Interim Benchmark Assessments: Are We Getting Our Eggs In the Right Basket?

Judith A. Arter
Pearson Assessment Training Institute

Overview

The following are the questions posed to presenters at this symposium, reordered and combined into the sequence in which I'll address each.

1. Do interim benchmark assessments produce the intended positive results of improved teaching and learning?
2a. Is the formative/summative distinction the most useful way to think about assessment at the district level?
2b. How do we best deal with the multiple purposes of these assessments? Can the same instruments serve both formative and summative purposes, or are these purposes mutually exclusive?
2c. What is the most useful way to think about the alignment and empirical relationship among benchmark assessments, standards, and state assessments?
3a. Is instructional utility more important than technical quality?
4. How do we evaluate the quality, relevance, and utility of these assessments for this context?

I'm going to discuss these questions from the perspective of a person who has spent many years trying to translate current best thinking about formative assessment into useful professional development materials and experiences for educators. My focus has been on improving the quality of the classroom level of assessment. Teachers and administrators want to do what's in the best interests of students; sometimes they just don't know what is best to do, especially when recommendations from experts differ. This is true of interim assessments.

I'm going to use the term *interim, benchmark* assessment to mean assessments or test-lets given at the same time to groups of students across teachers. They are standardized in content, timing, and test-taker. When they are developed by teams of educators, especially teachers, they are also called *common assessments* (Young, 2009; Ainsworth, 2007; Dufour, 2005; Reeves, 2005; Smoker, 2004).

Question 1: Do interim benchmark assessments produce the intended positive results of improved teaching and learning?

Evidence of Impact

Because we and others (e.g., Shepard, 2008) have had trouble locating evidence about the impact of interim assessments on student learning, my colleague, Rick Stiggins, recently asked Rick DuFour, a major proponent of common assessments developed and used in the context of *professional learning communities* (for example, DuFour, 2005), to provide references to the studies he cites. DuFour (personal communication, October 31, 2009) responded with three: Gallimore et al (2009), Odden and Archibald (2009), and Pashler, et al (2007). The Gallimore paper reports on the same research as that described in Saunders, et al. (2009).

The Pashler et al. report makes recommendations on the best "strategies for organizing both instruction and students' studying of material to facilitate learning and remembering information, and to enable students to use what they have learned in new situations" (p. 1) based on research on learning and memory. Of their seven recommendations, DuFour, in his personal communication, cites recommendations 1 and 5 and the connection between them to support interim assessments: It's useful to review key elements of course content after a delay of several weeks to several months after initial presentation (Recommendation 1), and quizzes, which require active retrieval of information, facilitate long-term memory (Recommendation 5). Therefore, the re-exposure to content can profitably be done by quizzes (p. 21).

The Pashler et al. report doesn't really offer evidence of the impact of interim assessments as we're discussing them in this symposium. However, I think the other two sources cited by DuFour do, indeed, provide such evidence provided interim assessments are implemented well.

Odden and Archibald conducted case studies of about 10 districts and summarized the common characteristics that enabled them to drastically improve student learning as measured by state tests. They describe similarities across their cases:

- Analyzing available data to see on which learning objectives students did well and where they need more help
- Determining where learning objectives were taught (or not taught) and refining curriculum and instruction accordingly
- Implementing formative assessments
- Conducting intensive professional development (through collaborative learning teams) on how to analyze assessment data and use it to plan instruction
- Restructuring to use time more efficiently, extending learning time for struggling students, and nurturing a collaborative culture
- Actively seeking out research evidence about how to improve schools

In his personal communication, DuFour cited the following quote in Odden and Archibald as supporting the impact of common, interim assessments:

> "Contrary to the widespread complaints in many education circles that there is too much testing in America's schools, the places we studied that doubled performance actually added another layer of testing—formative assessment. Formative assessments are instruments designed to provide detailed and concrete information on what students know and do not know with respect to discrete curriculum units. When teachers have this information . . . they know the goals and objectives they want students to learn, they know exactly what their students do and do not know with respect to those goals and objectives, so they craft instructional activities specifically to help the students in their classrooms learn the goals and objectives for the particular curriculum unit" (pp. 67-68).

Odden and Archibald note that the districts used "many sources and types of formative assessments," (p. 68), some of which were quarterly and others of which could be used "for shorter segments of instruction" (p. 68), so it was a little unclear if any particular assessment practices were more effective than others. (This same combination of interim and more frequent classroom assessment was seen in a couple of other studies—Larson and Kelleher, 2009, and Gonsalves et al., 2009—making it difficult to determine to what to attribute impact.) Also, the Odden and Archibald book cites no effect sizes; the results are anecdotal.

Gallimore et al. (2009) and Saunders et al. (2009) report statistically significant differences in student achievement between comparison schools and schools having grade-level teams that did the following (p. 1013):

1. Transform academic standards into explicit instructional goals
2. Identify assessment and indicators to assess the goals
3. Regularly evaluate school-wide achievement and determine next steps for instruction
4. Identify common instructional challenges
5. Organize professional development to address these challenges
6. Have regular (at least bi-monthly) meetings that focus on addressing identified student academic needs.

A significant part of their report discusses the support that teams needed to actually be able to accomplish these things. One important element of this support was the development of protocols for "analyzing standardized and periodic assessments, unit and instructional planning, and focusing on and addressing common student needs" (p. 1016). Other keys to effective teacher teams were giving teams enough time to focus collaboration on student learning and providing adequate support to those facilitating the collaborative teams.

