Educational Measurement: Issues and Practice
Fall 2009, Vol. 28, No. 3, pp. 5–13

Moving Toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments

Marianne Perie, Scott Marion, and Brian Gong, *National Center for the Improvement of Educational Assessment, Inc.*

Local assessment systems are being marketed as formative, benchmark, predictive, and a host of other terms. Many so-called formative assessments are not at all similar to the types of assessments and strategies studied by Black and Wiliam (1998) but instead are interim assessments. In this article, we clarify the definition and uses of interim assessments and argue that they can be an important piece of a comprehensive assessment system that includes formative, interim, and summative assessments. Interim assessments are given on a larger scale than formative assessments, have less flexibility, and are aggregated to the school or district level to help inform policy. Interim assessments are driven by their purpose, which fall into the categories of instructional, evaluative, or predictive. Our intent is to provide a specific definition for these "interim assessments" and to develop a framework that district and state leaders can use to evaluate these systems for purchase or development. The discussion lays out some concerns with the current state of these assessments as well as hopes for future directions and suggestions for further research.

Keywords: comprehensive assessment system, formative assessment, interim assessment

First encoded in federal law as a result of the Improving America's Schools Act of 1994 (IASA), the standards-based reform movement has resulted in the widespread use of summative assessments designed to measure students' performance at specific points in time. Under IASA, testing was required at three grades: once each at the elementary, middle, and high school levels. The enactment of the No Child Left Behind Act (NCLB) of 2001 required increasing the number of these large-scale summative tests to every grade 3–8 and at least once in high school. Policymakers' goal for these assessments generally has been to measure students' attainment of the state content knowledge and skills against some defined level of performance, such as attaining the level of *Proficient* or *Distinguished* or simply *meeting the standard*. While many had hoped that these once-a-year tests would provide instructionally useful information, educators and others know this is not occurring. This is not because there is something "wrong" with these summative accountability tests; rather it is that they were not designed to meet instructional purposes. For example, these tests—by design—usually are administered as late in the year as possible and the results are returned after the students are home for the summer. In addition, the reports are designed to provide reliable total score and performance level information for each student across a wide range of content within a minimum of testing time, at low cost, under standardized conditions common to the whole state. This design precludes these general survey tests from providing useful diagnostic information for individual students. Therefore, educators and policymakers have realized that other forms of assessments are necessary to inform instruction during the school year.

This need for measuring student performance throughout the year has resulted in a rapid influx of products.

Marianne Perie is a Senior Associate at the National Center for the Improvement of Educational Assessment, P.O. Box 351, Dover, NH 03821; mperie@nciea.org. Scott Marion is the Associate Director, and Brian Gong is the Executive Director, of the National Center for the Improvement of Educational Assessment.

Many vendors are marketing assessments to states and districts that they call "benchmark," "diagnostic," "formative," and/or "predictive" with promises of improving student performance and helping schools and districts meet the federal NCLB requirements or increasing pass rates on high school exit exams. All of these terms fit under the umbrella term "interim assessment." A good interim assessment can be an integral part of a state's or district's comprehensive assessment system used in conjunction with classroom formative assessments and summative end-of-year assessments. Unfortunately, there is little research indicating that many of these commercially available interim assessments positively affect student achievement. Furthermore, vendors for many of these products cite research on classroom formative assessment (e.g., Black & Wiliam, 1998) implying that their assessments will improve student learning even though few, if any, of these commercial products are the types of assessments or activities described in the Black and Wiliam (1998) meta-analysis. There is a growing concern among researchers and educators that states and districts are buying assessment systems that promise to provide information to improve learning without fully examining the validity of these claims.

The focus of this article is two-fold. First, we define interim assessments, distinguish them from formative assessments, and focus on their uses. Second, we provide a framework for evaluating these interim assessments to help state and district leaders thoughtfully examine the commercially available products, develop strong specifications for a customized system, or develop their own interim assessments. A final purpose of this article is to promote interest in further research in this area, and to that end, we conclude with a section describing our vision for this research.

Throughout this article, our discussion will focus on how interim assessments fit into the comprehensive system and what unique value, if any, interim assessments serve. We attempt to describe the characteristics of effective interim assessments, discuss the different purposes these assessments may serve, provide information on how to choose the best type of assessment for a given situation, and then offer guidance on evaluating the products that already exist in the marketplace. Although we believe that there are some organizations trying to sell item banks and reporting systems as interim assessments without thoughtfully integrating them into a state assessment system, our goal is not to condemn all currently available products, but rather to provide a framework for the consumer to use, in evaluating them.

