Measures that Matter

The Role of Interim Assessments in a Comprehensive Assessment System

By Marianne Perie, Scott Marion, Brian Gong (National Center for the Improvement of Educational Assessment) and Judy Wurtzel (The Aspen Institute)

Center for Assessment
Achieve
THE ASPEN INSTITUTE

ABOUT THIS REPORT

School districts increasingly see interim assessments as an important element in their instructional improvement strategy and are implementing interim assessments district-wide in multiple grades and content areas. Yet, they face significant design and implementation challenges to realizing the potential of interim assessments to improve teaching and learning. The Aspen Institute Education and Society Program and Achieve, Inc. partnered with the National Center for the Improvement of Educational Assessment to develop this policy brief. Our goals are to clarify how interim assessments fit into the landscape of formative assessment and to offer recommendations to districts on the appropriate use of interim assessments as part of a broader assessment system. The brief is informed by conversations with the Aspen Institute Urban Superintendents Network and the affiliated Dana Center-Achieve Urban Mathematics Leaders Network.

The report is available for download at the websites of all three organizations.

Achieve, Inc.

Achieve, Inc. was created by the nation's governors and business leaders in 1996. As a bipartisan, non-profit organization, Achieve helps states raise academic standards, improve assessments and strengthen accountability to prepare all young people for postsecondary education, work and citizenship. The American Diploma Project (ADP) Network is an Achieve initiative that includes 30 states working to raise high school standards, strengthen assessments and curriculum, and align expectations with the demands of college and work. Achieve provides policy and advocacy leadership, technical assistance and other support to the ADP Network states and is leading a number of initiatives aimed at meeting specific challenges states face while implementing the ADP policy agenda. *(www.achieve.org)*

THE ASPEN INSTITUTE

The Aspen Institute seeks to promote nonpartisan inquiry and an appreciation for timeless values through its seminars, policy programs, conferences and leadership development initiatives. For over two decades, the Aspen Institute's Education and Society Program has provided an informed and neutral forum to aid education practitioners, researchers and policy leaders in their efforts to improve student achievement and education policy. *(www.aspeninstitute.org/education)*

Center for Assessment

The National Center for the Improvement of Educational Assessment, Inc strives to increase student achievement through improved practices in educational assessment and accountability in the United States. The Center works with states and other educational agencies to design and implement effective assessment and accountability policies and programs. *(www.nciea.org)*

The Role of Interim Assessments in a Comprehensive Assessment System: *A Policy Brief*

The standards-based reform movement has resulted in the wide-spread use of summative assessments designed to measure students' performance at specific points in time. While many have hoped that these end-of-year tests would provide instructionally useful information for educators, they do not. This is not because there is something "wrong" with these summative accountability tests, rather that they were not designed to meet instructional purposes. Recognizing the inherent limitations of summative assessment, educators are looking for additional assessments to inform and track student learning during the year. Large numbers of vendors are now selling what they call "benchmark," "diagnostic," "formative," and/or "predictive" assessments with promises of improving student performance.

These systems often lay claim to the research documenting the powerful effect of formative assessment on student learning. However, the research in this area, including the seminal Black and Wiliam (1998) meta-analysis, evaluated formative assessments of a very different character than essentially all current commercially-available interim assessment programs.

This policy brief provides a definition to differentiate between true classroom formative assessment and the interim assessments currently in the marketplace. It also describes a framework for understanding what purposes interim assessments can serve and considering their appropriate role in a comprehensive assessment system. Our goal is to help district leaders thoughtfully examine the commercially-available interim assessment products, develop strong specifications for a customized system, develop their own interim assessments, or determine that interim assessments are not an appropriate tool for their assessment system. A second goal is to help state leaders consider what role they might play in supporting effective interim assessment practices. (The full paper from which this brief is drawn, *A Framework for Considering Interim Assessments,* February 13, 2007, is available at: www.nciea.org).

This Policy Brief is organized into seven sections:

Section 1
Distinguishing Among Assessment Types

Our schema recognizes three assessment types —summative, interim, and formative—and distinguishes among them based on their intended purposes, audience, and use of the information, rather than simply based on when the assessment is given. While this policy brief focuses on interim assessment, we define the three types of assessment here since we believe that clarifying the distinctions between these three types of assessments is a critical first step in determining the appropriate role of interim assessments in a comprehensive assessment system.

Summative assessments are generally given one time at the end of some unit of time such as the semester or school year to evaluate students' performance against a defined set of content standards. These assessments typically are given statewide (but can be national or district) and these days are usually used as part of an accountability program or to otherwise inform policy. They are the least flexible of the assessments.