Discussion of Impact Studies

So, if done well—assessing the right content, timing assessments to match instruction, taking appropriate action on the results, and persevering until success is achieved—there is some evidence that interim assessments can differentially impact student learning. These attributes of effective formative interim assessment instruments and processes are more or less those recommended by advocates of collaboratively designed interim assessments (for example, Young, 2009; Ainsworth, 2007; Ainsworth and Viegut, 2006; DuFour, 2005; Reeves, 2005; Schmoker, 2004) and have the flavor of Shepard's (2008) criteria for effective interim assessments.

Now, caveats. First, characteristics of effective formative interim assessment—both design and use—don't reflect much of current practice (Shepard 2008). Simply creating (or purchasing) and administering interim assessments doesn't increase student achievement (Henderson et al., 2007; 2008). You can't buy productive formative assessment.

Second, the amount of attention being put on having interim assessments in place saps resources from other formative practices supported by a much larger research base. Interim assessments were originally marketed because of the large research base showing that formative assessment has major impacts on student learning and motivation (e.g., Zimmerman, 2008; Morgon, 2008; Hattie and Timperley, 2007; Dweck, 2007, 2006, 2001; Costa and Kallick, 2004; Brookhart and Darkin, 2003; Assessment Reform Group, 2002; Crooks, 2001; Brookhart, 2001; Black and Wiliam, 1998).

My colleague Jan Chappuis likes to say that the following statement from Black & Wiliam (1998, p. 140) is the quote that launched a thousand products

> "Innovations that include strengthening the practice of formative assessment produce significant and often substantial learning gains."

But, impact happens not just because something is called "formative." Assessment activities and results work to improve student learning only if they include the attributes that caused these gains to begin with, such as the following (Arter, 2009; Hattie and Timperley, 2007; Militello, 2005; Herman et al., 2004; Popham, 2003; Black & Wiliam, 1998):

- On-going, continuous classroom assessment, frequent enough to guide on-going learning
- Assessing the building blocks to conceptual understanding and skillful performance; not always the final understanding or performance itself (for example there are prerequisites to being able to write for various audiences and purposes; these need to be separately practiced and assessed with feedback)
- Assessing the learning objectives currently being worked on by students in enough depth that is it possible to detect where a student is doing well and what a student needs to focus on next
- Providing accurate enough information that decisions and uses are productive
- Giving feedback to students that is descriptive—describing what the student has done well and advice on what to do to improve—and that avoids summative judgment ("you got an A," "you're smart;" "your performance is well below mastery"), or comparisons with other students
- Building in meaningful student self/peer-assessment and opportunities for students to communicate about their developing understanding during the learning

Notice the central importance of students as data-based decision-makers in creating the impact on learning. Students make life-altering decisions based on assessment results: what to study, whether to study, and whether it's best to try (and risk failing) or not try at all (and therefore have an excuse for failing that doesn't involve attribution to the worth of the self). Since students are important decision makers, we need to think about what information, delivered when, would be most useful to maximize the chances that students make productive decisions (I know what to do next and I choose to keep trying) rather than unproductive decisions (I don't know what to do next, I'll never get it, I give up). This is especially true of struggling students.

One might even go so far as to claim that students are the most important decision-makers—more important than teachers, administrators, or legislators. The decisions students make determine whether they will continue to try. If students give up, there is no way for others in the educational enterprise to succeed. Therefore, student involvement in assessment is not just something that is done with the "right" students or if there is time. Teaching students how to be productive users of assessment information, and providing them with the information they need to do so, is at the heart of effective formative assessment (Black & Wiliam, 1998, p. 143). Student-involvement in assessment appears nowhere in most interim assessment activities.

Because of these considerations, even if interim assessments prove to add value beyond classroom formative assessment (which they very well might), and even if they are designed and used in the best possible fashion (including formats other than multiple-choice and opportunities for involving students), they still *cannot take the place of* day-to-day classroom formative assessment because (a) they are simply not frequent or flexible enough to meet all the information needs of our most important decision-makers (students and teachers) and (b) they draw attention away from classroom assessment. Currently, there is stronger evidence supporting the large impact of classroom-level formative assessment practices than supporting the use of interim assessments, so, if we're going to use scarce resources wisely, we should focus on what the preponderance of evidence indicates is the best use of assessment in the service of student learning.

This doesn't mean we shouldn't try out interim assessments to determine if they add value, it means that we need to (a) devote more attention to the day-to-day classroom level of assessment, and (b) make sure that our constituencies understand that having interim assessments in place, even if well-designed and used productively, doesn't take care of the information needs of all important decision-makers.

Question 2a: Is the formative/summative distinction the most useful way to think about assessment at the district level?

Question 2b: How do we best deal with the multiple purposes of these assessments? Can the same instruments serve both formative and summative purposes, or are these purposes mutually exclusive?

Question 2c: What is the mot useful way to think about the alignment and empirical relationship between benchmark assessments, standards, and state assessments?

Many people have attempted to define *formative*.(e.g., CCSSO, 2009; Harlen & James, 1997, p. 369; Popham, 2008, p. 6; Sadler, 1998, p. 77; Shepard, 2008, p. 281). The common idea in all these definitions is the use of student assessment results to adjust teaching and learning (Chappuis, 2009, pp. 4-5). This is contrasted with *summative* assessment, the purpose of which is to document, or sum up, at a point in time how much learning has occurred. Teachers find this distinction useful because it provides a cognitive structure for categorizing and thinking about classroom assessment activities. Being clear on purpose—who will use the results of an assessment and how they will be used—is important at all assessment levels because purpose affects the design of assessment instruments and processes (as seen above).