Distinguishing Among Assessment Types

Before we can begin a thoughtful discussion on interim assessments, we need to agree on definitions. We prefer the schema that places assessments into three categories—summative, interim, and formative—and distinguishes among the three types based on the intended purposes, audience, and use of the information. Summative assessments are given one time at the end of the semester or school year to evaluate students' performance against a defined set of content standards. These assessments are typically given statewide (but can be national or district) and are usually used as part of an accountability program or to otherwise inform policy. They could also be teacher-administered end-of-unit or end-of-semester tests that are used solely for grading purposes. They are the least flexible of the assessments.

Skipping to the narrowest type, formative assessment is used by classroom teachers to diagnose 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. It can be a five-second assessment and is 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. Black and Wiliam (1998) defined formative assessment as just one part of formative instruction. In their seminal piece, *Inside the Black Box,* they argue that formative assessment cannot stand alone but must be a part of a whole system that uses the information from the assessment to adapt teaching to meet the learner's needs. 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.

Finally, interim assessments are considered medium-scale, medium-cycle assessments, falling between summative and formative assessments and usually administered at the school or district level. Typically, interim assessments are given several times a year, although a test that was administered once at some midpoint during the year could also be considered interim. While the results may be used at the teacher or student level, the information is designed to be aggregated beyond the classroom level, such as the school or district level. That is, they may be given at the classroom level to provide information for the teacher, but a crucial distinction is that these results 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, another critical feature separating these tests from formative assessments.

Although many others have used the term "interim assessment" to describe benchmark, diagnostic, predictive, and even some formative assessments, we offer the following definition:

> Assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level. The specific interim assessment designs are driven by the purposes and intended uses, but the results of any interim assessment must be reported in a manner allowing aggregation across students, occasions, or concepts.

By this definition, end-of-chapter tests available in most textbooks could be considered interim, if they are designed to be used to inform instructional decisions and reported in the aggregate. Teacher-created tests given at the end of a unit could be interim, formative, or summative, again depending on their purpose and design. The key components of the definition are that 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

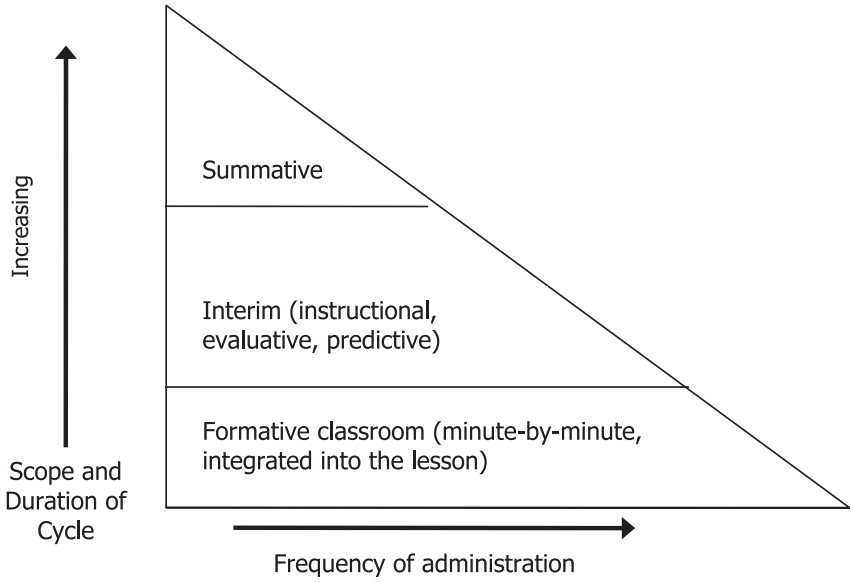

FIGURE 1. Tiers of assessment.

both at the classroom and beyond the classroom level, such as the school or district level. If the test is used simply for grading purposes, it is summative, while if it is used solely for the purpose of informing the teacher of a student's progress, it is most likely formative. 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 student's learning. It is these purposes that determine the necessary features of the assessments.