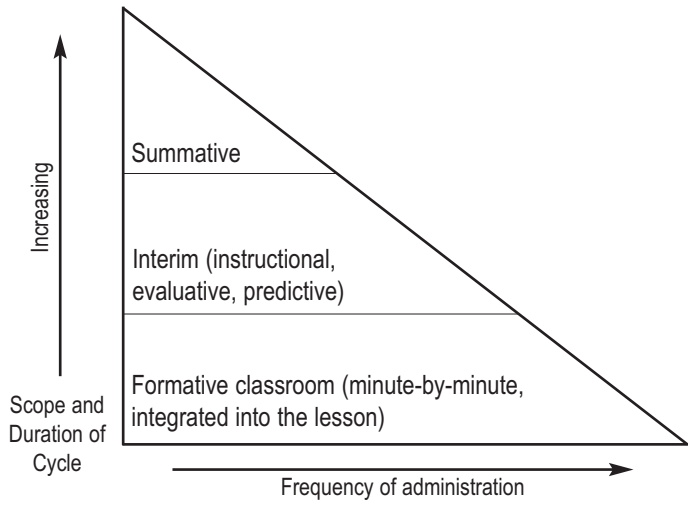
Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes. Thus, it is done by the teacher in the classroom for the explicit purpose of diagnosing where students are in their learning, where gaps in knowledge and understanding exist, and how to help teachers and students improve student learning. The assessment is embedded within the learning activity and linked directly to the current unit of instruction. The assessments are small-scale (a few seconds, a few minutes, certainly less than a class period) and short-cycle (they are often called "minute-by-minute" assessment or formative instruction). Furthermore, the tasks presented may vary from one student to another depending on the teacher's judgment about the need for specific information about a student at a given point in time. Providing corrective feedback, modifying instruction to improve the student's understanding, or indicating areas of further instruction are essential aspects of a classroom formative assessment. There is little interest or sense in trying to aggregate formative assessment information beyond the specific classroom.

Interim assessment is the term we suggest for the assessments that fall between formative and summative assessment, including the medium-scale, medium-cycle assessments currently in wide use. Interim assessments (1) evaluate students' knowledge and skills relative to a specific set of academic goals, typically within a limited time frame, and (2) are designed to inform decisions at both the classroom and beyond the classroom level, such as the school or district level. Thus, they may be given at the classroom level to provide information for the teacher, but unlike true formative assessments, the results of interim assessments can be meaningfully aggregated and reported at a broader level. As such, the timing of the administration is likely to be controlled by the school or district rather than by the teacher, which therefore makes these assessments less instructionally relevant than formative assessments. These assessments may serve a variety of purposes, including predicting a student's ability to succeed on a large-scale summative assessment, evaluating a particular educational program or pedagogy, or diagnosing gaps in a stu-

dent's learning. Many of the assessments currently in use that are labeled "benchmark," "formative," "diagnostic," or "predictive" fall within our definition of interim assessments.

The triangle in Figure 1 illustrates the distinctions between these three types.

Figure 1. Tiers of Assessment



We have started with these definitions of summative, interim and formative assessment because we believe that even assessment experts have been hobbled by the lack of clear definitions of assessment types and the abundant use of the term "formative assessment" to apply to assessments of very different design and purposes. This imprecision has led to a blurring of the differences between what we call "formative assessment" and what we call "interim assessment". Districts putting in place interim assessment systems may be getting important and actionable data, but they are rarely getting the power of true formative assessment practices.

Section 2
Questions to Start With

As policymakers decide whether to bring an interim assessment system to their state/district/school, we encourage them to articulate clearly the intended purpose and use of interim assessments and how the particular assessment system will work in the teaching-learning cycle. As a start, we think it will be helpful for educational leaders to address the following questions:

1. What do we want to learn from this assessment?
2. Who will use the information gathered from this assessment?
3. What action steps will be taken as a result of this assessment?
4. What professional development or support structures should be in place to ensure the action steps are taken and are successful?
5. How will student learning improve as a result of using this interim assessment system and will it improve more than if the assessment system was not used?

The answers to these questions will reveal a theory of action about how assessments will lead to improved student learning. The answers will dictate the type of assessment needed and will drive many of the design decisions including the types of items used, the mechanism for implementing it, the frequency with which it should be administered, and the types of reports that will need to be developed from the data. Importantly, these questions and the associated answers serve as the beginning of a validity argument in support of (or to refute) the particular assessment system. While this policy brief focuses primarily on helping policy makers answer the first question, we emphasize that it is essential to consider all five when making decisions about the purchase or design and implementation of an interim assessment system.

Section 3
Determining the Purpose for the Interim Assessment

A critical task for policy makers is to answer explicitly the first question posed above — "What do I want to learn from this assessment?" — and then find or develop a set of assessments that best fits that purpose. We see three general classes of purposes for interim assessments: *instructional, evaluative,* and *predictive*. Within each general class, there are myriad specific purposes. Each general class and examples within them are described below.

Of course, educational leaders are often trying to squeeze as many purposes as possible out of a single system. However, few assessments or assessment systems can serve more than two or three purposes well and they tend to work best when the various purposes have been prioritized explicitly. Thus, these descriptions of assessment purposes are followed by a discussion of the challenges inherent in attempting to use one assessment to serve multiple purposes.

A. Instructional Purposes

Interim assessments designed to serve *instructional purposes* should provide results that enable educators to adapt instruction and curriculum to better meet student needs. Within this general category of instructional purposes, policymakers and assessment leaders must go further and articulate and prioritize the specific instructional purposes for their interim assessment system to better guide the design and/or selection. For example, interim assessments can be used to enrich the curriculum, determine students' strength and weakness in a particular domain, or provide feedback to students for motivational and metacognitive reasons.

When the purpose is to enrich the curriculum, assessments are designed to have students explore concepts in greater depth or provide tasks that stretch students and teachers to do things at deeper cognitive levels than they might otherwise. The assessment itself provides the rich instruction.