Assessment results from the same instrument can sometimes profitably be used both formatively and summatively. For example, results from a high-stakes assessment developed for summative purposes should be squeezed dry for useful formative information. Using summative information formatively also occurs in the classroom, for example when a quiz or midterm exam is used by students to identify what they already know and what they need to work on before the next or final summative assessment, or when the ticket to a test re-take involves students' analyzing their strengths and weaknesses and acting to improve weaknesses.

Likewise, it's sometimes justifiable to use formative assessment results summatively. For example, say that a teacher uses a writing rubric to help students improve their writing. There is lots of formative assessment—teachers and students offering descriptive feedback, revising, further feedback, etc. until the writing is finished. These final products, especially if toward the end of a grading period, might be then used to assign a grade in writing or support a decision of level of student proficiency in writing.

However, just because sometimes it's possible and useful to use the results from the same assessment both formatively and summatively, doesn't mean one should. The relationship between formative and summative uses of information needs to be carefully planned because the design of the assessment instrument and process affects the type of formative or summative decisions that can be made, who benefits from the decisions, and subsequent impact on student motivation and learning.

Therefore, although it's useful to maintain the summative/formative distinction, what's really important is to be crystal clear on who the important decision-makers are, what decisions they need to make (some of which are formative and some of which are summative), what type of information would be of most use to them to make those decisions, and when that information needs to be delivered to enable the decision to be made.

I've found Table 1 (adapted with permission from Chappuis et al., 2010, pp. 14-15) useful in thinking through these considerations and how different levels of assessment productively align. Notice that Table 1 adds the classroom level of assessment to the two levels—interim and state level assessments—proposed in symposium question 3c, because it's impossible to meet the information needs of *all* important decision-makers without it.

Table 1. Framework for a Balanced Assessment System

Level of Assessment/Key Issues	Formative Applications	Summative Applications
Classroom assessment		
o *Key decision(s) to be informed?*	What comes next in the student's learning?	What standards has each student mastered or what grade does each student receive?
o *Who is the decision maker?*	Students and teachers	Teacher
o *What information do they need?*	Evidence of: where each student is now on the learning continuum toward each standard; building blocks necessary to master standards	Evidence of each student's level of mastery of each relevant standard
o *What are the essential assessment conditions?*	• Appropriate standards in learning progressions • Frequent information—minute-to-minute or day-to-day • Accurate assessment results • Results leading to next steps • Results as feedback	• Clear and appropriate standards • Accurate evidence • Focus on achievement only • Evidence well summarized • Grading symbols that carry clear and consistent meaning for all
Interim/benchmark assessment		
o *Key decision to be informed?*	Where are students struggling? Where can we improve instructional programs right away?	Did the program of instruction deliver as promised; should we continue to use it?
o *Who is the decision maker?*	Professional learning communities, district and building instructional leaders, and, maybe, students	Instructional leaders
o *What information do they need?*	Standards (and, perhaps, building blocks) students are to master	Evidence of student mastery of particular program standards
o *What are the essential conditions?*	• Clear and appropriate standards • Accurate assessment results • Results revealing how *each* student did in mastering *each* standard	Accurate assessments focused on particular program standards aggregated over students
Annual accountability testing		
o *Key decision(s) to be informed?*	What standards are our standards not mastering? Where and how can we improve instruction next year?	Are enough students meeting standards?
o *Who is the decision maker?*	School leaders, curriculum and instructional leaders	School and community leaders
o *What information do they need?*	Standards students are struggling to master	Percent of students meeting *each* standard
o *What are the essential assessment conditions?*	Accurate evidence of how *each* student did in mastering *each* standard, aggregated over students	Accurate evidence of how *each* student did in mastering *each* standard, aggregated over students

Question 3: Is instructional utility more important than technical quality?

The answer to this question depends partly on how *utility* and *technical quality* are defined. Utility, thought of as "fulfilling its purpose," is part of technical quality. By definition, if an assessment doesn't satisfy its purpose it lacks technical quality. In the context of this symposium, I'm taking technical quality to mean *quality of assessment items/tasks/rubrics* and *accuracy of results* and utility to mean *usefulness of results*.

I think it's a mistake to pose utility and accuracy as opposites. Instructional utility depends on accuracy. Teachers and others cannot usefully plan instruction and students cannot productively use assessment information to plan their own next steps in learning if information is not reasonably accurate. To say nothing of the impact on learners of making decisions based on inaccurate information. Additionally, assessment materials and processes (such as rubrics and multiple-choice items) can't be used as instructional aids unless they accurately represent the learning to be accomplished.

(While usefulness depends on accuracy, accuracy doesn't automatically result in usefulness. The most accurate information in the world is not useful if it is presented to students in a manner that causes them to conclude that trying to learn is not worth the pay-off, or if the items on the assessment don't match what was taught.)

Perhaps the issue raised in this question comes from the problem that there aren't enough professional test developers in the world to develop all the assessments needed to satisfy all the formative information needs of all users, and that, therefore, we have to rely on non-measurement experts (for example, teachers) to develop assessments. Since teacher-developed assessments are notoriously of poor quality, there must be a trade-off between accuracy and utility.

In the past one of the solutions to the low quality of teacher-made assessments was to, as much as possible, take assessment out of the hands of teachers. I think one of the major impetuses for centrally developed interim assessments is to provide teachers with at least *some* accurate interim information about student learning on valuable learning objectives.

