A Framework for Considering Interim Assessments

Marianne Perie

Scott Marion

Brian Gong

National Center for the Improvement of Educational Assessment

February 13, 2007

Previous versions of this paper were presented at the 2006 Reidy Interactive Lecture Series (RILS) and to various CCSSO-sponsored State Collaborative on Assessment and Student Standards (SCASS) groups.

A Framework for Considering Interim Assessments

Marianne Perie, Scott Marion, and Brian Gong

The standards-based reform movement, first encoded in federal law as a result of the Improving America's Schools Act of 1994 (IASA), has resulted in the wide-spread 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 a significant increase in the prevalence of these large-scale summative tests. Policymakers' goal for most of these assessments is to measure students' knowledge and skills against some level of desired performance, such as attaining the level of *Proficient* or *Distinguished* or simply *meeting the standard*. While many have hoped that these end-of-year tests would provide instructionally useful information for educators, educators and others know this is not occurring. This is not because there is something "wrong" with these summative accountability tests, rather 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, by no fault of the assessment vendors, are returned after the students are home for the summer. In addition, the reports often provide only total score and performance level information for each student. Therefore, educators and policymakers have realized that other forms of assessments are necessary to provide information to inform instruction during the school year. Educators want to measure student progress toward important end-of-year goals and to receive sufficient information to determine what steps can be taken to further students' learning and achieve these goals.

This need for measuring student performance throughout the year has resulted in a rapidly growing influx of products in the field. Large numbers of vendors are selling assessments to states and districts that they call "benchmark," "diagnostic," "formative," and/or "predictive" with promises of improving student performance and helping schools meet their goals of showing adequate yearly progress or increasing pass rates on high school exit exams. A good district-level assessment can be an integral part of a state's comprehensive assessment system, used in conjunction with classroom formative assessments and summative end-of-year assessments. Yet, there is little research that these commercially-available assessments positively affect student achievement. In fact, many of these products cite the research on classroom formative assessment to indicate that their assessments improve student learning. However, few, if any, of these commercial products are the types of products or activities described in the Black and Wiliam (1998) analysis—the research most commonly cited. There is a growing concern among researchers that states and districts are buying assessment systems that promise to provide information to improve learning without fully examining the validity of these claims.

Policymakers and educators using assessments need to understand the purpose and limitations of any assessment. Is the purpose of the assessment to predict how students are likely to perform on an end-of-year assessment, to diagnose gaps in learning, to indicate the extent of student mastery of specific content and skills, or to evaluate a particular program or pedagogy? Because these assessments cost money and instructional time, they must provide experiences and information that are not available on the state large-scale assessment. At the same time, they should provide information that can be aggregated across students, occasions, or concepts

to provide information to those outside of the classroom while still aligned with information gathered through formative assessments within the classroom. Assessments that fill the gap between classroom formative assessments and state summative assessments are an integral part of any comprehensive assessment system and should be evaluated as such.

The purpose of this paper is to provide a framework for evaluating these mid-level assessments, which we call interim assessments, to help state and district leaders thoughtfully examine the commercially-available products or develop strong specifications for a customized system. This is a very large field and we are focusing primarily on those products that are currently being marketed for use by schools and districts, but we intend for the guidelines provided here to be useful for states or districts who want to develop their own interim assessments.

We began this work with discussions with state assessment leaders, assessment researchers, and others and found there was a consistent call for definitions of these different types of assessments—formative, interim, etc. Definitions, by their very nature, encourage analysis of each word and phrase within the definition, so creating a definition for interim assessment was not an easy task, particularly because we believe the definition for interim assessments should be driven by their purpose. We offer a definition in the next section, but throughout the paper our primary focus is on how these assessments are used. As such, we argue that the actual definition is always contextualized within specific purposes and uses. Consider, for example, how an assessment can be used summatively (to evaluate learning at the end of a unit), formatively (to inform subsequent instruction), or predictively (as an early warning system for future performance), to name a few purposes. We believe the purpose should be the main driver in defining and evaluating an assessment. Our discussion will focus on how the interim assessment fits into the comprehensive system and what unique purpose(s) the interim assessment serves. Therefore, after attempting to define these assessments and provide a classification system to consider their uses, we examine 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.