When the purpose is to illuminate strengths and weaknesses of individuals or groups of students, an assessment system often contains a bank of items aligned with the state curriculum 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 standard[1] allowing teachers to identify strengths and weaknesses in the students' learning. Ideally, to provide actionable information the assessment is fully aligned with the specific classroom or at least school curriculum and provides more in-depth analyses of student misconceptions or lack of understanding along with instructional tools and strategies for improving instruction. An interim differs from a true formative assessment even when used for similar instructional purposes in terms of the timing and the capability for aggregating the results beyond a single classroom.

1 Unless of course the assessment is focuses on only one content standard (or less) in depth.

Finally, if the specific instructional purposes are targeted toward motivating and providing feedback to students about their learning, tasks should be designed to engage students and encourage them to wrestle with challenging subject matter knowledge. Quick feedback afforded by computer-based testing programs can likely motivate students and provide them with feedback about their strengths and weaknesses in particular content domains. Rich tasks that make student thinking and productions explicit, such as exhibitions and projects, can also provide motivation and feedback. The goal, in any case, is for policymakers to ensure that the feedback is worthwhile and accurate.

B. Evaluative Purposes

The primary goal of interim assessments designed for *evaluative purposes* is to provide information to help the teacher, school administrator, curriculum supervisor, or district policymaker learn about curricular or instructional choices and take specific action to improve the program, affecting subsequent teaching and thereby, presumably, improving the learning. Think of this as a program evaluation designed to change curriculum and instruction not necessarily in mid-term but over the years.

Assessment systems designed to serve evaluative purposes must provide detailed information about relatively fine-grained curricular units. However, not every student needs to be assessed in order for the teacher or administrator to receive high-quality information from an evaluative assessment. Examples of assessments used for evaluative purposes include assessments given at various points throughout the year to provide more details about student performance on instructionally-relevant subdomains (e.g., adding simple fractions)—not with the intention of intervening but for evaluating the effectiveness of a program, strategy, or teacher. Moreover, assessments can be used to learn which concepts students understood well and which were less clearly taught by teachers within one or more schools with the goal of helping them modify the curriculum and instructional strategies for future years. Similarly, they can be used to provide a more in-depth understanding at the school level on how the test items link to the content standards and how instruction can be better aligned with improved performance on the test.

Another set of important evaluative purposes for interim assessments can be to enforce some minimal quality through standardization of curriculum and pacing guides, centralizing coordination for highly mobile urban student populations and high teacher turn-over, or as a lever to overcome differences in learning expectations and grading standards.

C. Predictive Purposes

Predictive assessments are designed to determine each student's likelihood of meeting some criterion score on the end-of-year tests. End users should be able to aggregate the results to the classroom, subgroup, school, and even district level. We suspect that there are few assessment systems where the only purpose of the system is to predict performance on some later assessment. Nevertheless, the predictive purposes of interim assessments are important to many users, and this interest could increase as the annual NCLB targets continue to rise and there is increasing emphasis on high school exit exams.

In addition, we have received anecdotal reports from certain users that scores from these predictive-type assessments can serve as a screener. That is, in some districts, predictive tests are 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, and/or tutoring. This scenario could be an example of how interim and formative assessments work together to help improve student performance on a summative assessments. It also highlights the value of having all three of these assessment types aligned in a comprehensive assessment system.

D. Multiple Purposes

Given constrained resources, it is no wonder that educational leaders are tempted to use a single assessment system for as many purposes as possible. Unfortunately, one of the truisms in educational measurement is that when an assessment system is designed to fulfill too many purposes—especially disparate purposes—it rarely fulfills any purpose well. This does not mean that certain interim assessment systems cannot fulfill more than one purpose. If the system is intended to provide rich information about individual students' strengths and weaknesses tied to a particular set of curricular goals, then these results can likely be aggregated to the subgroup, school, and/or district level to provide evaluative and predictive information. On the other hand, if the primary goal is to gather predictive or early warning information, it is unlikely that the assessment will contain rich enough information for full instructional or even evaluative purposes.

Therefore, if users want to fulfill multiple purposes, they must design a system to fulfill the finest grain purposes first. Users can then consider approaches to aggregate the results to more general levels in the educational system. Alternatively, users can carefully and explicitly differentiate between the assessments designed for use with all students and those for use with students for whom more fine grained information is needed.

Section 4
Characteristics of an Effective Interim Assessment System to be used for Instructional Purposes

Once educational leaders are clear about purposes for the interim assessment system, they still face myriad additional considerations and decisions. This section provides guidance in sorting through those considerations and decisions by identifying characteristics of effective interim assessment systems.

It is important to note at the outset that there is little research on what kinds of educational results can reasonably be expected from interim assessment and thus little evidence about the characteristics of an effective interim assessment system. Our sense, however, is that it is reasonable to expect that interim assessments, when well designed, can accomplish evaluative, predictive and instructional purposes. Our work with many states and districts and conversations with state, district and assessment leaders suggests some common sense guidance that can help guide decision making until better evidence is developed.

Here we focus on the characteristics of interim assessments for instructional purposes. This is because most districts appear to want to use assessments for this purpose, most vendors say their assessments can meet that purpose, and we have more concerns about claims for *instructional purposes* then for evaluative and predictive purposes.