These three tiers of assessment—summative, interim, and formative—are shown in Figure 1. The triangle illustrates that formative assessments are used most frequently and have the smallest scope (i.e., the narrowest curricular focus) and the shortest cycle (i.e., the shortest time frame, typically defined as 5 seconds to 1 day), while summative assessments are administered least frequently and have the largest scope and cycle. Interim assessments fall between these other two types on all dimensions.

Overview of Interim Assessments

We encourage the reader to think broadly about the possible forms of interim assessments, from commercially purchased, computer-based sets of multiple-choice items to more locally created sets of extended performance tasks administered commonly throughout a school, district, or state. We do not intend to tout one type of interim assessment as being the best—although we argue that some are clearly superior for improving learning than others—but to encourage users to be explicit about the desired purpose of the assessment and then find the assessment that best fits that purpose. For example, an interim assessment may be given in order to

(1) Evaluate how well the student has learned the material taught to date.
(2) Predict students' performance on a summative assessment.
(3) Determine whether one pedagogical approach is more effective in teaching the material than another.
(4) Provide aggregate information on student achievement at a district level.
(5) Provide specific feedback on where the gaps in a particular student's knowledge are at the classroom level.
(6) Determine whether students are on track to succeed on the summative assessment.
(7) Diagnose and provide corrective feedback to help a group of students get on track to succeed on the summative assessment.
(8) Motivate and provide feedback to students about their learning.
(9) Provide information to help the instructor better teach the next group of students by evaluating the instruction, curriculum, and pedagogy.
(10) Ensure that teachers are staying on track in terms of teaching the curriculum in a timely manner (i.e., pacing).
(11) Provide a more thorough analysis of the depth of students' understanding.
(12) Determine whether students are prepared to move on to the next instructional unit.

Summarizing this large list brought us to three general classes of purposes for interim assessments: instructional, evaluative, and predictive. Although this categorization is not perfect, it seems to capture the essence of most of the goals of using an interim assessment system. We recognize that many assessments are not designed to serve only a single purpose, but we argue that 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, an important additional step is to check not only whether the assessment is being used for its intended purposes, but to check the quality with which it meets those purposes.

Instructional Purposes

The primary goal of an interim assessment designed to serve instructional purposes is to adapt instruction and curriculum to better meet student needs. Of the three purposes, this one aligns most closely with the previous definitions of formative assessment. That is, the results of these assessments are used to adjust instruction with the intent of helping the students assessed meet the learning goals. However, the testing and reporting time frame of these interim assessments is typically medium cycle, whereas classroom formative assessments tend to operate on shorter cycles.

Subsumed under this purpose are other types of assessment that certainly would not meet the definition of *formative* presented earlier, but are instructional nonetheless. Consider, for example, features included in many commercially available systems. A typical system contains a bank of items nominally aligned with the state curriculum that teachers can use to create a test to evaluate student learning on the concepts taught to date. Results

may be reported immediately, and data are disaggregated by content standard allowing teachers to identify strengths and weaknesses in the students' learning. This type of interim assessment might be labeled formative, but we would argue that to be truly formative it must be timed appropriately for adjustments to instruction to occur, be aligned with specific local curriculum, provide more in-depth analyses of student misconceptions or lack of understanding, lead to strategies for improving instruction, and lead the teacher to modify instruction. Nevertheless, this type of assessment falls under the instructional category.

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. 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. Whether that professional development support is or should be included in the instructional interim assessment package will be debated among policy makers, developers, and educators. We feel strongly that no matter what the source of professional development, an assessment system purchased for instructional purposes will be effective only when used by educators who have the knowledge and tools to use the assessments and the results appropriately. Ideally, we believe that promoting informed use would be supported by development and training by both the developer and the user.

Evaluative Purposes

Another type of purpose an interim assessment might serve is to provide evaluative information about the curriculum or instruction. Think of this as a programmatic assessment designed to change instruction not necessarily in mid term but over the years. The students benefiting from the information gleaned from these assessments would not necessarily be the students assessed, but the students receiving the instruction in the future. Many had hoped that summative end-of-year assessments would fulfill this purpose, and in many cases these end-of-year tests have provided useful evaluative data, but most are too short and designed to cover too much content to provide the depth of information required for most evaluative purposes.

District-level policymakers are often interested in interim assessment systems for reasons other than to inform modifications to instruction. For instance, their goals may be to enforce some minimal quality through standardization of curriculum and pacing guides, to centralize coordination for highly mobile urban student populations and high teacher turnover, or as a lever to overcome differences in learning expectations and grading standards. These types of purposes are evaluative in nature.

