BUILDING VALID AND USEFUL INTERIM ASSESSMENTS

Considering Purpose in Design

The Smarter Balanced Assessment Consortium has produced comprehensive summative assessments as well as a resource library of formative tools for English Language Arts and Science. Banks of items aligned to the Common Core State Standards are available to build interim assessments' to use throughout the school year. This paper describes multiple approaches to building those interim assessments based on the intended purpose and use. Blueprints, item types, length and frequency of assessing students is discussed.

Building Valid and Useful Interim Assessments

Marianne Perie, University of Kansas

A report prepared September 30, 2014 for the Smarter Balanced Assessment Consortium

A comprehensive assessment system, with cohesion among the components, is needed to successfully promote student learning of the college- and career-ready standards. Comprehensive assessments systems typically consist of three components: summative, interim, and formative assessments.

Summative assessments are administered after learning has occurred to determine the amount of knowledge and skills a student has acquired. Stiggins, Arter, Chappuis, and Chappuis (2006) consider summative assessments to be "assessments *of* learning" and contrast them to formative assessments which are "assessments *for* learning."

The Smarter Balanced Assessment Consortium (Smarter) is developing summative assessments aligned to the Common Core State Standards that are intended to measure knowledge and skills in English language arts and mathematics at the end of the year. The summative tests are comprised primarily of machine-scorable items, either multiple-choice or technology-enhanced items, some short-constructed response items, as well as performance tasks with an extended constructed-response item. Additionally, the consortium provides access to a Digital Library of formative assessment tools and practices that supports classroom instruction. The Digital Library is an online collection of professional development and instructional resources aligned to the Common Core that support educators in their use of formative assessment strategies. It is intended to complement the interim and summative assessments by providing resources that teachers can use throughout the year in their classrooms.

Interim assessment fits between summative and formative assessment and is intended to provide periodic evaluations of student knowledge and skills. Interim assessments would generally be considered "medium-cycle" assessments under the Wiliam and Leahy (2006) typology. They are administered at regular intervals, usually three to four times per year or every 6 to 10 weeks (Goertz, Olah, & Riggan, 2010). However, it is not just the cycle that defines interim assessment. According to Perie, Marion, and Gong (2009), interim assessments are administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform s policymaker's or educator's decisions at the classroom, school, or district level. The specific interim assessment designs are driven by the purpose and intended uses, but the results of any interim assessment are reported in a manner allowing aggregation across students, occasions, or concepts (Perie et al., 2009). Aggregating the data is a key distinction between interim and formative assessments. Table 1 summarizes the differences among the three assessment types.

Table 1. ***Three Assessment Types***

	Formative	Interim	Summative
Typical Use	feedback to adjust ongoing teaching and learning	monitoring student progress	student placement; school and district accountability
Frequency of Administration	continual; multiple times a day	generally two to six times per school year	usually once a school year
Scope of Administration	student and classroom	usually school or district	usually state

Note. Sourced from *Interim assessment practices and avenues for state involvement* by the TILSA SCASS Interim Assessment Subcommittee, 2008, p. 4.

Although data from all three kinds of assessment can be used to improve teaching and learning, Marshall (2006) argues that interim assessment presents the most powerful entry point for principals to work to improve instruction and boost student achievement.

Interim Assessment Purposes and Uses

Perie et al. (2009) emphasized that the purpose of the interim assessment drives its design. They defined three purposes: instructional, predictive, and evaluative. Since 2009, there has been less interest in using interim assessment for evaluating education programs. Instead, the increased emphasis on using student growth to evaluate teachers has led to the need to assess within-year growth of students. Interim assessments that can quantify the increase in knowledge and skills throughout the year can be used to augment teacher evaluation systems. Thus, the three purposes and uses discussed here include instruction, prediction, and growth.

Instructional Uses

One purpose of an interim assessment is to monitor student progress, identify skill gaps in individual students or across groups of students, and either modify instruction accordingly or provide additional support, such as after-school tutoring, to targeted students. The overarching goal of interim assessments used for instructional purposes is to help students meet their learning goals. A typical system contains a bank of items aligned with the state content standards that teachers can use to create a test to evaluate student learning on the concepts taught to date. Results are reported immediately, and data are disaggregated by content target or strand, allowing teachers to identify strengths and weaknesses in the students' learning.

