegory of the search for the "just right" fit. Goldilocks's metrics for fit included the temperature of porridge, size of chair, and comfort of bed. As educators and policymaker metrics for fit of an assessment system should include both the *purpose of the assessments* (e.g., properties of assessment including validity) and *the intended uses by school educators* (e.g., lesson planning) (Militello, Sireci, & Schweid, 2008).

Educational assessments come in many shapes and sizes. Large-scale assessments can be either criterion-referenced (e.g., NCLB state-level assessments) or norm-referenced (e.g., TIMSS, NAEP, SAT, ACT). Small-scale assessments tend to be conducted and analyzed at the classroom level and may include valuable qualitative understandings of the teachers. Somewhere in the middle reside formative assessments. Wiliam (2006) describes, an assessment is formative only if information about what is being assessed results in change that would otherwise not occur. Under this guideline, "An assessment of curriculum is formative if it shapes the development of that curriculum. An assessment of a student is formative if it shapes that student's learning" (p. 3). Here the definition of formative assessment becomes more detailed to the utility for teacher to understand if and how students are learning. This type of formative assessment has been called cognitively diagnostic. Leighton and Gierl (2007) define cognitively diagnostic assessment as a means to "measure specific knowledge structures and processing skills in students so as to provide information about their cognitive strengths and weaknesses" (p. 3). While "educators are demanding . . . they receive instructionally relevant results from any assessment . . . and that these assessment be sufficiently aligned with classroom practices to be of maximum instructional value," (Huff & Goodman, 2007, p. 24) there exist little evidence that there are assessment systems available to meet these local demands.

The issue of assessment fit is more difficult when looking at teacher as the end-users. While most formative assessment systems claim utility for classroom teachers, there is little evidence that the right assessment data is being provided them. Two frameworks help further unpack the search for an assessment system that meets the needs of teachers. First, the National Research Council's "assessment triangle" offers anchors of assessments including the ability to: (1) diagnose student cognition within a specific subject area, (2) conduct student observations that elicit responses from students and offer evidence of competencies, and (3) make fair and accurate inferences about student achievement (National Research Council, 2001). Additionally, the literature has generated normative features of the meaningful, effective use of teacher-level formative assessments including (see Brunner et al., 2005; Coburn & Talbert, 2006; Heritage, 2007; Kerr, Marsh, Schuyler Ikemoto, Darilek, & Barney, 2006; Marshall, 2008; Militello & Sykes, 2006; Murnane, Sharkey, & Boudett, 2005; Streifer & Shumann, 2005; Wayman & Stringfield, 2006; Wylie & Ciofalo, 2008):

- Assessments that are linked to a curriculum that is aligned with the district scope and sequence and state curricular benchmarks;
- Assessments that provide timely, student diagnostic-level data;
- Ability to disaggregate data with other datasets (e.g. other student achievement data, perceptional data, etc.) and to easily access and communicate reports with a variety of audiences, and
- Availability of on-going professional development and immediate on-site assistance to translate data into instructional knowledge.

In an effort to illustrate why this issue of fit is important for current and future educational leaders we describe the characteristics of three formative assessment systems currently available. Next, we analyze the characteristics of each system in relationship to the concept of fit.

## 3 Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP)

The MAP Math test consists of 52 multiple-choice items. According to the NWEA Technical Manual (2005), there are at least 1,500 items in the item pool for each subject area tested in a state with a minimum of 200 items per goal area. MAP tests are supposed to be un-timed, but the manual recommends allotting 75 minutes to ensure all students will finish without being rushed. A one-month testing window is necessary because each student needs access to a computer terminal. MAP utilizes computerized-adaptive testing technology. This assigns test items to students by matching the difficulty of an item to the achievement level of the student. Another key feature is that all items within each subject area pool are calibrated onto a common scale. This scaling allows students to be placed onto this same scale even though they respond to different items. The common scale also allows for analysis of student growth across time. State-specific item pools are created from the "universe" of all MAP items so that the pools used for a particular state are best matched to the state's curriculum frameworks.

## 4 ATI's Galileo

Galileo is a system for building benchmark assessments available through Assessment Technology Incorporated (ATI). The Galileo system has a large bank of items and is designed to enable ATI to work collaboratively with a district to design an assessment system that is aligned with local instruction and informs curriculum planning. ATI begins by working with a district to identify the number of benchmark assessments to be administered in a given year and the standards (objectives) to be measured at each subject area in each grade level. Districts are able to custom order assessments through Galileo's "Educational Management System." ATI has an online Benchmark Planner in which the district defines the assessment goals, specifies the standards to be measured and the number of items per standard, and reviews preliminary versions of the assessments.