Assessments used for evaluative purposes could be given district wide to compare the effectiveness of various instructional programs for improving student learning. Consider, for example, a district that is experimenting with more than one reform program or pedagogical strategy across different schools. The use of interim assessments in this context could be an effective way of monitoring the relative efficacy of each program. Similarly, assessments could be given at various points throughout the year to measure growth—not with the intention of intervening but for evaluating the effectiveness of a program, strategy, or teacher.

The assessments could also be used on a smaller scale, providing information on which concepts the students understood well and which were less clear. Teachers within one or more schools could use this information with the goal of helping them modify the curriculum and instructional strategies for future years. Other purposes could be 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. Of course, teachers can and should always learn from their experience. Any instructional interventions that could improve instruction in a current year should be implemented.

In our definition, an *evaluative* assessment would be designed explicitly 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. 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 the assessment. A matrix sample could be used to maximize the information while minimizing the time spent on assessments in the classroom.

Predictive Purposes

Predictive assessments are designed to determine each student's likelihood of meeting some criterion score on the end-of-year tests. Predictive purposes of interim assessments are important to many users and this interest could increase as the annual NCLB targets continue to rise. In addition, assessments in this category could be used to predict performance on a high school exit exam or success with postsecondary curriculum. Although predictive purposes are important in high-stakes testing situations, we suspect that there are few assessment systems where the sole purpose for the system is prediction. Rather, most users want additional information to help them improve the performance of students for whom failure is predicted. This additional information might come from the assessment itself or from further probes to determine areas of weakness in those not on track to succeed. This scenario could be an example of how interim and formative assessments work together to help improve student performance on a summative assessment. It also highlights the importance of aligning all components of a comprehensive assessment system.

A 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 on track to perform at the basic level, and then appropriate interventions are used to bring the student to proficient, 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 questions should be raised if too many students predicted to pass the summative test actually fail it.

Identifying the Goal

As policymakers decide to bring an interim assessment system to their state/district/school we encourage them to have a theory of action for how the particular assessment system will work in the teaching-learning cycle. Policymakers and educators using assessments need to understand the limitations of any assessment for fulfilling particular purposes. As a start, we think it will be helpful for educational leaders to address the following questions:

(1) What do I 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 appropriately?
(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 were not used?

The answers to these questions 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.[1] Importantly, these questions and the associated answers serve as the beginning of a validity argument to support (or refute) the particular assessment system.

Answering these questions also may suggest that it might be appropriate to consider primary and secondary purposes in designing or choosing an interim assessment system. For instance, while the primary purpose of giving an interim assessment may be evaluative, we would hope that given the results for a specific set of current students, teachers and school leaders would attempt to provide remediation programs for those students not understanding key concepts. Similarly, even when the primary purpose of an interim assessment is to predict success on the end-of-year assessment, a policymaker may also want the predictive assessment to provide some diagnostic information so that educators can intervene with students predicted to score below a critical level. Of course, the assessment may only fulfill secondary purposes if certain factors associated with a primary purpose—such as having a very short test—do not overly constrain other uses.

Finally, the answers to the above questions should help policymakers to determine whether the best approach is to adopt a currently existing system or to build their own. There are many vendors currently selling interim assessments under various labels. These assessments are marketed to serve a plethora of purposes, including serving as a diagnostic tool, 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, test assembly supports, administration tools, and customized reports. These systems often are computer- 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,[2] and results typically are reported in terms of number correct or as scale score developed by the publisher. The "diagnostic" portion tends to be a summary of results by content standard. Often, these systems provide a variety of options for reports, with different levels of aggregation.

Other states and districts have experimented with developing in-house local assessments. These tend to be computer-based systems that include teacher-developed items linked directly to instructional units. They give quick feedback to teachers and produce in-depth reports at the student and classroom levels. It seems that most of these systems have been developed for instructional purposes rather than as predictive or evaluative.

There is no one-size-fits-all assessment, only a best design for a desired use and existing constraints and resources. We believe that many educational leaders consider a cost-benefit relationship before investing in such a system, but we fear that the equation often tips in favor of low costs and short testing time. 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.

Characteristics of an Effective Interim Assessment System

This section of the article is intended to help educational leaders either choose or develop a strong interim assessment system for their schools. We provide evaluative criteria to help policymakers critically appraise their local assessments and also provide suggestions for the type of validity evidence to collect over time. 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 needs.