Stecher et al. (2008) found that 53% to 73% of study districts in California, Georgia, and Pennsylvania used interim tests in elementary and middle school mathematics in 2006. About three quarters of the teachers indicated that the results of these progress tests helped them identify and correct gaps in curriculum and instruction. Similarly, 86% of teachers in Providence, Rhode Island, reported that they modified instruction because of interim assessment tests (Clune & White, 2008). Many of these teachers reported that interim assessment results helped them monitor student progress and identify skill gaps for their students and led them to modify curriculum and instruction (Christman et al., 2009; Clune & White, 2008).

To serve instructional purposes, an assessment system must go beyond simply providing data. Educators must have strategies for interpreting and using the data to effectively modify classroom instruction. It is worth noting a tension between the need for professional development to accompany these assessment systems and the ownership of that responsibility. That is, should whoever designed the interim assessments also be responsible for developing associated professional development resources, or does that responsibility remain at the local level with district or perhaps even with school leadership?

Predictive Uses

Predictive assessments are designed to determine each student's likelihood of meeting some criterion score on the end-of-year tests. In some districts, predictive tests have been used solely to identify students who are not on track to succeed on the end-of-year assessment. Then, once those students are identified, they are given further probes to determine areas of weakness and provided with remedial instruction, extra support, tutoring, or a combination of these three aids. As described by Perie et al. (2009), one confounding variable on any predictive test is that, if it provides good feedback on how to improve a student's learning, then its predictive ability is likely to decrease. That is, if the test predicts that a student is not on track for college and career readiness, and then appropriate interventions are used to bring the student to that standard, the statistical analysis of the test's predictive validity should underpredict student performance over time. However, it is important to track the performance of students predicted to succeed on the summative test. And, although it should not be considered a strike against the predictive test if a student predicted to fail the summative test actually passes it after some intervention, questions *should* be raised if too many students predicted to pass the summative test actually fail it.

Henderson and colleagues from the Regional Educational Laboratory Northeast and Islands (REL-NEI) examined the effect of quarterly benchmark exams in mathematics on eighth-grade state test scores. Using an interrupted time-series design, they found that, after two years of implementation of the benchmark assessment program, participating schools posted small, non-significant gains over comparison schools (Henderson, Petrosino, Guckenburg, & Hamilton, 2008). Although the interventions were not studied in detail, the information from the predictive assessments was clearly used to provide interventions.

Growth Purposes

Given the recent emphasis on using test scores in teacher evaluation models, some states or districts may wish to use interim assessments to gather data on within-year growth. When the interim assessments are scaled with the summative assessment, growth in student performance can be calculated from the first administration of the interim assessment through the summative assessment. A key component is ensuring assessments are linked to allow for the measurement of growth.

Research suggests that there are differential learning losses over the summer related to income. As reported in Baker et al. (2010), a research summary concludes that, while students overall lose an average of about one month in reading achievement over the summer, lower-income students lose significantly more, and middle-income students may actually gain in reading proficiency over the summer, creating a widening achievement gap (Cooper et al., 1996). Therefore, some states and school

districts are looking for assessments to provide within-year growth that can more fairly be attributed to the teacher. Interim assessments provide a good mechanism for this purpose if designed with the intent of conducting growth calculations.

Blueprint Design

The purpose of the assessment will influence the design, and different designs are better suited for specific purposes. Consider an example from Gong (2010), modified to exemplify designs for the three purposes described earlier. Say the summative assessment measures 10 targets with two to five standards within a target. Figure 1 shows a possible set of content that might be assessed on a summative assessment. Although it does not provide information on the relative emphasis or distribution of the content, it will suffice as an example.

Figure 1. **Example summative design.**

Target	A	B	C	D	E	F	G	H	I	J
Standard	12	123	12	1234	123	12345	12	12	123	1234

The summative blueprint would have more detail on the number and types of items covering each target and standard but would include all of the content shown in Figure 1. There are several approaches to designing an interim assessment that would best prepare students for the content assessed by the summative assessment.

First, we could design a shorter test that mimicked the blueprint for the summative assessment. The forms could either match the summative blueprint exactly or strategically sample the content on the summative form. Parallel forms would be created off of the same blueprint that include different items but items that are parallel in terms of content and complexity. Administering parallel forms at multiple times during the school year will allow students to show growth toward mastering the content on the summative test and predict performance on the summative test, but it will also test them on content they have not yet been exposed to.