This can no longer be our solution just because there *aren't* enough professional test developers in the world to provide all the assessments teachers and students need. Our only solution is to make teachers better assessors. But, not like we've done in the past by telling them what they need to know or trying to turn them into apprentice psychometricians. Rather, we have to be ready to demonstrate how better quality assessments (and more skill in using assessment processes and results) will make their lives easier and better (part of which is demonstrating the pay-off in terms of student learning and motivation). In my experience, teachers are more than willing to expend large amounts of effort to improve their practice if they have a clear view of what needs to be done and they can see how their effort will pay off.

Fortunately, there are a couple of things that will help us help teachers be better assessors. First, teachers probably don't need to understand issues of reliability and validity in the same way that test developers understand them. Because the assessments teachers develop are relatively low-stakes, having an understandable cognitive structure for thinking about issues of quality and rules of thumb to implement them might be enough. Along this line, it behooves us to consider the three (or four or five) things that teachers could do that would most improve the quality of the assessments they use—both interim and day-to-day. Then we need to help them not only do these things well, but *want* to do these things well.

For example, I think that assessments developed by teachers would make a quantum leap in quality if teachers would simply make sure that their assessments cover what they taught and that teaching and assessment match the intended learning objectives. Rather than lecturing about the importance of test specifications, one of the most useful things we've done is this:

1. Ask teachers to discuss the question: "What happens when our assessments don't match what was taught?" They always agree that assessments should match what was taught because if not teachers might reteach content that doesn't need to be retaught or not reteach something that needs to be, students are frustrated or bored, and students might come to the conclusion that they're not getting it and give up in hopelessness. Having teachers come up with these conclusions for themselves builds felt need.

2. Show teachers the steps they could take to make sure this doesn't happen (see Attachment 1). The process requires that teachers analyze a test item by item to see what the test covers and then think about if that's what was intended. We illustrate this using a concrete example, like the one in Attachment 2. Teachers feel this is very useful but always bring up the issue of time. So we suggest they initially focus on the one assessment they are most dissatisfied with. (Which is frequently, in our experience, a common interim assessment.)

3. Increase teacher desire to do the required analysis by showing a nifty student self-assessment and goal-setting idea that is possible once they know what each item on a test measures (see Attachment 3). Teachers love this idea because it's a do-able, concrete example of what meaningful student self-assessment and goal setting looks like. (Notice that this set of activities relate to learning objectives that can be assessed using selected response or short answer assessment methods. We show teachers other techniques to use if learning objectives must be assessed using a rubric.)

I realize that the procedure described above only works well to improve student learning if items match important learning outcomes, test items are well written, and the test samples appropriately. But, the process of analyzing a test for match to instruction always brings out these points, as well as others such as "mastery of what" and "what happens when standards-based reporting gets ahead of standards-based assessment and record-keeping." Such contextualized learning is much more effective than merely telling teachers they need to develop test plans, sample well, and write good quality test items.

I know I've drifted some from interim assessment to classroom assessment. But, the point is that teachers need to be better assessors to do either, and if they are, maybe we don't have to fret as much about trading off accuracy against utility.

Question 4: How do we evaluate the quality, relevance, and utility of these assessments for this context?

The value of any assessment depends on the extent to which the assessment reflects the learning objectives it is supposed to measure, serves the purpose for which it was intended, provides dependable information, and results in a positive impact on student motivation and learning.

Interim assessments, to the extent they are intended to be formative, need to incorporate those features research has shown to maximize formative usability, as described elsewhere in this paper: timing, coverage, and quality of items, tasks, and rubrics. If intended to be summative, interim assessments might need to have other characteristics such as ability to predict performance on year-end assessments.

In either case, interim assessments need to be able to demonstrate they add value to assessment systems already in place or beyond that possible by increased focus on high quality classroom assessment.

Conclusion

There is some evidence that, if done well, interim assessments can result in increased student learning. It also makes logical sense that three levels of assessment—classroom, interim, and year-end—can lead to a productive/balanced system, one that gives *all* users of assessment information and processes (including teachers and students) the information they need to make productive formative and summative decisions.

It's still unclear, however, the extent to which interim assessment can produce more student learning than if the same resources were instead used to help teachers become better classroom assessors. In other words, the value-added question still needs to be answered so that we can decide on the best balance between classroom, interim, and end-of-year assessment.

In any case, there are some immediate needs. Most importantly, we need to make sure that constituents understand those attributes of assessment instruments and uses of results that originally lead to the conclusion that formative assessment can create large gains in student achievement and motivation. Interim assessments, with results used only by teachers, represent, at best, a small part of the active ingredients.

More broadly, we need to help constituents understand the various formative and summative decisions made at all levels by all decision-makers so that they can design assessment systems that meet everyone's needs. This is the meaning of balance.

References

Ainsworth, L. (2007). Common formative assessments: The centerpiece of an integrated standards-based assessment system. In D. Reeves (Ed.), Ahead of the curve: The power of assessment to transform teaching and learning. Bloomington, IN: Solution Tree.

Ainsworth, L.& D. Viegut (2006). Common formative assessments: How to connect standards-based instruction and assessment. Thousand Oaks, CA: Corwin Press.

Arter, J.A. (2009). Setting the context for: Examination of formative uses of day-do-day classroom and common assessments. Paper presented at the annual meeting of the American Educational Research Association, 2009, San Diego. Available from <judy.arter@pearson.com>.

Assessment Reform Group (2002). *Testing, motivation and learning.* Cambridge, UK: University of Cambridge, Faculty of Education.

Black, P. & D. Wiliam (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80(2)*: 139-148.