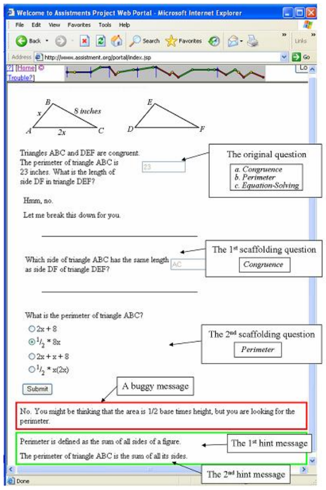
Galileo has several reports that summarize students' performance at the individual and aggregate levels. Aggregate level reports can be produced at the class, school, and district level, and many can be created interactively to suit the user's needs. A primary report for the benchmark assessments is the Developmental Profile Report. This report lists and describes all the standards (objectives) measured and provides an achievement level classification for each student for each standard.

# 5 The ASSISTment System

## 5.1

The ASSISTments system is an Internet based tool developed at Worcester Polytechnic Institute (WPI) and Carnegie Mellon University. The ASSISTment System blends assess**ment** and instructional **assist**ance. The system **assesses** student learning asking middle school level mathematic questions (drawn from a bank of more than 1400 questions generated from project staff, local teachers, and state-level assessment release items). The system **assists** students by providing immediate feedback and scaffolding questions if a wrong answer is given.

The Screenshot below shows a sample ASSISTment item that involved understanding algebra, perimeter, and congruence. Like a human tutor, the ASSISTment System breaks an item down to the individual steps needed to solve the problem. By tracking where students' understanding ends, the system can track the specific components of a student's learning and give teachers data on the precise pieces of knowledge students have and don't have mastered. For example, a student is initially presented a real state assessment item that has several skilled needed (congruence, perimeter and equation solving) to solve it correctly. If the student had gotten the question correct he would have gotten credit toward all three skills. If a student gets the question incorrect he would be given a question that probes his understanding of congruence. If the student gets that correct, the student gets credit towards understanding congruence. If he does not get the question correct, he can ask for help. The student is forced to eventually ask for as much help as needed so they can answer the question and move to the next scaffolding question focused on perimeter. The screenshot shows a buggy message (triggered by common student misconceptions) as well as two hint messages. After the student gets the second scaffolding question correct he will be given one focused on the algebra question solving x (Not shown in the Screenshot).


Welcome to Assistments Project Web Portal - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Links
Address
Go
[?] [Home]
[Trouble?]
Lo
B
x
8 inches
A
2x
C
E
D
F
Submit
Done

Address: http://www.assistment.org/portal/index.jsp

Triangles ABC and DEF are congruent. The perimeter of triangle ABC is 23 inches. What is the length of side DF in triangle DEF? 23

The original question
a. Congruence
b. Perimeter
c. Equation-Solving

Hmm, no.

Let me break this down for you.

Which side of triangle ABC has the same length as side DF of triangle DEF? AC

The 1st scaffolding question
Congruence

What is the perimeter of triangle ABC?

○ $2x + 8$
● $^1/_2 * 8x$
○ $2x + x + 8$
○ $^1/_2 * x(2x)$

The 2nd scaffolding question
Perimeter

A buggy message

No. You might be thinking that the area is 1/2 base times height, but you are looking for the perimeter.

Perimeter is defined as the sum of all sides of a figure.

The 1st hint message

The perimeter of triangle ABC is the sum of all its sides.

The 2nd hint message

A number of reports can be immediately accessed by teachers. An item-level report can be generated to see student responses on specific questions. Teachers can also generate student-level reports. In this report, teacher can analyze individual student progress and provide narrative communication to the student and their parents/guardians. Teachers can see the number of items taken, percentage correct, time interval, and hint requests. Finally, the ASSISTments system has a parental and teacher notification features. Parents are sent an email from the system that describes their child's performance on ASSISTments. Additionally, when teachers assign nightly homework with the system, they are sent an email that describes who completed the homework and specific items each student got right or wrong.

## 6 Assessments In Action or Inaction

The actual use of the three aforementioned FAS were studied (see Militello & Schweid, 2009; Militello, Sireci, & Schweid, 2008). Table 1 below provides a summary of a number of features of the three FAS. In regard to NWEA's MAP, this FAS is most appropriate for use at the district-level. Specifically, because the data generated by the assessment gives district administrators longitudinal scores, patterns from year-to-year can assist them in their decision-making (e.g., professional development opportunities). However, because the MAP data does not generate item-level data reports, teachers find little utility in its use. ATI's Galileo provides a rich set of interim or benchmark assessments that school-level educators are able to use to monitor student gains on what was recently taught. This FAS also puts pressure on teachers to teach to the curriculum. Of the three systems, teachers used only one, the ASSISTment System, in a real-time, cognitive diagnostic manner.