A. General Design Characteristics

There is no one-size-fits-all assessment, only a best design for a desired use and the existing constraints and resources. We recognize that some districts or states will be looking to purchase an already-available assessment system, while others will be looking to create a system customized to their needs. The considerations described below are appropriate for both approaches.

The general characteristics of any interim assessment that is to be used for instructional purposes include:

- Not *all* multiple-choice.
- Provision for *qualitative* insights about understandings and misconceptions and not just a numeric score.
- Immediate implications for what to do besides re-teaching every missed item.
- Rich representation of the content standards students are expected to master
- High quality test items that are directly linked to the content standards and specific teaching units
- A good fit within the curriculum so that the test is an extension of the learning rather than a time-out from learning
- A good fit with curriculum pacing so that students are not tested on content not yet taught

- Clear reporting that provides actionable guidance on how to use the results
- Validation of the uses of and information provided by the assessment
- Administration features (speed, availability of normative information, customization, timing flexibility; adaptive) that match the assessment purposes
- Professional development for teachers

While each item on this checklist could be discussed in-depth, we believe that five merit particular attention: reporting, item type, scoring, item quality, and professional development.

B. Reporting Results

One strategy for defining the desired characteristics is to focus on reporting. What do we want the tests to tell us? Score reports serve to make the results actionable. We recommend visualizing and designing the intended reporting system as a way of clarifying all the information desired from the assessment. Assessments serving an instructional purpose will have different features in their reports than those serving predictive or evaluative purposes.

A score report should go beyond indicating which questions were answered incorrectly; it should inform a plan for action to further student learning. Technology has helped make it possible to "slice and dice" assessment data in myriad ways, and unfortunately it has become easy and tempting to provide voluminous reports to teachers, administrators, parents and students. However, most people can only absorb and act on small amounts of data at a time. We suggest that reporting systems should be built so that initial reports provide only the most important actionable data – with provisions for easy access to additional on an "as wanted" basis. Over time, as users become more comfortable using reports and demand for more data builds, reports might be redesigned to provide additional data. Further, to judge a reporting system's effectiveness, it must be vetted with those who need to use the information: teachers in most cases but also school leaders.

A particularly important issue in the reporting of interim assessment data is whether and how to make the data public and whether and how to incorporate the data into formal or informal accountability systems. While there is no hard evidence on the best approach, our sense is that the results of interim assessments should be made public within the district (among teachers, administrators and parents) but should not be used for accountability purposes. This is particularly true if assessments are to be used for instructional purposes and the goal is for teachers to use assessment results as the basis for conversations among themselves and with their students about the nature of students' work and the changes in their own practice that are needed to improve this work. For such conversations and analyses to take place, teachers must believe in – and not fear – the assessment results.

C. Item Type

Only after the purpose and desired form of results are clear, can we begin to think about the types of items that would be appropriate for the interim assessment. One of the goals of this paper is to broaden the discussion of interim assessments to include more locally-developed or other customized approaches. Adopting this viewpoint allows us to consider a wider range of item types than is typically the case with commercial systems. Performance tasks, particularly extended tasks, can serve instructional purposes more readily than other interim assessment item types. They enrich the curriculum, provide opportunities for more in-depth focus on the content area, and provide opportunities for teachers to learn about student thinking as they observe students working on the tasks. These tasks can be relatively short, such as graphing the results shown in a table, or more complex, such as designing, carrying out, and analyzing a science experiment. Again, as long as the results can be aggregated and used at a level beyond the classroom (which can be done through systematic observations, rubrics, and other scoring methods), an assessment with these types of tasks falls under our definition of interim.

Extended performance tasks such as a research paper, science laboratory, or historical debate, have the advantage of helping to erase the familiar boundaries between assessment and instruction. When students are engaged in such tasks, an observer struggles to determine whether there is an assessment underway or simply an interesting instructional unit. Perhaps most importantly, these types of performance tasks, when well designed, increase student motivation by engaging them in meaningful interactions with rich subject matter.

D. Item Quality

An assessment can be only as good as the quality of the items, no matter how much thought and good intentions were part of the design. Good assessment items represent in-depth coverage of the specific content standards and curricular goals to be assessed, provide information about students' depth of understanding, and identify students' learning gaps and misconceptions. To accomplish these goals, items must provide opportunities for students to demonstrate their thinking and they must be built on research related to how students' learning progresses — that is how they develop proficiency in a particular domain.

Because so many widely used interim assessment systems rely primarily on multiple-choice and short constructed-response items, the accompanying box describes some of the design features that allow these types of items to yield instructionally relevant information.

Potential users of commercial systems should conduct structured, expert-led reviews of the quality of items that will be used on their tests. Consider developing a set of criteria for including items in a centralized item bank. Similarly, those developing a customized or other local system need structured expert reviews to ensure that the assessments can validly fulfill the intended purposes. It sounds overly obvious to say that the quality of the interim assessment system is dependent upon the quality of the items included in such systems, but this point often gets overlooked.