Again, we emphasize that the purpose must be clearly stated before one can truly determine or evaluate the necessary characteristics of the assessment. Consideration should be given to all parts of the interim assessment, including item quality, administration requirements, and reporting elements. This last piece is important because the report is the mechanism for translating the assessment data into decisions, which then translate into action and should be one of the first considerations in designing a new assessment. It serves to transform raw data into results that can be interpreted meaningfully and acted upon appropriately. Time should be spent discussing the question: what do we want the tests to tell us? Assessments serving an instructional purpose will have different features in their reports than those serving predictive or evaluative purposes.

Evaluative Criteria

To help guide the evaluation of commercially available interim tests and the development of custom interim assessment systems, we have provided the following criteria for states and/or districts to consider prior to purchasing or developing an interim assessment system. We find that most, if not all, of these criteria fit under Standard 15.8 of the *Standards for Educational and Psychological Testing*

(American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999):

> When it is clearly stated or implied that a recommended test use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence.

Following our argument that the interim assessment design must be linked to the purposes and intended uses, we present evaluation criteria for the three major purposes articulated earlier: instructional, evaluative, and predictive. To avoid redundancy, we first present several general criteria that cut across all three purposes.

General

(1) A test can be no better than the quality of the items it contains. Therefore, the quality of the items needs to be evaluated against professional standards and expert opinion. The types of items/tasks may vary depending on the specific purposes and intended uses, but all should be of high quality as documented through traditional reviews for content and bias and sensitivity as well as pilot testing and data reviews.
(2) Alignment evidence should be provided to document the relationship of the items and sets of items in a test "form" to the knowledge and skills (including depth of knowledge) called for in the target content standards.
(3) The inferences resulting from the test scores should be validated for the intended uses and purposes.
(4) The test publisher must include clear guidelines regarding the appropriate uses of the assessment results, as well as indicating either potentially inappropriate uses of the results or limitations of the validity evidence.
(5) Tasks should be applicable to the target student populations; in most schools and districts these may include English language learners and students with disabilities.
(6) There should be evidence that the professional development associated with the assessment system facilitates educators' appropriate interpretation and use of the assessment results for the specified purposes. Clearly, assessments serving instructional purposes will require different professional development than is required for evaluative and predictive purposes, and the audiences (teachers, building administrators, district leaders) may differ for each.
(7) 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 performance tasks embedded in instruction, the issue of "testing time" is less critical.

Instructional

(1) To the extent possible, interim assessments for instructional purposes should fit as seamlessly with instruction as possible and represent an opportunity for student learning during the assessment experience.
(2) Ideally, the system should provide evidence, based on scientifically rigorous studies, demonstrating that the assessment system has contributed to improved student learning in settings similar to those in which it will be used.
(3) There should be evidence that the results of the assessment and the associated score reports have been designed to facilitate meaningful and useful instructional interpretations.
(4) Clear guidelines should be provided explaining how the results of the assessment, including the results of particular tasks/items or sets of items, should be used to help inform instructional decisions.
(5) Each particular assessment in the system must link closely to the curricular goals taught prior to the assessment administration, preferably quite proximal to the assessment event. The assessment should include only content and skills for which the students have had a legitimate opportunity to learn, unless the purpose of the assessment is as a pretest to determine readiness for some learning in the near future or as a placement test.
(6) To best serve instructional purposes, each interim assessment should assess only a limited number of important curricular goals to make it more likely that instructional adjustments can be timely and targeted appropriately.
(7) In general, to serve instructional purposes interim assessments intended to support diagnosis of students' understanding and misconceptions should include high-quality open-ended tasks. All items, whether open ended or multiple choice, should be developed so that useful information about students' understanding and cognition can be gleaned from specific incorrect answers.
(8) Instructional interim assessments should measure instructional and curricular goals, provide information not easily gleaned from the state's large scale assessment such as more in-depth understanding demonstrated through extended tasks or synthesis works.

Evaluative

(1) The collection of tasks administered through the year should represent a technically sound range of difficulty and appropriate breadth, dependent on the focus of the evaluation.
(2) The assessments should comprise items and tasks with a mix of formats to provide users a deep understanding of the relative effectiveness of educational programs.
(3) The assessment must be targeted to the content standards that are the focus of the educational program(s) being evaluated or studied and/or to the expected domain of transfer.
(4) The reports must be designed to facilitate the intended evaluation and accurately portray the error associated with the scores and subscores.