Finally, we believe that there can be a difference between formative assessment and assessments used for formative purposes. Although we are strong proponents of the use of good formative assessment strategies, we also argue that interim assessments can serve formative purposes without meeting all the requirements for formative assessment. Furthermore, we recognize that not all interim assessments are designed to serve formative purposes and that other purposes may be legitimate, depending on the user's needs. Our goal is to provide a framework to help state and district-level consumers critically evaluate these systems.

Distinguishing Among Assessment Types

Before we can begin a thoughtful discussion on interim assessments, we need to agree on some definitions for the various names and types of assessments being used with the promise of improving student learning. The simple definition most first-year graduate education courses teach is that there are two types of assessments: summative, which are given at the end of instruction to provide information on what was learned, and formative, which are given at the beginning or in the middle of instruction to provide information about what the student knows

or doesn't know relative to what s/he should know at that point. However, this distinction of naming an assessment based on whether it's given at the beginning, in the middle, or at the end of instruction has led to many assessments being called formative even when they serve purposes that have little to do with providing useful information to teachers or students on improving student learning. In actuality, this distinction of formative and summative was initially used in the field of program evaluation with the express purpose of formative evaluation to provide mid-cycle corrections while summative evaluation was used to determine whether a program's results matched its goals (Scriven, 1967).

Our schema places assessments into three categories—summative, interim, and formative—and distinguishes among the three types based on the intended purposes, audience, and use for the information, not simply as a result of when the assessment is given. 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 usually given statewide (but can be national or district) and are often used as part of an accountability program or to otherwise inform policy. They are the least flexible of the assessments.

Interim assessments may be administered on a smaller scale, typically school- or district-wide. While the results may be used at the teacher or student level, the information is designed to be aggregated at a level beyond the classroom level, such as to 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. 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.

The final type, formative assessment, is one given in the classroom by the teacher 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. It can be a five-second assessment and is often called "minute-by-minute" assessment or formative instruction. 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.

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 one hour), while summative assessments are administered most frequently and have the largest scope and cycle. Interim assessments fall between these other two types on all dimensions.