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education, 8*(2), 153–169.

Brookhart, S. M., & D. Durkin (2003). Classroom assessment, student motivation, and achievement in high school social studies classes. *Applied Measurement in Education, 16*(1), 27–54.

Chappuis, J. (2009). *Seven strategies of assessment* for *learning.* Portland, OR: Assessment Training Institute.

Chappuis, S., C. Commodore, & R. Stiggins (2010). Assessment balance and quality: An action guide for school leaders. Portland, OR: Assessment Training Institute.

Costa, A. L., & B. Kallick (2004). *Assessment strategies for self-directed learning.* Thousand Oaks, CA: Corwin.

Council of Chief State School Officers (2009). Mission and history of the formative assessment for students and teachers SCASS. Available from <www.CCSSO.org> (downloaded March 9, 2009).

Crooks, T. 2001. The validity of formative assessments. Paper presented at the 2001 Annual Meeting of the British Educational Research Association, Leeds, UK, September 13–15.

DuFour, R. (2005). What is a professional learning community? In R. DuFour, R. Eaker, R. DuFour (Eds.) On common ground: The power of professional learning communities. Bloomington, IN: National Educational Service, pp. 31-42.

Dweck, C. S. (2001). *Self-theories: Their role in motivation, personality, and development.* Philadelphia: Psychology Press.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York: Random House.

Dweck, C. S. (2007). The secret to raising smart kids. *Scientific American Mind*, November 28, 2007. Retrieved November 12, 2008 from http://www.sciam.com/article.cfm?id=the-secret-to-raising-smart-kids&print=true

Gallimore, R., B.A. Ermeling, W.M. Saunders, & C. Goldenberg (2009). Moving the learning of teaching closer to practice: Teacher education implications of school-based inquiry teams. Elementary School Journal 109(5): 537-553.

Gonsalves, P., D. Kravin, & W. Conrad (2009). Paper presented at the annual meeting of the American Educational Research Association, 2009, San Diego. Available from <pgonsavles@acoe.org>.

Harlen, W., & M. James (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Educational Assessment: Principles, Policy and Practice, 4*(3): 365-379.

Hattie, J., & H. Timperley (2007). The power of feedback. *Review of Educational Research*. Retrieved October 9, 2007 from http://rer.sagepub.com

Herman, J.L., E.L. Baker, & R.L. Linn (2004). Accountability systems in support of student learning: Moving to the next generation. *CRESST LINE*, Spring 2004, pp. 1-7.

Henderson, S., A. Petrosino, S. Guckenburg, & S. Hamilton (December 2007.) Measuring how benchmark assessments affect student achievement, Report REL 2007-No. 39. Washington DC: U.S. Department of Education, Institute of Education Sciences,

National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. <http://ies.ed.gov/ncee/edlabs> (downloaded March 3, 2010).

Henderson, S., A. Petrosino, S. Guckenburg, & S. Hamilton (April 2008). A second follow-up year for measuring how benchmark assessments affect student achievemen, Report REL 2008-No. 002. Washington DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. <http://ies.ed.gov/ncee/edlabs> (downloaded March 3, 2010).

Larson, A.F. and M.E. Kelleher (2009). Error identification assessments: Diagnostic instruments *for* instruction. Paper presented at the annual meeting of the American Educational Research Association, 2009, San Diego.

Morgan, A.(2008). *Feedback: Assessment for rather than of learning.* Retrieved September 3, 2008 from http://www.bangor.ac.uk/adu/the_scheme/documents/FEEDBACKJanuary06_000.ppt

Militello, M. (2005). Two much information: A case study of assessment accountability in an urban school district. Paper presented at the annual meeting of AERA, Montreal, Canada.

Odden, A.R. & S.J. Archibald (2007). Doubling student performance . . . and finding the resources to do it. Thousand Oaks, CA: Corwin Press.

Pashler, H., PM. Bain, B.A. Bottge; A. Graesser; K. Koedinger, M. McDaniel, and J. Metcalfe (2007). Organizing instruction and study to improve student learning: IES practice guide. National Center for Education Research, Institute of Education Sciences, US Department of Education. <Available at http://ncer.edu.gov>.

Popham, W. J. (2008). Transformative assessment. Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J. (2003). *Are your state's NCLB tests instructionally insensitive? Here's how to tell!* Paper prepared for the National School Boards Association, February 2003. Similar points are made in: *Building tests that support instruction and accountability: A guide for policymakers* (<www.aasa.org>) & *Crafting curricular aims for instructionally supportive assessment (*<http://education.umn.edu/NCEO/Presentations/CraftingCurricula.pdf>).

Reeves, D. (2005). Putting it all together: Standards, assessment, and accountability in successful professional learning communities. In R. DuFour, R. Eaker, R. DuFour (Eds.) On common ground: The power of professional learning communities. Bloomington, IN: National Educational Service, pp. 45-63.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education, 5*(1): 77-84.

Saunders, W.M., C.N. Goldenberg, & R. Gallimore (2009). Increasing achievement by focusing grade-level teams on improving classroom learning: A prospective, quasi-experimental study of Title I schools. *American Educational Research Journal, 46(4)*: 1006-1033.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Shepard, L. A. (2008). Formative assessment: Caveat emptor. In C.A. Dwyer (ed), *The future of assessment: Shcaping teaching and learning*. New York: Lawrence Erlbaum, pp. 279-303.

Schmoker, M. (2004). Start here for improving teaching. *School Administrator*. November, 2004, <www.aasa.org/School Administrator Article> (downloaded March 8, 2010).