A second option is similar to the first, but instead of developing fully parallel items, the items would cover parallel content but increasing levels of complexity. This design follows the theory that students may know or learn the breadth early but gain a greater depth of understanding over the school year. This approach may, therefore, be more logical for an English language arts test than for a mathematics test. The number of items on the form and the number measuring each target would remain constant, easing some developing costs, but it would be important to select an appropriate increase in complexity over the forms. This design would work well if the purpose was either predictive or measuring growth.

If the content is the same as the summative, then the content covered by the interim assessment would be the same as shown in Figure 1. However, if the decision is made to sample that content, then the design might look something like Figure 2.

***Figure 2*. Example of an interim design sampled from the summative design.**

Target	A	B	C	D	E	F	G	H	I	J
Standard	12	13	1	124	12	125	2	1	23	34

A third option can be considered an end-of-unit type of test. That is, each form might only have two to four targets on it that correlate to the unit taught (see Figure 3). Teachers would administer the tests when the students have completed the related unit(s) of study. Districts could use these assessments as a pacing guide by requiring them to be administered at certain times or leave the decision to the teachers to individualize the assessment timing. Either the items would need to be pre-equated or linking items would be needed if student growth measures were desired. Otherwise, real-time scores would provide information on the degree to which students learned the material taught. A benefit of this approach is that students would only be tested on material that had been covered in the classroom. This approach would work well for an instructional purpose.

***Figure 3*. Example of an interim design focused on end-of-unit structures.**

Interim test #1

Target	A	B	C
Standard	12	123	12

Interim test #2

Target	D	E	F	G
Standard	1234	123	125	12

Interim test #3

Target	H	I	J
Standard	12	123	1234

A fourth design would be similar in theory to the previous one, but would be additive in design. That is, students would be tested on what they have learned so far in the school year. As shown in Figure 4, each form would cover recent and prior targets. Again, the benefit is that students would only be tested on material that had been taught, but, in this case, they would also be required to show that they retained the information from earlier in the year. This design would work well for both instructional and predictive purposes.

***Figure 4.* Example of an interim design that adds content throughout the year.**

Interim test #1

Target	A	B	C
Standard	12	123	12

Interim test #2

Target	A	B	C	D	E	F	G
Standard	12	123	12	1234	123	12345	12

Interim test #3

Target	A	B	C	D	E	F	G	H	I	J
Standard	12	123	12	1234	123	12345	12	12	123	1234

A fifth design combines the third and fourth to create overlapping but moving targets over time. That is, instead of covering everything that had been learned previously, the test might measure the previous unit and the current one only. This design works well for instructional purposes with some review and some new material. If all of the items are scaled with the summative assessment, it can also meet predictive and

growth purposes as secondary purposes. Figure 5 shows an example of how this overlapping and moving content might work.

Figure 5. **Example of an interim design focused on overlapping but moving content.**

Interim test #1

Target	A	B	C
Standard	12	123	12

Interim test #2

Target	B	C	D	E	F	G
Standard	3	12	1234	123	12345	12

Interim test #3

Target	F	G	H	I	J
Standard	45	12	12	123	1234

This was not an exhaustive list of possible assessment designs, but should provide examples of how to consider the design relative to the intended purpose. Table 2 summarizes the designs detailed in this section, describing the primary purpose, benefits, and weaknesses. It is worth noting that many designs could be used to fulfil multiple purposes but usually maximizes one purpose over the others.

Table 2. ***Summary of interim assessment designs***

Design	Purpose	Benefits	Weaknesses
Same blueprint with parallel items given multiple times a year, typically using a sample of the content to keep the test shorter in length	Most helpful for predictive and growth	Parallel forms are easy to construct and results make it simple to calculate growth. To keep the assessment shorter, the content targets most highly correlated with strong performance on the summative assessment can be emphasized.	Students are exposed to content that they have not yet been taught. If content is not sampled, these assessments are likely to be quite time consuming.
Same breadth blueprint given multiple times a year, but items vary in terms of complexity	Growth and some instructional feedback	This design may work well for ELA if the content is similar and what changes over the year is the complexity of the text with which students work.	Determining appropriate levels of complexity for each administration is a difficult task, and students are exposed to content that they have not yet been taught.