**Summary of Formative Assessment System Technical Features**

| Feature | NWEA's MAP | ATI's Galileo | ASSISTments |
|---|---|---|---|
| Score Reporting (Time & Access) | Quick-24 hrs. (teacher access) | Quick- 24 hrs. (teacher access) | Immediate(teacher access) |
| Content Validity/Alignment | Moderate | Excellent | Excellent |
| Diagnostic Scores | Moderate | Good | Excellent |
| Norm-Referenced Info | Good | Unavailable | Unavailable |
| Measurement Precision | Excellent | Unknown | Unknown |
| Equated scales | Excellent | Unknown | No |

**Table 1**

Teachers using the ASSISTment System create *customized homework* problem sets. This can be as simple as a teacher asking student to input answers from textbook problem into the ASSISTments System. In class assignments can also be administered for *instant response*. The ASSISTment System can be used by all students simultaneously and the results can be displayed anonymously. This included open response items. Teachers also use the system to make student *work more accessible* with the variety of reports available. These reports include reports that are automatically generated and attached to emails to both parents and teachers. The parent notification email provides a detailed account of ASSISTment completion and success. For teachers, reports are emailed regarding homework completion and success.

Teachers also have *open source coding* access to develop their own items. This is an iterative process where questions along with scaffolding sub-questions and hint messages are constantly created and vetted. In summary, this system collects data efficiently and provides student-level diagnostic results as a means to impact a teachers teaching and students learning[2]. The data allows for classroom-level decision-making by teachers for day-to-day instruction (e.g., reteaching strategies) as well as long-term state-assessment goals (short-term regrouping). ASSISTments also develops teacher ownership of the data as they spend time on analyzing data not grading and creating their own local questions to be uploaded. For students, ASSISTments is not just another test. Rather an important benefit of the system is that students receive tutoring while being assessed.

[2] A number of studies have been conducted to evaluate the impact of ASSISTment use. Initial studies found that ASSISTments in a good assessor of student knowledge (Feng & Heffernan, 2006; Feng & Heffernan, 2007; Razzaq, Heffernan, & Linderman, 2007). ASSISTments also have prognostication powers in regard to future state-level examinations (Feng, Beck, Heffernan, Koedinger, 2008). Perhaps most importantly, ASSISTments provides teachers with useful student-level knowledge they can reflect on, talk to colleagues about, and adjust their pedagogy.

## 7 Conclusion

The Goldilocks fable helps the readers understand the search process educators must undertake to find the "just right" fit of a formative assessment system. The brief description of the three assessment systems and their use in schools spotlights that fit (and misfit) matters. The answer to the question, "Which formative assessment should we use?" should be: It depends. The utility of a formative assessment system is predicated on the end-use. That is, how educators want to use assessment data should guide their consumerism in the selection of a product. Figure 2 below is a model of assessment fit. In this model the specific assessments fit specific data generated and end-use.

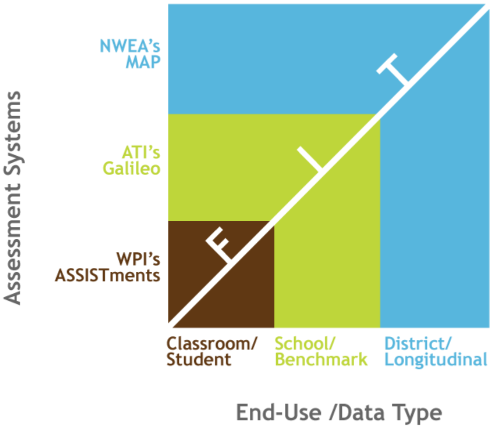


**Figure 1:** Fit of the Three Assessment Systems

Currently, the ubiquitous assessment data are over-hyped and under-utilized; yet, schools continue to be fed a steady diet of tests. The real power of assessments lies in the transformation of raw data and disseminated information into explicit knowledge to guide instructional improvement. The high demands to use data, coupled with the inadequate training and pervasive fear, result in the phenomenon of pedagogical practices geared toward tests and less on good instructional practice (Earl & Katz, 2002). As schools send out search parities to find and implement these systems, they would be well off to take with them the importance of fit.

Formative assessment companies aren't inherently bad. They have filled a niche. However, schools and educational policymakers need to be better consumers. Importance must be placed on the intended use of FAS and the characteristics of the systems must be assessed. Armed with such information will allow them to make more informed decisions. School leaders have a role to play in the process of formative assessment understanding and implementation. School leaders would be well served if they (1) understood the concept of assessment fit, (2) build teachers' capacity to use assessments that provide student-level diagnostic data, (3) provide adequate resources and support mechanisms, and (4) monitor the use of assessment data.