Predictive

(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. However, in order to justify the additional testing and cost, the predictive assessment should be significantly more related to the criterion measure than other measures (e.g., teachers' grades) that could be used.
(2) The predictive assessment should comprise items with a similar mix of item types as the criterion measure.
(3) The predictive assessment should be designed from the same or similar blueprint as the criterion measure.

(4) The reports should be designed to facilitate the intended predictions, including an honest and accurate characterization of the error associated with the prediction, both at the total score and subscore levels.
(5) If the purpose of the assessment goes beyond solely predicting performance to identifying areas of weakness, 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.

We are not suggesting that interim assessment systems must meet all the criteria listed above before being purchased for a district or state, but we recommend that educational leaders consider the criteria when evaluating which, if any, system to purchase or when evaluating a proposal to create a customized system.

Validity Evidence

Approaching this from a validity perspective, we argue that the interim assessment system should be validated for the specific purposes and uses. Validity evidence would include:

(1) A clearly articulated goal or target. An interim assessment serving an instructional purpose, for example, must include a rich representation of the content standards students are expected to master.
(2) High-quality items that elicit and assess what is intended. Items should be directly linked to the content standards and specific teaching units.
(3) Useful and clear interpretations to support the intended uses.
(4) Operational feasibility and low negative unintended consequences. A predictive interim assessment should minimize the loss of instructional time.

Additionally, any provider should be required to provide evidence of the validity of the system for the intended purposes. Once the system has been implemented, the sponsor—whether districts and/or states—should periodically evaluate the system to ensure that it is meeting intended purposes and uses. While any evaluation will have to be tailored to the specific purposes and uses, we offer the following general suggestions for exploring the validity of an interim assessment system:

(1) If the test is used for instructional purposes, follow up with teachers to determine how the data were used, if they provided useful information, and whether there was evidence of improved student learning, including evidence of generalizability and transfer, for current students.
(2) If the test is used for evaluative purposes, gather data from other sources to triangulate results of interim assessment and follow up to monitor if evaluation decisions, such as changes to curriculum and/or instruction, are supported.
(3) If the assessments are used for either instructional or evaluative purposes, look for evidence of increases in teacher knowledge of content, pedagogy, and student learning.
(4) If the test is used for predictive purposes, do a follow-up study to determine that the predictive link is reasonably accurate, provides more predictive power than information such as grades and teacher judgments, and that the use of the test contributes to improving criterion (e.g., end-of-year scores).
(5) Regardless of the purpose of the assessments, the manageability, including the quality of implementation, should be monitored.
(6) Finally, any unintended negative consequences should be monitored for all interim assessments including any adverse effects on student motivation as a result of engaging with the tasks, a narrowing of the curriculum, or a decreased focus on formative assessment.

Matching the Purpose with the Assessment

The main impetus for this article was to provide advice on how to evaluate the suitability of commercially available or locally created products for states and districts considering implementing some sort of interim assessment system. We have continued to emphasize the need to articulate the purpose(s) of such a system.

We recognize that in most instantiations of interim assessment educational leaders are trying to squeeze as many purposes as possible out of a single system. Unfortunately, one of the truisms in educational measurement is that when an assessment system purports to fulfill too many purposes—especially disparate purposes—it rarely fulfills any purpose well.[3] This does not mean that certain interim assessment systems cannot fulfill more than one purpose, depending on the level addressed by the primary 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 likely cannot be aggregated to the subgroup, school, and/or district level to provide evaluative 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 to serve instructional or even evaluative purposes. Therefore, users should design a system that will adequately fulfill the more important and finest grain purpose first and then consider whether additional purposes can be fulfilled well within the same assessment, or whether it would be more appropriate to use multiple assessments—including formative assessment—within a comprehensive system.

We recommend that educational leaders considering purchasing a commercially available system follow the advice offered in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), specifically in Standard 11.1:

> Prior to the adoption and use of a published test, the test user should study and evaluate the materials provided by the test developer. Of particular importance are those that summarize the test's purposes, specify the procedures for test administration, define the intended populations of test takers, and discuss the score interpretations for which validity or reliability data are available.