Organized by units/reporting categories	Instructional although each unit could predict performance on that unit in the summative assessment	Each assessment can be matched to an instructional unit or period of time. Also, it could serve another purpose as a district pacing guide.	Unless an equating plan is implemented, it is more difficult to measure growth in this model.
Shorter tests that build throughout the year in length and breadth	Primarily growth and can be instructional mainly on the most recent material	Students will only be assessed on material they have been taught and the assessment requires review of previous units in addition to the current unit.	The tests vary in length, which may make scheduling more complicated.
Overlapping but moving content	Through course as each piece is instructional, and overall the data can be used for predictive purposes		

Content Alignment

For interim assessments to meet any of the purposes, they need to align with the same content standards as the summative assessment. Assessments designed for instructional purposes must also be closely aligned with district curriculum (Goertz, Olah, & Riggan, 2010). Assessments designed for predictive purposes should align with key targets that most influence the total score on the summative assessment. So, alignment is a key component in designing interim assessment, but the purpose will help define the target of the alignment. In addition, if the assessment will serve a secondary purpose of a pacing guide, it should be tightly linked not only to the Common Core but to the instructional pacing guide adopted by the district.

Another consideration related to the design approach is whether each form should align to the full set of standards or whether the system as a whole should show full alignment. That is, if the comprehensive assessment system is comprised of three interim assessments and a summative assessment (and hopefully formative tools), the purpose and design of the interim assessment will determine whether the alignment study should be conducted on each form individually or on the interim item bank or set of forms. Alternatively, one design purposefully focuses on key targets that are deemed to make the greatest difference in student performance and long-term learning. If this design is used, we would expect a one-way alignment: The interim assessments would align with the Common Core, but not all of the Common Core would be assessed by the interim assessments.

Construction

Once the contents of the assessments are determined, other design and construction decisions need to be made. First, it will be important to determine whether the assessments should be fixed form or adaptive. Currently, Smarter has provided access to fixed-form assessments with blocks that can be

arranged in many of the different configurations described above. Conceivably, however, if states had access to an item pool, they could create an adaptive interim assessment. Particularly if the second design option was determined to be most beneficial, tracking a student's understanding of increasingly complex content would be well suited to an adaptive model.

A second consideration is who builds the forms. Should Smarter build forms at the consortium level? Should states build forms to follow the design approach that best suits their needs? Should that decision be made at the district level? Should teachers be able select items or blocks of items that best represent what they have taught? The answers to these questions will need to start at the consortium level, and then the degree of flexibility desired may lie within a state.

Regardless of who designs the tests, it is important to consider both the number and types of items to be included in the pool, on the blocks, or on the full form.

Types of items

If used for predictive purposes, types of items should parallel the types on the summative assessment. That is, if the summative test is comprised of 80% multiple-choice, 15% technology-enhanced, and 5% constructed-response items, then the interim assessment should follow that same distribution. If the interim assessment is intended to be used for instructional purposes, more constructed-response, or "open-ended" items will shed more light on student thinking. Open-ended test items that ask students to explain their answers increase the diagnostic value of benchmark tests. Students' responses reveal their thinking, helping teachers to refine their instructional strategies and design targeted instruction for individual students (Herman & Baker, 2005). However, do not assume that multiple-choice items cannot also provide instructionally-useful information. Multiple-choice items can also yield important diagnostic information, particularly when they are purposely designed so that distractors—the incorrect answer options—reflect common student misunderstandings (Herman & Baker, 2005). Even if multiple-choice items are written to provide instructionally useful information, such information still remains an inference made on the part of the teacher. Open-ended or constructed-response items allow students to reveal their own understandings and misunderstandings, although they again require the teacher to recognize, interpret, and respond to them.

Number of items

The number of items is affected by the purpose of the assessment, the frequency of administration, and the type of items. For example, 20 items may be sufficient to obtain adequate reliability coefficients at the district level, but teachers who want to use subscale scores for individual students for instructional purposes may require additional items. Fewer items per objective and a larger scope of objectives might be useful for predictive assessments measuring the aggregated performance of a group of students or for high-stakes summative assessments (Lalley & Gentile, 2009; Niemi et al., 2007). However, this approach would not provide much instructionally useful information.

Another approach is to administer an assessment every X number of weeks. One model is to assess every nine weeks, providing for three interim assessments prior to the summative assessment. In this scenario of assessing every nine weeks, 30 items should suffice to gauge where students fall on the

learning continuum. But all of these models assume the items are multiple-choice, technology-enhanced, or short constructed-response. Teachers could also use a very few number of items if the forms were comprised of performance tasks or extended constructed responses. However, this approach would work only for instructional purposes or to prepare students for the performance tasks and essay on the Smarter summative assessment.