Stiggins, R., J. Arter, J. Chappuis, & S. Chappuis (2006). *Classroom assessment for student learning: Doing it right—using it we*ll. Portland, OR: Assessment Training Institute.

Young, A. (2009). Using common assessments in uncommon courses. In T.R. Guskey (Ed.) *The teacher as assessment leader*, Bloomington, IN: Solution Tree Press, pp. 135-153.

Zimmerman, B. (200)8. Investigating self-regulation and motivation: historical background, methodological development, and future prospects. *American Educational Research Association Journal*, 45(1), 166–183.

Attachment 1: Steps to Make Sure Assessment and Instruction Align

1. *Analyze your test item by item.* Identify and write down what learning each item assesses. Describe the learning in whatever terms you want. If two or more items address the same learning, use the same terms to describe that learning.

2. *Organize the learning targets into a test plan*. Transfer the information from Step One to this chart:

Learning Target	Item #s	Points

3. *Question your test plan: Is this a representative sample of what you taught and what you expected students to learn? How does it relate to standards?*

   - Does the number of points for each learning target represent its relative importance within the whole? If not, which ones are out of balance? Are some learning targets overrepresented? If so, which one(s)? Are some learning targets underrepresented? If so, which one(s)?
   - Does the number of points for each learning target represent the amount of time you spent on it relative to the whole? If not, which ones are out of balance?
   - Are some of the important learning targets you taught left out? If so, which one(s)?
   - Do all items on your test align directly with the content standards you have taught?

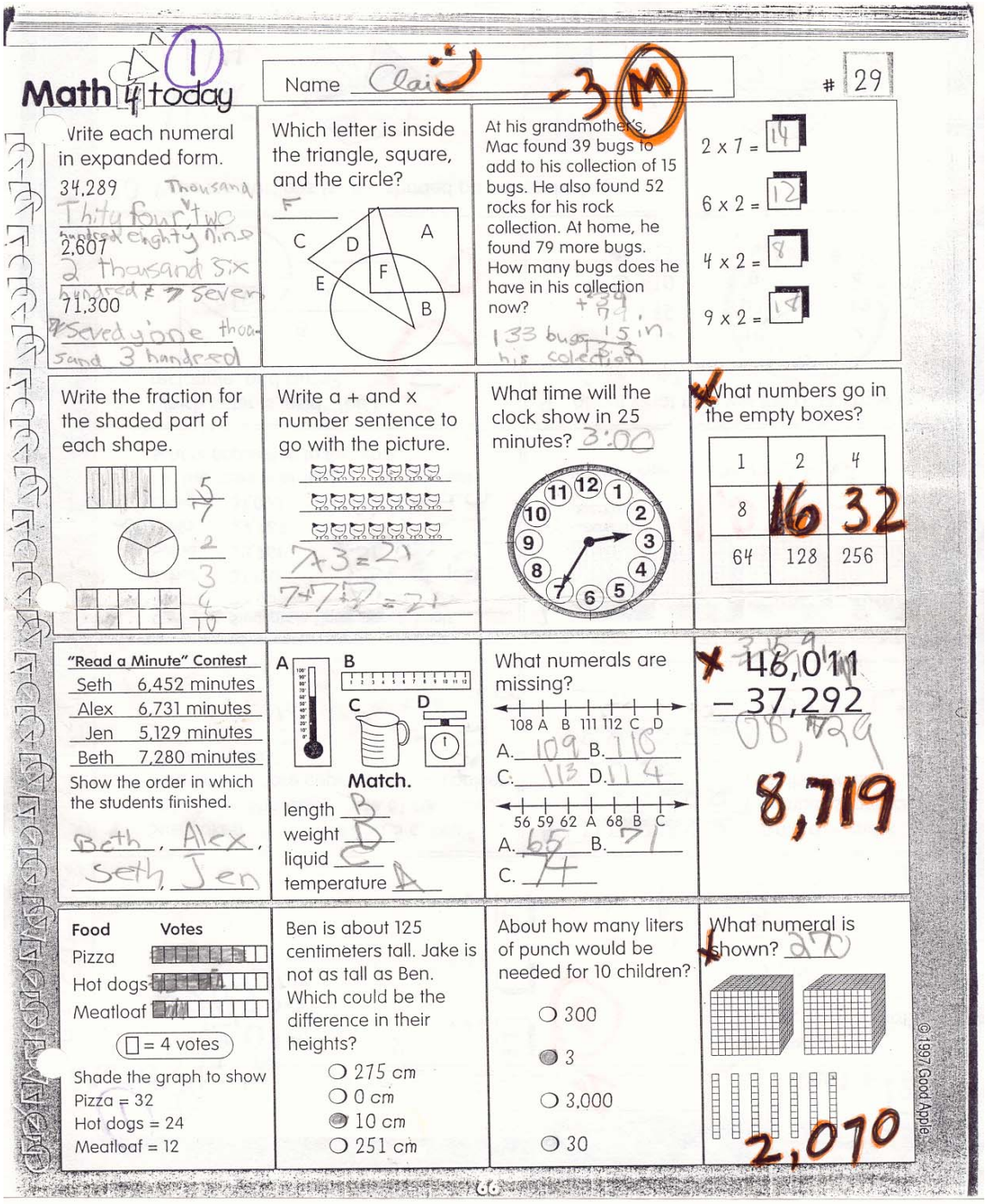
4. *Adjust your test plan.* As needed, adjust the numbers in the "# of points" column on the previous page to reflect the amount of time you spent teaching each learning target and each target's relative importance to the content as a whole. As needed, add or delete learning targets to reflect what you taught and what you deemed most important to learn and assess.