Future Areas of Research Needed

Clearly, this field is rich for further research. New studies funded by the U.S. Department of Education's Institute of Education Sciences are exploring areas that may inform the field of formative uses of assessment. Many of these studies focus on interim assessments, sometimes as part of a tutoring session or computer-based learning. In general, they examine how testing a particular unit of instruction relates to retention of information after an extended period of time. One common finding across studies was that student performance

on the "repeated testing" was not nearly as important as the corrective feedback they received as a result. That is, a student who guessed incorrectly on an item on a unit test, but who received good corrective feedback, was just as likely to answer a similar item correctly on a future test as a student who had answered it correctly the first time. Another common finding we found interesting was that the repeated testing, in and of itself, contributed to retention. And this was particularly true when the short tests required students to generate their own responses on short-answer items (Viadero, 2006). We look forward to seeing the results of these studies when published.

We feel it is important to continue to examine how the use of interim assessments can help further student learning. Education leaders can find themselves in a difficult position if they do not want to adopt a test without validity evidence, while there is little validity evidence available. So, the first area we see the need for strong research efforts is in validating the use of these types of assessment. In general, we see the need for research in the following areas:

(1) Score-based inferences from interim assessments need to be validated for the use of improving performance on summative assessment and gather evidence to evaluate this argument. Choose several types of interim assessments and validate their uses.
(a) Are predictive assessments truly predicting student performance on end-of-year assessments more so than other readily available data? Of course, the results of this question could be confounded by the use of appropriate interventions, but those interventions may provide evidence of the validity of the consequences.
(b) Are instructional assessments actually improving instruction? Are there any unintended consequences?
(c) Are evaluative interim assessments effectively identifying differences in various pedagogies or instructional approaches? What characteristics make them more useful?
(2) Studies are needed to examine differential effects of interim assessments on students' intrinsic motivation to learn. Consider the concern that frequent assessments may diminish intrinsic motivation by shifting the effort and purpose from learning "to know" to learning "to display one's knowledge" (Lave & Wenger, 1991). How can we use the interim assessments constructively to further students' desire to learn rather than to further their desire for a high score?
(3) Kluger and DeNisi (1996) and others found that normative types of feedback or feedback that focuses on the person rather than on the task can actually have a negative effect on student performance. Their research showed that the most effective types of feedback were ones in which students were told not only what they needed to learn but how to get there. How does this research apply to the interpretation of results from interim assessment?
(4) It has been argued that evidence collected for summative purposes can rarely be disaggregated to support learning, but evidence collected for formative purposes can be aggregated to support summative inferences (Wiliam, 2006). However, we need to learn more about how to aggregate results of formative assessments before pursuing this path. What are the requirements for building a system that provides teachers the information they need but can still be scaled to compare results across students, teachers, and/or schools?
(5) What are the effective strategies for implementing interim assessments and presenting results so that teachers use the data appropriately for making effective educational decisions?
(6) What types of professional development are necessary to influence effective use of interim assessments and what factors (e.g., teacher qualifications) interact with various professional development models? What approaches are most effective for providing this type of professional development on a large scale?

There are a host of other lines of inquiry areas that one might pursue to build the research base on interim assessment, but we think that the ones listed above are an important starting point.

Discussion

We first approached this article from the perspective of investigating "formative" and "benchmark" assessments being used at the district and state levels. Because many assessments now in the field are marketed under the appropriated term "formative assessment," we realized that there needed to be a discussion regarding the current types of assessments being sold for formative purposes. Then, we turned to developing a framework to better understand interim assessments: how they are used and why they are proliferating at such a rapid rate. Furthermore, we were interested in the role that state and district leaders play in selecting/developing these assessments and how we might be able to help these leaders with this task. When asked why we chose to focus on interim assessments rather than the purer and research-based formative assessments, our answer was simple: states and districts are spending considerable resources to implement such systems.

We recognize the difficulty of developing, at a state level, strong formative assessment strategies as advocated by Black, Wiliam, Shepard, and others. Components such as weaving the assessment seamlessly into the curriculum and providing useful feedback that leads to appropriate modifications in instruction is difficult when the agent (state department of education personnel) is several steps removed from the classroom. While states can support professional development programs that help educators develop and use such tools, they could also help by purchasing a preexisting system, if such a system supports formative and professional learning needs. In addition, states may have other requirements for an assessment program, such as developing an early warning system to identify students who are not on track to succeed in order to help with additional supports. Or, the states may wish to use these interim assessments as evaluation tools for different schools, instructional programs, or pedagogies. That is why we chose to define interim assessments, focused on specific purposes and uses, as tools to evaluate students' knowledge and skills that are designed to inform decisions at the classroom level and above.