Evaluative Criteria

When considering the design of an interim assessment, one should also consider the validity evaluation. The following criteria are taken from an article by Herman and Baker (2005) on building effective benchmark assessments. Although the focus is on interim assessments used for instructional purposes, these guidelines are applicable to all assessments used in a classroom.

1. Align standards and interim assessments from the beginning of test development. Decide what specific content to assess and at what level of intellectual demand. Include the application of complex learning. To create interim tests that enrich student learning opportunities, focus on the big ideas of a content area and counteract curriculum narrowing by designing interim tests that allow students to apply their knowledge and skills in a variety of contexts and formats.
2. Enhance the diagnostic value of assessment results through initial item and test structure design. Use extended-response items to reveal student thinking and potential misconceptions. Build distractors into multiple-choice items that reveal common student misunderstandings.
3. Ensure the fairness of interim assessments for all students, including English language learners and students with disabilities. Avoid unnecessarily complex language or specific contexts that could unfairly confound some students' ability to show what they know.
4. Use the data to examine the tests' technical quality, such as reliability. Reliability and accuracy are necessary but not sufficient prerequisites to validity (that is, the extent to which a test accomplishes its intended purposes). If the interim assessment is used for predictive purposes, there should be a strong predictive relationship between students' performance on the interim assessments tests and students' performance on the summative assessments. (Unless, as stated earlier, interventions improve performance on the summative assessment.)
5. Build in utility. Design reports of test results to be user-friendly and to provide guidance on how to appropriately interpret and use the results. Teachers need to understand how to understand student misconceptions and to know the next steps to use in teaching.
6. Hold interim testing accountable for meeting its purposes. Crafting good interim tests and ensuring their wise use for improving student learning requires systematic design and continual evaluation.

Criteria specific to assessments used for predictive or growth purposes can be found in Perie et al. (2009):

1. The assessment should be highly correlated with the criterion measure (e.g., the end-of-year state assessment). The technical documentation should include evidence of the predictive link between the interim assessment and the criterion measure.

2. The predictive assessment should be comprised of items with a similar mix of item types as the criterion measure.
3. If used for growth, the items in all interim assessment forms should be calibrated and scaled with the summative assessment.
4. The predictive assessment should be designed from the same or similar blueprint as the criterion measure. To be maximally useful, each test should allow for an analysis of results by what has been covered for that student in school and what has not.
5. Even when used for predictive or growth purposes, the assessment should contain enough diagnostic information so that remediation can be targeted for students predicted to score below the cut on the criterion measure. If the assessment is unable to provide such information, additional guidance should be included in the system to help with remediation.

Other Conditions for Strong Interim Assessments

Designing and building a strong interim assessment is only the first step. Working with teachers on implementing the assessment, interpreting the results, and using the data to effectively modify instruction are key components for a successful system. Research has shown that, for interim assessments to provide the most benefit to student learning, appropriate conditions must be in place at the district and school level, and that teachers must be trained appropriately.

First, the state role is to provide access to an interim assessment system. This goal will be accomplished through the Smarter consortium although, if more than one option is provided by the consortium, states may want to provide guidance for their districts. Because one part of the definition of interim assessment is that data can be aggregated at many different levels, districts are often heavily involved in the data analysis. Goertz et al. (2010) followed interim assessment practices in multiple districts and schools and found several conditions at the district level that fostered positive results in the use of interim assessments.

First, districts that communicate strong expectations for data use are more likely to see schools administering them and spending time analyzing results. If the district is responsible for the analysis of the assessment data, then it is important that they generate that data in a timely fashion. In addition, the analyses must be presented in an accessible manner that can inform instruction. Beyond providing the data, the district will need to ensure that schools build in dedicated time for teachers to analyze data and plan instruction. These activities are often more effective in a team environment. Any pacing guide also has to allow for time to reteach students on content they did not learn. Finally, districts should provide instructional support for teachers and students. This support could be in the form of materials, a consultant, or resources to support learning before and after school.

Second, conditions must be favorable at the school level to promote good usage of interim assessments. These conditions start with strong leadership that fosters an environment where teachers review and analyze results in teams and develop a shared understanding of how to adjust instruction accordingly. It takes time and leadership to develop a culture of data use. Ideally, there would be time devoted to these tasks during a professional development day. Moreover, for interim assessment systems that

require a break from instruction in order to test, educational leaders should consider the time required for assessment, which should be as short as possible to provide the desired information. For certain performance tasks that are less distinguishable from instruction than more formal tests, the issue of "testing time" is less of an issue but still must be considered.