5. *Draw conclusions about your assessment.* What does this tell you about the matches among what's written in your curriculum, what you taught, and what you assessed?

From Stiggins, et al. (2006), pages 108 – 109.

Attachment 2: Example

ANALYZING ASSESSMENTS FOR CLEAR TARGETS



Math 4 today

Name Clai

29

Write each numeral in expanded form.

34,289 Thousand Thirty four, two hundred eighty nine

2,607 2 thousand Six hundred Seven

71,300 Seventy one thousand 3 hundred

Which letter is inside the triangle, square, and the circle?

F

C D A F E B

At his grandmother's, Mac found 39 bugs to add to his collection of 15 bugs. He also found 52 rocks for his rock collection. At home, he found 79 more bugs. How many bugs does he have in his collection now?

133 bugs in his collection

2 x 7 = 14

6 x 2 = 12

4 x 2 = 8

9 x 2 = 14

Write the fraction for the shaded part of each shape.

5/7 2/3 4/10

Write a + and x number sentence to go with the picture.

7 x 3 = 21

7 + 7 + 7 = 21

What time will the clock show in 25 minutes? 3:00

What numbers go in the empty boxes?

1	2	4
8	16	32
64	128	256

"Read a Minute" Contest

Seth	6,452 minutes
Alex	6,731 minutes
Jen	5,129 minutes
Beth	7,280 minutes

Show the order in which the students finished.

Beth, Alex, Seth, Jen

A B C D

Match.

length B

weight D

liquid C

temperature A

What numerals are missing?

108 A B 111 112 C D

A. 109 B. 110

C. 113 D. 114

56 59 62 A 68 B C

A. 65 B. 71

C. 74

46,011 − 37,292 = 8,719

Food	Votes
Pizza	
Hot dogs	
Meatloaf	

□ = 4 votes

Shade the graph to show

Pizza = 32

Hot dogs = 24

Meatloaf = 12

Ben is about 125 centimeters tall. Jake is not as tall as Ben. Which could be the difference in their heights?

- ○ 275 cm
- ○ 0 cm
- ● 10 cm
- ○ 251 cm

About how many liters of punch would be needed for 10 children?

- ○ 300
- ● 3
- ○ 3,000
- ○ 30

What numeral is shown? 2,070

Attachment 2 (cont.)

ANALYZE YOUR OWN ASSESSMENT FOR CLEAR TARGETS

1. Analyze your test item by item.

Identify and write down what learning each item assesses. Describe the learning in whatever terms you want. If two or more items address the same learning, use the same terms to describe that learning. For "Claire's Math," you would begin like this:

NCTM Standards Represented on "Claire's Math"

1. NUMBER SENSE (3)*	2. REPRESENTATION (1)	3. PROBLEM SOLVING / NUMBER OPERATIONS (1)	4. NUMBER OPERATIONS (4)
5. NUMBER SENSE (3)	6. ALGEBRA (2)	7. MEASUREMENT (1)	8. ALGEBRA (2)
9. NUMBER SENSE (1)	10. MEASUREMENT (4)	11. NUMBER SENSE (7)	12. NUMBER OPERATIONS (1)
13. DATA ANALYSIS & PROBABILITY (3)	14. MEASUREMENT (1)	15. MEASUREMENT (1)	16. NUMBER SENSE (1)

* Numbers in parentheses indicate the number of answers called for in each box, which adds up to the total possible score on this assignment.

2. Organize the learning targets into a test plan.

Transfer the information from Step One to this chart. (Example is from "Claire's Math.")

Learning Target	Item #s	Points
Number Sense: Place value	1, 5, 9, 11, 16	15
Representation	2	1
Number Operations: Fractions, multiply by 2, subtract with borrowing	4, 12	5
Problem Solving/Add with carrying	3	1
Measurement: Identify correct units	7, 10, 14, 15	7
Data Analysis & Probability: Tables, charts, and graphs	13	3
Algebra: Number patterns, number sentences	6, 8	4

From Stiggins, et al. (2006), pages 108 – 109.

3. Question your test plan: Is this a representative sample of what you taught and what you expected students to learn? How does it relate to standards?

- Does the number of points for each learning target represent its relative importance within the whole? If not, which ones are out of balance? Are some learning targets overrepresented? If so, which one(s)? Are some learning targets underrepresented? If so, which one(s)?
- Does the number of points for each learning target represent the amount of time you spent on it relative to the whole? If not, which ones are out of balance?
- Are some of the important learning targets you taught left out? If so, which one(s)?
- Do all items on your test align directly with the content standards you have taught?

4. Adjust your test plan.

As needed, adjust the numbers in the "# of points" column on the previous page to reflect the amount of time you spent teaching each learning target and each target's relative importance to the content as a whole.

As needed, add or delete learning targets to reflect what you taught and what you deemed most important to learn and assess.

5. Draw conclusions about your assessment.

What does this tell you about the matches among what's written in your curriculum, what you taught, and what you assessed?

From *CASL*, Activity 4.4, pages 108 – 109. A blank form can be found on the *CASL* CD, in the Chapter 4 file, "Analyze for Clear Targets."

Activity directions:
With a partner, discuss possible answers to the questions posed in Step 3. Then discuss your overall conclusions about Claire's math test.

From Stiggins, et al. (2006), pages 108 – 109.

Attachment 3: Student Self-Assessment and Goal Setting
"You Be George"

- Students use test plans as a basis for evaluation of strengths and areas of study.
- Students complete self-evaluation and goal-setting form on the basis of test or quiz results.