That said we are concerned that many of the commercially available systems are quite different from what the research currently supports, and those selling such systems promise far more than they can deliver. For example, these systems often lay claims

to the research documenting the powerful effects of formative assessment on student learning when it is clear that Black and Wiliam's (1998) meta-analysis evaluated studies with formative assessments of very different character than essentially all currently available commercial interim assessment programs.

We believe it is not worth spending scarce resources on interim assessments that simply administer a series of minisummative assessments. The assessments should be linked to specific instructional units to provide teachers with useful information. While pre- and posttest designs may be useful for some purposes, testing students on material they have not yet learned rarely provides teachers with helpful information. We have seen systems where shorter versions of the end-of-year assessment are given periodically throughout the school year. The items on these assessments are placed on the same scale as the items on the end-of-year assessment, so the results can be used to show progress toward the goal. A cursory examination of several of these systems revealed that they do not meet the criteria discussed in this article and suffer from such technical and content shortcomings that we believe they are a poor use of money and instructional time.

A good interim assessment can be an integral part of a state's or district's comprehensive assessment system, used in conjunction with classroom formative assessments and summative end-of-year assessments. As such, we believe that there are valid purposes for giving interim assessments beyond informing instruction at that point in time. However, the policymakers and educators using the assessment need to understand the purpose of the assessment and what it can and cannot do. If policymakers want an assessment to help educators improve instruction, they should look for one that ties directly to the classroom instruction and provides in-depth examination of not just which items students miss but why they miss them. Actually, if this is the sole goal of the assessment, we argue that resources would be better spent helping teachers learn formative assessment techniques, including using the information to intervene with students who do not yet understand key concepts.[4] If policymakers want an assessment to tell them how students are likely to perform on an end-of-year assessment, they need to examine the reliability of the predictions and the information describing what to do next.

At a minimum, we argue that any expenditure of resources (teacher time, money, etc.) for an interim assessment system must provide experiences and information that are not available on the state large-scale assessment or in the classroom through daily instructional activities, including formative assessment. Finally, any of these assessment types need to provide evidence of their validity. Are they demonstrating their intended positive consequences and are there any unintended negative consequences of their use? For instance, do additional assessments solidify a student's understanding of a concept or inure him to tests in general? Such validity evidence should be examined prior to adoption of the assessment program and should also be generated for the specific populations and context of the state's or district's program. These interim assessments can be an integral part of any comprehensive assessment system and should be considered as a piece of a whole and evaluated as such.

Notes

[1]For more information about these types of design decisions within the context of interim assessments, please see Perie, Marion, and Gong, 2007.

[2]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.

[3]This should also be a red flag to any educational leader considering purchasing a system that promises to fulfill many purposes or to solve all educational problems.

[4]However, this assumes that the curriculum is sound; our experience has been that often considerable attention needs to be paid to the curriculum before fine tuning any instruction through formative assessment.

Acknowledgment

The authors wish to acknowledge the work of the Council of Chief State School Officers' formative assessment working group (FAST SCASS) that helped us to clarify the distinctions between formative and interim assessment. In particular, we received influential feedback from Jim Popham, Margaret Heritage, and Fritz Mosher. We also would like to thank others that reviewed earlier drafts including Bob Linn, Sue Brookhart, and three anonymous reviewers. Finally, we wish to acknowledge our colleagues at the Center who focused our thinking on the definition and use of interim assessments, specifically Rich Hill, Charlie DePascale, Karin Hess, and Jennifer Dunn.

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.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

No Child Left Behind Act of 2001, Pub. L. No.107–110, 115 Stat.1425 (2002).

Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments*. National Center for the Improvement of Educational Assessment. Dover, NH: NCIEA. Available at www.nciea.org.

Viadero, Debra. (2006). Cognition studies offer insights on academic tactics: U.S.-funded projects eye ways of helping students remember more material. *Education Week, 26*(1), 12–13.

Wiliam, D. (2006). Assessment for learning: Why, what and how. *Orbit: OISE/UT's Magazine for Schools*, *36*(2), 2–6.