Characteristics of multiple choice and short answer interim assessment items that can yield instructionally relevant information

Well-designed multiple-choice items can provide constructive feedback on the breadth of students' knowledge as well as providing a quick check on misconceptions or incomplete understandings. A classic example of a multiple-choice item that provides evidence of student learning is the following:[2] *Consider the four diagrams shown below. In which of the following diagrams, is one quarter of the area shaded?*

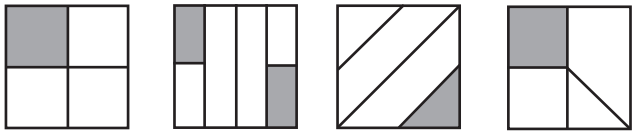


Diagram A is the obvious answer, but B is also correct. However, some students do not believe that one quarter of B is shaded because of a belief that the shaded parts have to be contiguous. Students who believe that one quarter of C is shaded have not understood that one region shaded out of four is not necessarily a quarter. Diagram D is perhaps the most interesting here. One quarter of this diagram is shaded, although the pieces are not all equal; students who rely too literally on the "equal areas" definition of fractions will say that D is not a correct response. By crafting questions that explicitly build in the under- and over-generalizations that we know students make, we can get far more useful information about what to do next.

For items such as these to be instructionally useful, there should be a clear theoretical basis related to how students develop proficiency in the domain when developing the item distractors. There must be a sound research basis for linking each wrong answer to a specific gap in learning. In addition, as good as any one item may be, usually many multiple-choice items are required to gain real insight into why a student answers incorrectly.

Another way to enrich the information gained from multiple-choice items or from short constructed response items is to ask the student to justify their response to each item. Asking questions such as "why did you select that answer?" or "what rule or information did you use to arrive at your decision?" or simply asking students to explain their answer can provide additional insights into the student's thinking. This strategy also allows the test to serve two purposes: to provide quick aggregate data on student performance and to provide more in-depth information about a student's thinking process that a teacher can use to inform instruction.

In some cases more information can be obtained from a related set of items than from a single item. Consider, for example, a simple arithmetic problem. If a student answers a subtraction problem as follows

584

-68

524

all we know is that the student answered the item incorrectly. However, if we look at a *set* of student responses

584	495	311	768	821
-68	-73	-82	-34	-17
524	422	371	734	816

we now have more information to process. On the surface, we know that the student answered 2 of the 5 items correctly, but if we look closely, we see that the student makes the same error on the three items answered incorrectly. On items 1, 3, and 5, where the number on the bottom contains a digit that is higher than the one on the top, the student is simply reversing the order. That is, in the first item, the student should be subtracting the 8 from the 4 (carrying a 1 to make it 8 from 14), but instead, the student flips it so that s/he is subtracting the 4 from the 8. The same error is made on all three items, providing richer information on the type of corrective instruction needed.[3]

---

2 This item is taken from "Does Assessment Hinder Learning?" Paper given by Dylan Wiliam at ETS Europe Breakfast Salon, July 11, 2006, Institute of Civil Engineers, London, and available online at: http://www.uk.etseurope.org/home-corpo-uk/free-resources/free-resources-detail/?doc=1077

3 An expanded version of this and other task sets can be found in Pellegrino, Chudowsky, and Glaser's (eds.) *Knowing what students know: The science and design of educational assessment.*

E. Scoring

Scoring is also an important consideration. While electronic scoring allows results to be produced quickly and aggregated easily across classrooms and schools, students self-scoring or teachers scoring student work is a learning opportunity in itself. This is particularly true for open-ended items where examination of the raw student work may enable teachers to observe and interpret patterns in student responses that may be lost in a scoring guide. Scores can then be recorded and uploaded for aggregation across classrooms or schools.

F. Professional Development

To serve instructional purposes, an assessment system must go beyond simply providing data. It must include strong supporting materials for interpreting and using the data to effectively modify classroom instruction. Specifically, it should include guidance on the types of errors to look for, how to detect errors, possible interpretations for each incorrect answer, and what corrective steps can be taken once a problem is identified. Again, these common errors on student thinking need to be identified based on modern conceptions of learning and empirical research on the nature and distribution of such misconceptions.

It is worth noting a tension between the need for professional development to accompany these assessment systems and the ownership of that responsibility. It is the contention of many assessment developers that tools and strategies for improving instruction are the teacher's responsibility, not the instrument provider's. Many policymakers, however, want to see that professional-development support included in the package that they purchase. We lean toward the latter viewpoint and believe an assessment system purchased for instructional purposes must include professional development to ensure that educators have the tools to use the assessments and the results appropriately. We believe that this should be a shared responsibility among the developer and the user.

Section 5
Current Commercially Available Interim Assessment Systems

Having determined their purposes and intended uses and thoughtfully considered the test design issues in the prior section, educational leaders can choose among many assessment systems in the market. In this section, we offer some general observations about criteria policy makers might consider when deciding whether to purchase commercial available assessment systems.

Many test publishing companies offer interim assessment products, often labeled "formative" or "benchmark" assessment products. These assessments are marketed to serve a plethora of purposes, including serving as diagnostic tools, providing information that can be used to guide instruction, determining student placement, measuring growth or progress over time, and predicting success on a future assessment. Typically these systems consist of item banks, administration tools, and customized reports. These systems often are computer-based and even web-based, allowing students to take the test whenever they wish (or their teacher wishes) and wherever a computer with an internet connection is available. Others also have the option of creating pencil-and-paper tests. Teachers can construct the tests, the tests can be fixed by an administrator, or the tests can be adaptive.

The items are "linked" to content standards[4], and results typically are reported in terms of number correct. The "diagnostic" portion tends to be a summary of results by content standard, allowing the teacher to see which standards students perform well on and which they do not. Often these systems provide a variety of options for reports, with different levels of aggregation. A student-level report indicates which items students answered correctly or incorrectly, while a classroom report might indicate the percentage of students answering each item correctly or the average percent correct for each content standard.