On the following pages you will find an activity students can do to help them know what the results of a test mean about what they have learned and what they still need to work on. It is a way of providing descriptive feedback to students that engages them in self-assessment and goal-setting.

Here is the process:

1. The teacher identifies the learning target each item on the test represents and fills out the first two columns on the form.

2. The teacher corrects the tests and hands them back.

4. Students mark the next two columns—right or wrong—by looking at their corrected tests.

5. Students mark the last two columns—simple mistake or further study—by reviewing the items they got wrong. To make this decision, they ask themselves, "Do I know what I did wrong? Could I correct this myself?" If the answer is "Yes," then they mark the "simple mistake" column. If the answer is "No," they mark the "I don't get it" column.

6. Students then transfer each learning target to one (or more) of three categories on the next page—strengths, further study, and review.

7. Finally, they use the form of your choice to make a plan to improve.

From Stiggins, et al. (2006), pages 158-163.

IDENTIFYING MY STRENGTHS AND AREAS FOR IMPROVEMENT

Name: George **Assignment:** Math Test #7 **Date:** December 1, 2004

Please look at your corrected test and mark whether each problem is right or wrong. Then look at the problems you got wrong and decide if you made a simple mistake. If you did, mark the "Simple mistake" column. For all the remaining problems you got wrong, mark the "Don't get it" column.

Problem	Learning Target	Right?	Wrong?	Simple mistake?	Don't get it?
1	Place Value: Write numerals in expanded form to 10 thousands place				
2	Place Value: Write numerals in expanded form to 10 thousands place				
3	Place Value: Write numerals in expanded form to 10 thousands place				
4	Place Value: Identify place value to the thousands place				
5	Place Value: Put numbers in order through the thousands				
6	Place Value: Put numbers in order through the thousands				
7	Place Value: Put numbers in order through the thousands				
8	Write fractions to match models				
9	Write fractions to match models				
10	Write fractions to match models				
11	Write fractions to match models				
12	Subtract 3-digit numbers with borrowing				
13	Subtract 3-digit numbers with borrowing				
14	Subtract 3-digit numbers with borrowing				
15	Subtract 3-digit numbers with borrowing				
16	Measurement: Read time to the nearest minute				
17	Measurement: Read a thermometer				
18	Measurement: Know how much a liter is				
19	Measurement: Know how long a centimeter is				
20	Measurement: Choose the right tool to measure length, weight, liquid, and temperature				

From Stiggins, et al. (2006), pages 158-163.

IDENTIFYING MY STRENGTHS AND AREAS FOR IMPROVEMENT

Name: George **Assignment:** Math Test #7 **Date:** December 1, 2007

Please look at your corrected test and mark whether each problem is right or wrong. Then look at the problems you got wrong and decide if you made a simple mistake. If you did, mark the "Simple mistake" column. For all the remaining problems you got wrong, mark the "Don't get it" column.

Problem	Learning Target	Right?	Wrong ?	Simple mistake?	Don't get it?
1	Place Value: Write numerals in expanded form to 10 thousands place	**x**			
2	Place Value: Write numerals in expanded form to 10 thousands place	**x**			
3	Place Value: Write numerals in expanded form to 10 thousands place	**x**			
4	Place Value: Identify place value to the thousands place	**x**			
5	Place Value: Put numbers in order through the thousands	**x**			
6	Place Value: Put numbers in order through the thousands	**x**			
7	Place Value: Put numbers in order through the thousands		**x**	**x**	
8	Write fractions to match models	**x**			
9	Write fractions to match models		**x**		**x**
10	Write fractions to match models	**x**			
11	Write fractions to match models		**x**		**x**
12	Subtract 3-digit numbers with borrowing	**x**			
13	Subtract 3-digit numbers with borrowing		**x**	**x**	
14	Subtract 3-digit numbers with borrowing	**x**			
15	Subtract 3-digit numbers with borrowing		**x**	**x**	
16	Measurement: Read time to the nearest minute		**x**	**x**	
17	Measurement: Read a thermometer	**x**			
18	Measurement: Know how much a liter is		**x**		**x**
19	Measurement: Know how long a centimeter is	**x**			
20	Measurement: Choose the right tool to measure length, weight, liquid, and temperature	**x**			

From Stiggins, et al. (2006), pages 158-163.

.

YOU BE GEORGE

George, a third-grader, filled out the form on the previous page after receiving his corrected test from his teacher. Please imagine you are George—do a little self-analysis and goal setting by completing the form on this page.

NAME: George **TEST DATE:** December 1, 2007

I AM GOOD AT THESE

Learning targets I got right:

I AM PRETTY GOOD AT THESE, BUT NEED TO DO A LITTLE REVIEW

Learning targets I got wrong because of a simple mistake: What I can do to keep this from happening again:

I NEED TO KEEP LEARNING THESE

Learning targets I got wrong and I'm not sure what to do to correct them: What I can do to get better at them:

From Stiggins, et al. (2006), pages 158-163.

You can follow this or any self analysis with a goal-setting activity.

To get better at _____, I could…

-
-
-

One thing I am going to start doing is…

-

I'll start doing this on _____ (date) and work on it until _____ (date)

One way I'll know I'm getting better is …

Goal	Steps	Evidence
What do I need to get better at?	How do I plan to do this?	What evidence will show I've achieved my goal?

Time Frame: Begin _____ End _____

Date _____ Signed _____

From Stiggins, et al. (2006), pages 369 – 371.