Exposing the imbalance in "balanced assessment"

There are three elements to 'balanced assessment', but W. James Popham argues that only two deserve their place

IT IS DIFFICULT TO ATTEND any sort of assessment-relevant educational conference these days without hearing someone extol the virtues of "balanced assessment." In the U.S., what's typically being described by the proponents of balanced assessment is the application of three distinctive measurement strategies: *classroom assessments*; *interim assessments*; and *large-scale assessments*.

Balanced assessment, as is the case with "balanced" anything, sounds so delightfully defensible. Those who oppose balanced assessment are apt to be the sorts of villains who want "low standards" instead of "high standards" and who applaud "unreliable tests" instead of tests reeking of reliability. Whatever is balanced seems, *a priori*, to be wonderful, but in this case the term may be misleading.

The 'Blessed Trinity' of balanced assessment

Briefly, the three measurement strategies of balanced assessment are:

Classroom assessments, typically teacher-made, are currently employed by most teachers for the purpose of grading their students or as motivators when urging students to "study hard for the upcoming test." Classroom assessments can also supply timely evidence whenever teachers use formative assessment.

Interim assessments are usually purchased from commercial vendors, but are sometimes created locally. These are standardized tests, typically administered by a district or a state, perhaps two or three times during the school year. Interim tests are intended to fulfill one of the following measurement missions: (1) a predictive function, such as identifying students who are at risk of failing a subsequent high-stakes test, (2) an evaluative function, such as appraising the effectiveness of a recently concluded educational program, or (3) an instructional function, such as supplying teachers with instructionally useful diagnostic data regarding their students. Occasionally, interim tests are intended to supply evidence for more than one of these functions.

Large-scale assessments are almost always created by assessment organizations, either for-profit or not-for-profit groups. In the U.S., the most common examples of these sorts of tests are the annual accountability assessments administered by all U.S. states. Although large-scale assessments are used for purposes other than accountability, for instance, as college entrance exams, the large-scale tests associated with the balanced assessment are typically achievement tests intended for use in an accountability context.

A party crasher

Two of these types of assessment are supported by strong evidence, but one is trying to crash the measurement party without the proper admission credentials.

Classroom assessments

It's not classroom assessments. Classroom assessments are supported by a formidable array of empirical evidence showing that, when used properly, they trigger substantial growth by students. When classroom assessments are used as part of formative assessment – a process wherein assessment-elicited evidence is used by teachers and/or students to make necessary

adjustments in what they are doing – there is an abundance of empirical evidence to show that the formative-assessment process is remarkably effective. In their seminal 1998 review of classroom-assessment studies, Paul Black and Dylan Wiliam concluded that formative assessment works conclusively, it produces powerful gains in students' achievement, and it is sufficiently robust so that teachers can use it in a variety of ways, yet still get glittering results. Subsequent empirical investigations continue to support the instructional payoffs of appropriately employed classroom assessments.

Large-scale assessments

Nor are large-scale assessments the party crashers. Large-scale tests, particularly those employed for accountability purposes, enjoy enormous support among both educational policy makers and the public at large. The public are increasingly demanding hard evidence that their schools are being successful, and that their taxes are being well-spent. Not placated by educators' reassurances, educational policy makers at all levels, local to national, are demanding hard, test-based evidence regarding students' achievement. Large-scale accountability tests supply such evidence, and will remain in place until an incredulous citizenry becomes convinced that our schools are working.

> **“Those who oppose balanced assessment are apt to be the sorts of villains who want “low standards” instead of “high standards” and who applaud “unreliable tests” instead of tests reeking of reliability”**

Interim assessments

However, in contrast to the other two types of assessment, interim assessments are neither supported by research evidence, nor are they regarded by the public or policy makers as being of particular merit. Indeed, most members of the public and most policy makers don't even know that interim assessments exist.

The chief advocacy for including interim assessments as one of the three strategies of balanced assessment, not surprisingly, comes from the vendors who sell them. Many district-level administrators are desperate to prevent their schools from getting low scores on annual state accountability tests, and so are swayed by the glowing words about the positive instructional payoffs that accompany commercially peddled interim tests. It is not surprising that many district officials purchase interim assessments for their teachers.

Yet, at the 2010 annual meeting of the National Council on Measurement in Education in Denver, Judith Arter – based on her careful review of research studies regarding interim tests – concluded that no meaningful empirical support currently exists for interim assessments. Regretfully, she noted that “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.”

Accordingly, when it comes to the support associated with these three assessment approaches, one of them is blatantly out of balance with its assessment cousins.

A serious shortcoming?

Interim tests, other than being seen by some armchair analysts as “rounding out” the balanced-assessment picture, come to us without compelling support, either empirical or political. In the U.S. where almost any TV-advertised health product is accompanied these days by an allusion to “clinical evidence” supporting the product's virtues, the promotional literature accompanying America's interim assessments is particularly light on evidence, of any sort, that they are worth what they cost. And their costs are not trivial, either in terms of money spent or in classroom time taken.

Perhaps, in the future, research evidence supporting the instructional dividends of interim assessments will be available. However, it's possible that there is an inherent, but unrecognized flaw in the interim-assessment approach, a flaw that dooms these tests to be ineffectual, particularly in improving instruction. Interim tests, you see, are administered at a given time during the school year, for instance, in the middle of or near the close of every three-month segment. So, in order for the results of these tests to help teachers instructionally, the timing of the teacher's instruction must mesh with what's covered in a given interim test. A test covering yet-untaught content, or content that was treated weeks ago, will hardly inform a teacher's decision-making. Accordingly, either teachers allow the curricular pacing of their instruction to be regimented by what's to be assessed on these interim tests (and few teachers relish such regimentation), or teachers will find their instruction is out of line with what's being tested. Perhaps this is why no evidence regarding the profound payoffs of interim tests has yet been seen. Perhaps, for most teachers, interim assessments just don't work.

What we know

- Classroom formative assessment works well, and the process can be successfully used by classroom teachers in diverse ways.
- Society now demands evidence from large-scale accountability tests to evaluate the success of tax-supported schooling.
- Interim assessments, at the moment, are supported neither by research evidence nor by a societal demand.

Conclusion

Nonetheless, we continue to see ardent advocacy for the installation of balanced-assessment approaches. Much of this advocacy can be traced back to the very folks who sell such tests. If balanced assessment comes to be seen as *necessarily* including interim assessments, then those who sell such assessments can be assured of a serious slice of assessment's fiscal pie. Yet, until suitable support for interim tests arrives, balanced assessment will most definitely remain out of balance.

About the author

W. James Popham is Professor Emeritus at UCLA Graduate School of Education and Information Studies. He spent the bulk of his educational career as a school teacher, and later taught courses in instructional methods for prospective teachers as well as courses in evaluation and measurement for graduate students at UCLA. He has written and co-written 30 books, 200 journal articles, 50 research reports, and 175 papers.

Further reading/resources

Arter JA (2010). *Interim Benchmark Assessments: Are We Getting Our Eggs In the Right Basket*? Paper presented at the annual meeting of the National Council on Measurement in Education, Denver www.assessmentinst.com.

Heritage M (2010), *Formative Assessment: Making It Happen in the Classroom*. Corwin Press: Thousand Oaks, CA.

Heritage M (2010), *Formative Assessment and Next-generation Assessment Systems: Are We Losing an Opportunity*? Council of Chief State School Officers: Washington, DC. www.ccsso.org/resources/publications.html?name-search=Heritage A

Popham WJ (2010), *Everything School Leaders Need to Know About Assessment*, Thousand Oaks, CA: Corwin Press.