Some of the products have been linked to state end-of-year assessments, allowing them to serve a predictive function. Some of these systems have reported correlations and other indices to document the statistical link between the interim and summative assessments as evidence of the interim assessment's predictive ability. Most products include at least some measures of their reliability as well.

These products are marketed as being very flexible, giving instant feedback, and providing diagnostic information on which areas need further instruction. However, these systems generally fail in providing rich diagnostic feedback regarding student thinking. That is, few provide any information on why a student answered an item incorrectly or how best to provide corrective feedback. For instance, many of these computer-based assessments rely primarily on multiple-choice items. Unless each wrong answer provides insight into the nature of the student's incorrect thinking, the only information received from this type of item is essentially a correct/incorrect dichotomous response. Likewise, open-ended items need to result in more than a score, preferably in a summary report of the types of errors a student is making or of the areas of strength and weakness in a given performance (e.g., his/her writing).

4 Unfortunately, the strength of the alignment between such commercial tests and the state content standards is rarely evaluated by independent analysts, so the "link" between the two is often based on the publishers' claims.

In spite of these caveats and concerns, the best current commercially-available systems can:

- Provide an item bank reportedly linked to state content standards,
- Assess students on a flexible time schedule wherever a computer and perhaps internet connections are available,
- Provide immediate or very rapid results,
- Highlight content standards in which more items were answered incorrectly, and
- Link scores on these assessments to the scores on end-of-year assessments to predict results on end-of-year assessments.

Many of the better commercially-available interim assessment products can address questions such as:

- Is this student on track to score Proficient on the end-of-year NCLB tests?
- Is the student improving over time?
- What proportion of students is at risk of scoring below Proficient on the end-of-year NCLB tests?
- On which content standards are the students performing relatively well (or poorly)[5] (for a student, classroom, school, district, state)?
- How does this student's performance compare to the performance of other students in the class?

We must note that most of the features described above are not necessarily instructional and while most systems meet some of the requirements for an effective interim assessment, few, if any fully meet all of the criteria. Again, the focus remains on the purpose. If the purpose of these assessments is to enrich the curriculum, challenge the students to self-diagnose their own learning, provide insights into any misconceptions the students have, or provide additional professional development for the teachers, many of these types of assessment systems are woefully inadequate.

We find that most commercially-available interim assessment systems currently do not:

- Address well multiple purposes, i.e., instructional, evaluative, or predictive
- Provide rich detail about the curriculum assessed,
- Help teachers understand the nature of a student's misconception(s),
- Report detailed information on the student's depth of knowledge on a particular topic,
- Further a student's understanding through the type of assessment task,
- Give teachers the information on how to implement an instructional remedy.

5 This assumes that there are enough items for given strands or standards to determine if differences are reliably different.

Furthermore, these systems typically do not answer the following questions:

- Why did a student answer an item incorrectly?
- What are possible strategies for improving performance in this content area?
- What did the student learn from this assessment?
- What type of thinking process is this student using to complete this task?

While we emphasize these concerns about the weakness of available interim assessment systems for instructional purposes, we want to be clear that interim assessments can play a productive role in this area and in others. In particular, going to the other classes of purposes we noted at the beginning – predictive and evaluative – there is much that interim assessments might offer such as helping districts determine whether all schools have similarly high standards, whether highly mobile students have exposure to a common curriculum, and predicting which students are "off track" so they can intervene.

Section 6
Implications for District and State Decision Makers

We believe that a good interim assessment can be an integral part of a comprehensive assessment system, used in conjunction with classroom formative assessments and summative end-of-year assessments. However, policy makers need to carefully consider what they want to accomplish through interim assessments and whether the interim assessments and support for effective use that they put in place make it likely that they can meet their goals.

The policy brief has focused primarily on analysis and decisions that should be done by districts considering the use of interim assessments. In this section, we highlight three additional steps district decision-makers should consider in designing and using interim assessments. We then turn to the question of the role state policy makers can play in developing and implementing interim assessments.

A. Additional Implications for District Decision Makers

Three additional steps for district decision makers to consider in the design and implementation of an interim assessment system are: conducting a cost-benefit analysis, building capacity for effective formative assessment practices, and evaluating effectiveness. Each is considered below.

Conduct a Cost-Benefit Analysis

Any expenditure of resources (instructional time, teacher time, money, etc.) for an interim assessment system must provide valuable experiences and information that is not available on the state large scale assessment, on another district or school assessment, or in the classroom through daily instructional activities. For example, if the primary purpose of the assessment is evaluative, an end-of-year assessment may be sufficient. If the primary purpose is to inform instruction, it may be that investments in improving daily classroom assessment practices are a more effective strategy.

Districts may sometimes find it difficult to determine whether a proposed interim assessment provides experiences and information not otherwise available on another assessment. Large school districts often have many "legacy" assessments – a plethora of diagnostic, formative, interim and summative assessments as well as assessments for special populations put in place for a variety of schools for a variety of reasons over years. We suggest that districts conduct an assessment audit that examines what assessments exist in the district, the intended purpose of each, the data they are collecting and the utility of the data. Based on the results of the audit, districts may be able to significantly pare down their assessment burden and reclaim instructional time.