Finally, attention needs to be devoted to increasing teacher capacity in data analysis and use. Guidance should be provided to teachers to teach them how to interpret results correctly. Good coaching can help them learn how to identify student misconceptions or lack of understanding and reteach effectively. Goertz et al. (2010) identified different techniques for probing deeply on conceptual compared to procedural learning. Teachers need to understand those differences and timing for applying different interventions on teaching what students have not yet learned in previous instruction.

Summary

In conclusion, states need to determine the primary purpose for the interim assessments and select a design that best suits that purpose. Smarter can facilitate that effort by providing blocks or "testlets" of interim items that can be combined in multiple ways. Then, instructions for assembling those forms should be provided to districts along with suggested analyses of the results. Most importantly, districts and schools need to work with the teachers on interpreting the results and using them to better meet the learning goals of their students.

Ideally, usage of interim and formative assessments should be tracked and compared to performance on summative assessment. Understanding what types of information and interventions have the most impact will help educators maximize their time and efforts with the students. Smarter can also facilitate the data analyses across states for research purposes with the goal of better help students graduate ready for college and careers.

References

Baker, E., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., Ravitch, D., Rothstein, R., Shavelson, R., & Shepard, L. (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing Paper #278). Washington, DC: Economic Policy Institute.

Christman, J., Neild, R., Bulkley, K., Blanc, S., Liu, R., Mitchell, C., & Travers, E. (2009). *Making the most of interim assessment data: Lessons from Philadelphia*. Philadelphia, PA: Research for Action.

Clune, W. H., & White, P. A. (2008). *Policy effectiveness of interim assessments in Providence public schools* (WCER Working Paper No. 2008-10). Madison, WI: University of Wisconsin-Madison, Wisconsin Center for Education Research.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*, 227–268.

Crane, E. (2010). *Building an interim assessment system: A workbook for school districts*. Washington, DC: Council for Chief State School Officers. Retrieved from http://www.ccsso.org/Documents/2010/Building_an_Interim_Assessment_2010.pdf

Goertz, M. E., Olah, L. N., & Riggan, M. (2010). *From testing to teaching: The use of interim assessments in classroom instruction* (CPRE Research Report No. RR-65). Philadelphia, PA: Consortium for Policy Research in Education.

Gong, B. (2010). *Some implications of the design of balanced assessment systems for the evaluation of technical quality of assessments*. Dover, NH: National Center for the Improvement of Educational Assessment, Inc. Retrieved from http://www.nciea.org/publication_PDFs/RILS_Gong2010.pdf

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (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, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands.

Herman, J. L., & Baker, E. L. (2005). Making benchmark testing work. *Educational Leadership, 63*(3), 48–54.

Lalley, J., & Gentile, J. (2009). Classroom assessment and grading to assure mastery. *Theory into Practice, 48*, 28–35. doi:10.1080/00405840802577577

Marshall, K. (2006). *Interim assessment: Keys to successful implementation*. New York: New Leaders for New Schools. Retrieved from http://www.marshallmemo.com/articles/Interim%20Assmt%20Report%20Apr.%2012,%2006.pdf

Niemi, D., Vallone, J., Wang, J., & Griffin, N. (2007). *Recommendations for building a valid benchmark assessment system: Interim report to Jackson Public Schools* (CRESST Report 723). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/R723.pdf

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*(3), 5–13. doi:10.1111/j.1745-3992.2009.00149.x

Stecher, B. M., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., McCombs, J. S., Russell, J., & Naftel, S. (2008). *Pain and gain: Implementing No Child Left Behind in three states, 2004–2006*. Santa Monica, CA: RAND Corporation.

Stiggins, R., Arter, J., Chappuis, J., & Chappuis, S. (2006). *Classroom assessment for student learning: Doing it right—using it well*. Merrill Education/ETS college textbook series. Upper Saddle River, NJ: Pearson Education, Inc.

TILSA SCASS Interim Assessment Subcommittee (2008). *Interim assessment practices and avenues for state involvement*. Washington, DC: Council for Chief State School Officers. Retrieved from http://www.ccsso.org/Documents/2008/Interim_Assessment_Practices_and_2008.pdf

Wiliam, D., & Leahy, S. (2006, April). *A theoretical foundation for formative assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.