Finding the "right fit" between the purpose of an assessment system and the intended uses by local educators is an important issue. Asking teachers to use data to inform their teaching in order to advance student achievement requires careful consideration. The constant press to use "data" may result in the use of any data that is readily available. Such misfit leads to inappropriate uses and, at worst, it can lead to poor pedagogy and student confusion. Appropriate uses of formative assessment data will require local educators to develop efficacy toward assessments. We posit that this is a function of *utility* (how teachers can actually use the data in their practice) and *outcomes* (teachers can see student growth as a function of using the data in their practice). As more and more assessments bombard schools, we should not embrace a Luddite mentality, railing against all tests. Rather we should develop our capacity to discriminate among assessment types to embrace, train, and use those assessments that are "just right" for our students. The future of formative assessment systems may be bright if they are: technology-based, curricula aligned, readily accessible to parents and educators, useful to students, and giving teachers information about what students are thinking, how they are learning, and strategies they are employing. Only when educators find the assessment that is "just right" will we be able to feed the practice of teachers and improve the achievement of students.

## 8 References

Assessment Technology Incorporated. (2006). Building reliable and valid benchmark assessments. Tucson, AZ: Assessment Technology Incorporated.

Brunner, C., Fasca, C., Heinze, J., Honey, M., Light, D., Mandinach, E., et al. (2005). Linking data and learning: The Grow Network study. Journal for Students Placed at Risk, 10(3), 241-267.

Coburn, C., & Talbert, J. (2006). Conceptions of evidence use in school districts: Mapping the terrain. American Journal of Education, 112(4), 469-495.

Earl, L., & Katz, S. (2002). Leading schools in a data-rich world. In K. Liethwood & P. Hallinger (Eds.), Second international handbook of educational leadership and administration, Part Two (pp. 1003-1023). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Friedman, T. (2007). The world is flat 3.0: A brief history of the twenty-first century. New York: Farrar, Staus and Giroux.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? Phi Delta Kappan, 89(2), 140-145.

Huff, K., & Goodman, D. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), Cognitive assessment for education: Theory and applications (pp. 19-60). New York: Cambridge University Press.

Kerr, K. A., Marsh, J. A., Schuyler Ikemoto, G., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. American Journal of Education, 112(4), 496-520.

Marshall, K. (2008). Interim assessments: A user's guide. Phi Delta Kappan, 90(1), 64-68.

Militello, M., & Schweid, J. (2009). WPI PIMSE Annual Report. Washington, DC: National Science Foundation, Graduate STEM Fellows in K-12 Education (GK-12).

Militello, M., Schweid, J., & Sireci, S. (Under Review). Formative assessment systems: Evaluating fit between intended use and product characteristics. Educational Assessment.

Militello, M., Sireci, S., & Schweid, J. (2008, March). Intent, purpose, and fit: An examination of formative assessment systems in school districts. Paper presented at the American Educational Research Association, New York City, NY.

Militello, M., & Sykes, G. (2006, April). Why schools have trouble using data. Paper presented at the National Council on Measurement in Education, San Francisco, CA.

Murnane, R., Sharkey, N. S., & Boudett, K. P. (2005). Using student-assessment results to improve instruction: Lessons from a workshop. Journal of Education for Students Placed at Risk, 10(3), 269-280.

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academic Press.

Northwest Evaluation Association. (2005). Technical manual: For use with Measures of Academic Progress and achievement level tests. Lake Oswego, OR: NWEA.

Popham, W. J. (2004). Curriculum, instruction, and assessment: Amiable allies or phony friends? Teacher College Record, 106(3), 417-428.

Sharkey, N. S., & Murnane, R. (2006). Tough choices in designing a formative assessment system. American Journal of Education, 112(4), 572-588.

Streifer, P. A., & Shumann, J. S. (2005). Using data mining to identify actionable information: Breaking ground in data-driven decision making. Journal of Education for Students Placed at Risk, 10(3), 281-293.

Wayman, J., & Stringfield, S. (2006). Data use for school improvement: School practices and research perspectives. American Journal of Education, 112(4), 463-468.

Wiliam, D. (2006). Formative Assessment: Getting the focus right. Educational Assessment, 11(3&4), 283-289.

Wylie, E. C., & Ciofalo, J. (2008). Supporting teachers' use of individual diagnostic items [Electronic Version]. Teachers College Record. Retrieved October 13, 2008 from http://www.tcrecord.org/PrintContent.asp?ContentID=153