Another key component of the cost-benefit analysis is, of course, the up-front financial costs of purchasing and implementing the interim assessment system. Decision makers often look first at this financial cost when determining the cost-benefit relationship of investing in an assessment system. We fear that

the equation often tips in favor of the obvious up-front costs. For instance, it is cheaper to score multiple-choice items than constructed-response items or performance tasks, and it often costs less to buy a computer-based testing system than to invest in professional development for all teachers. We recognize the reality of constrained budgets, but argue that saving a few dollars on an assessment system might actually "cost" more in terms of opportunities for learning that may be lost as a result of cutting up-front purchase costs.

Build Capacity for Effective Formative Assessment Practices

We suspect that one reason school districts are investing in interim assessment systems that they hope will serve instructional purposes, rather than promoting formative classroom assessment, is that they lack the capacity to do formative assessment well at scale or they are hoping that using assessments will magically fix problems with incoherent curricula, lack of instructional expertise, or incoherent policies that address the core learning challenges of the district. There simply are not sufficient numbers of teachers and administrators with the requisite skills and knowledge and districts lack the professional culture, teacher networks, professional development resources and other supports needed to foster effective practices. As Black and Wiliam have noted, "…the improvement of formative assessment cannot be a simple matter. There is no quick fix that can alter existing practice by promising rapid rewards. On the contrary, if the substantial rewards promised by the research evidence are to be secured, each teacher must find his or her own ways of incorporating the lessons and ideas set out above into his or her own patterns of classroom work and into the cultural norms and expectations of a particular school community. This process is a relatively slow one and takes place through sustained programs of professional development and support." (Black and Wiliam, 1998)

Districts interested in improving formative assessment practices should design their *interim* assessment systems with explicit attention to increasing teachers' ability to do *formative* classroom assessment. The choice of item types, the format of reports and data analysis, and the structure and content of professional development can be carried out in ways that help teachers learn how to embed assessment within a learning activity, provide immediate corrective feedback, and modify instruction to meet students' needs. Over the long term, the focus of assessment effort in a district can move from interim assessment to the formative assessment practices that appear to have the most pay off for student learning.

Evaluate Effectiveness

Given the lack of research supporting the use of interim assessment and the many questions about the power and validity of different types of interim assessment, we suggest that decision makers deliberately and continuously evaluate the effectiveness of their interim assessment strategies within and across districts and adjust accordingly. This evaluation should include teacher surveys or focus groups to determine how the data were used and if there is evidence that the information gained from interim assessments improved student learning. Other important information includes the extent to which these assessments improve teacher understanding of the link between curriculum and assessment, the ease with which the tests are administered, scored, and interpreted, and the link between the interim and summative assessments.

B. State Role in Developing and Implementing Interim Assessments

Data from interim assessments are most often used to inform instruction at the classroom level or other educational decisions at the school or district level. They are typically not used for accountability or other policy purposes at the state level. So, what role do state educational leaders have in developing and implementing interim assessments?

State assessment leaders can provide invaluable help to districts by building policy-structures to support the use of comprehensive assessment systems including the appropriate use of interim assessments. The possibilities for state policies in support of effective interim assessment use are considerable. We limit our discussion here to the following which we feel are particularly important:

- Philosophical leadership—establishing and maintaining the vision
- State assessment design decisions
- Accountability policies
- Funding and other resources
- Professional development policies
- Quality control

Leadership: Establishing and maintaining the vision

State leaders must be clear and coherent as they articulate their vision for learning, instruction, assessment, and school quality and then work to ensure consistency among the various state initiatives and policies. Articulating support for comprehensive assessment systems on the one hand and throwing up policy roadblocks, even if inadvertent, to its effective use on the other hand will lead to fragmented implementation. This broad vision should support the meaningfulness of formative and interim assessments. For example, leaders who dwell on the large scale summative results as the only important measure of school quality can undermine the use of other assessment approaches.

Effectively using comprehensive assessments to improve student learning is the type of intervention that can take considerable time for successful implementation. Unfortunately, state policy makers tend to jump from initiative to initiative. Prioritizing and sustaining challenging interventions can be difficult in response to political pressures, but it is a critical role for state leaders to play. This requires that state policy makers develop a deep enough understanding of interim and formative assessment so they can successfully articulate support for their use.

State Assessment Design Decisions

All assessments in a standards-based environment start from the content standards. State leaders can insist on having high quality content standards that support the assessment of specific learning targets in ways that best support learning and instruction. Content standards developed according to the most up

to date learning theories—as almost none are—would better support assessments used to gauge students' progress on learning continuum. The states can then either play a direct role in developing curriculum or work with the districts in developing curriculum, aligned with the content standards.

The state assessment is one of the most visible programs run by any state department of education. It is particularly important for the influence it can have over formative and interim assessment practices. Certain types of large-scale assessment designs—such as those that provide clear focus on a limited number of meaningful outcomes and that use rich item formats—can better signal to the field about the types of learning outcomes and tasks that are valued. Many state leaders make decisions about large scale summative assessment designs based on key considerations of cost and time for both testing and report turnaround. These are certainly legitimate considerations, but leaders must be cognizant that school leaders will focus on the assessment results for which they will be held publicly accountable. Therefore, state policy leaders—if they are interested in supporting formative and interim assessment use—should work to ensure that the large-scale summative assessment is conceptually coherent with these other closer to the classroom assessments.

Accountability Policies

The types of inputs and outputs that get included in school and/or district accountability can have tremendous repercussions throughout the system. State policies—such as those found in Wyoming and Rhode Island —where local assessment results are used for graduation certification can direct attention and efforts toward improving local assessment practices. Of course, there is always a potential risk when using assessments designed to improve teaching and learning for accountability for corruptibility where the assessments would not serve either purpose well. Given this potential risk, at this time we recommend limiting the use of interim assessments in accountability systems to student-based systems (e.g., graduation) and not for school accountability purposes.

Funding and other resources

Money matters! Creating legislative funding priorities and requests is one of the most important aspects of state leaders' jobs. These requests operationalize the state's educational values and vision in very public ways. Beyond new money, state leaders can reallocate existing resources to support formative and other local assessment initiatives in ways that make clear their importance to the state's educational program. This leadership may include funding to support districts purchasing an interim assessment system or developing one in-house. It could be a direct flow-through to purchase a test, or more general funds that could also be used for professional development or other means of supporting local assessments.

Professional Development Policies

Although there has been little research on the efficacy of interim assessments in improving student learning, one finding that seems consistent is that the effectiveness of the test is dependent on how the teacher uses the information to give feedback to the students. Professional development linked to the use of the

interim assessment is a necessary component. Teachers need to understand not only how to administer the assessment and interpret the results but how to learn from the results and adjust instruction accordingly. The state can play a valuable role in providing professional development for teachers on adapting instruction based on information from an assessment thus turning results into actions. This is no small task. Providing this type of professional development support at the state level could maximize the efficiency of such an effort compared with having each individual school or district mount such a program. Professional development courses directly related to state standards could provide additional resources on teaching specific units. The state may also provide funding for teachers to receive training related to the specific assessment selected by their district. In the long run, the state k-12 department of education should work with teacher credentialing and teacher education institutions to make standards-based education and assessment an earlier and more salient part of teacher pre-service.

Quality Control

The state department of education can provide invaluable information about interim assessments to its districts. First, the department can vet potential vendors of interim assessments and provide information on the characteristics of the assessments available, quality of the items, and the degree of alignment with state curriculum. The state department may choose to allow state funds to be spent only on interim assessment systems that meet a specific qualifications or it may simply provide the information to the districts and allow them to use their own judgment. Secondly, the state department of education can serve as a warehouse for research studies on the utility of various interim assessments. This research could be about specific assessment systems, necessary features of any assessment, or best practices in using assessment systems. Thirdly, the state department of education can track which assessments various districts are using and create connections between districts with similar needs and goals. Providing these links will help districts learn from each other and replicate success without having to re-invent the wheel. Finally, the state DOE can evaluate the relative efficacy of the interim assessments used in the various districts in the state and share the results of such evaluations broadly.

Section 7
Conclusions

Our hope is that policy makers will take at least six points from their reading of this policy brief.

First, interim assessments, as we have defined them here, are distinct from formative assessment. While a definition may seem trivial, we believe that the many terms currently used to describe interim assessment (benchmark, periodic, predictive, formative) have impeded clear discussions and decisions about whether and how interim assessments should be used.

Second, the research supporting the efficacy of assessment to improve teaching and learning is based on formative assessment -- the minute-by-minute, classroom assessment that makes up the bottom layer of the triangle in Figure 1 on page 4 of this policy brief. There simply is no research base to support the claim that interim assessments improve student learning. While interim assessment has considerable intuitive appeal, experience shows that there is often a considerable gap between intuition and research evidence.

Third, we believe that there are useful and valid purposes for interim assessments within a comprehensive assessment system. However, in deciding whether interim assessment is an appropriate strategy, and more specifically, what interim assessment design is appropriate, policymakers must consider the following:

- What purpose (predictive, instructional or evaluative) are the interim assessments to serve? No assessment system can serve multiple purposes equally well.
- What is our theory of action about how interim assessments improve teaching and learning? That is, who will use the information, what changes in behavior are expected as a result of the information, and how teachers, administrators and students will be motivated and supported to do this work?

Fourth, policy makers should evaluate commercially available assessments cautiously. In particular, if policy makers desire interim assessments to serve instructional purposes, they should ask whether they meet the criteria suggested in this policy brief.

Fifth, as with any assessments, policy makers should ask at the outset whether the benefits of interim assessments outweigh their costs in terms of instructional time, teacher time and fiscal resources. Moreover, as interim assessments are put in place, they should be evaluated to determine the answers to this question.

Finally, policy makers should seek to eliminate the "zone of wishful thinking" in the design and implementation of interim assessment systems. Policymakers often hope that data will automatically lead to improved practice. However, experience shows that data must be accompanied by the reporting systems, professional development, support structures, and management practices that will impact teacher and student beliefs and behaviors. Each of these elements should be considered at the initial phases of designing or selecting and implementing an interim assessment system.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice. 5*(1), 7-74. Also summarized in an article entitled, Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan.* 80(2), 139-148.

Perie, M., Marion, S.F., Gong, B. (2007). *Moving Towards a Comprehensive Assessment System: A Framework for Considering Interim Assessments.* Dover, NH: The National Center for the Improvement of Educational Assessment, Inc. Available at: www.nciea.org.