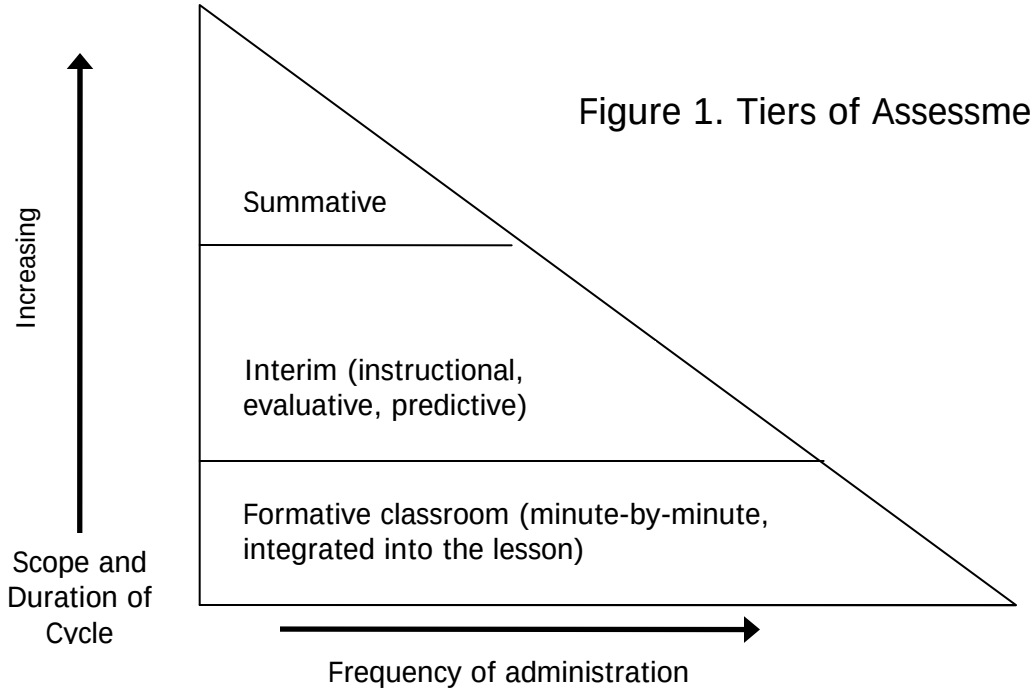


Figure 1. Tiers of Assessment

Defining Formative and Interim Assessments

Because many of the commercial products available are marketed as "formative assessment" we feel it is important to go further in the definition and clearly distinguish between our use of the terms formative assessment and interim assessment. We understand that there is no current consensus on the definition of formative assessment, although leaders in our field have been working on a definition for some time now. Even less time and effort have been spent defining interim assessments. We offer a definition of interim assessment on the next page, but we fully expect it to be challenged and revised over the next several months. However, we felt it was important to lay out our current thinking on the definitions for both types of assessments before exploring the characteristics and evaluative criteria for interim assessments.

The definition of formative assessment, as proposed by a group of educational researchers wrestling with this issue[1] is as follows:

> *An assessment is formative to the extent that information from the assessment is used, during the instructional segment in which the assessment occurred, to adjust instruction with the intent of better meeting the needs of the students assessed.*

[1] The group of educational researchers includes Jim Popham, Dylan Wiliam, Lorrie Shepard, Rick Stiggins, Scott Marion, Phoebe Winter, Don Long, Stuart Kahl, and Brian Gong.

It was further revised by state assessment and other education leaders attending the Formative Assessment for Students and Teachers State Collaborative in Assessment and Student Standards (FAST SCASS) meeting in Austin, Texas in October 2006 to read:

> *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.*

Although other definitions exist, both of these definitions fit nicely with the work done by Black and Wiliam who defined formative assessment as just one part of formative instruction. In their seminal piece, *Inside the Black Box,* Black and Wiliam (1998) 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. In this definition of formative assessment, the assessment system consists of three phases:

1. Assessment (item development and delivery)
2. Diagnosis (analysis and reporting)
3. Prescription (pedagogy and professional development)

That is, a true formative assessment system does not stop with the development and administration of a test, but includes analyses that probe more deeply into what an incorrect answer implies about student learning and what should be done next or in the near future to further that learning. As stated by Black, et al. (2002), the first priority of a formative assessment is "to serve the purpose of promoting pupils' learning."

Both the classroom assessments, as described by Black and Wiliam and some commercially-available assessment systems may fit this definition of formative. However, we will distinguish between these small-scale, short-cycle assessments described most recently by Wiliam (2006) and the medium-scale, medium-cycle assessments currently in the field, which we are calling *interim assessments*. Although some have commented that since there are so many names for these assessments already in the field, the last thing we need is another name, we propose the term "interim assessment" as an umbrella term for all of these other assessments. We chose this term because one of the dictionary definitions of "interim" is, simply, "at an interval." Thus, benchmark, diagnostic, predictive, and even some formative assessments are considered interim assessments under our definition as follows:

> *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 purpose and intended uses, but the results of any interim assessment must be reported in a manner allowing aggregation across students, occasions, or concepts.*

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. By this definition, end-of-chapter tests available in most textbooks could be considered interim, if they are used in the aggregate. Teacher-created tests given at the end of a unit could be interim or formative, 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 at both the classroom and beyond the classroom level, such as the school or district level. We will argue that the academic goals should be linked to the curriculum that a student has been taught (or should have been taught) at the time the assessment has been given.[2] Typically, these goals are aligned with state content standards. In addition, a key distinction among the assessment types is the recipient of the information. For interim assessments, although the results may be used by the teacher to adjust instruction, another recipient could be and educational leader such as a school administrator, curriculum supervisor, or district policymaker. Unlike formative assessments, interim assessments can be aggregated easily, and the results can be examined across classrooms, schools, or even districts.

Overview of Interim Assessments

There are many forms of interim assessments currently available, often labeled "benchmark," "formative," "diagnostic," or "predictive." They can be given early in the school year, mid-way through, or periodically throughout the year. The one common thread is that they are designed to give information about the students' level of knowledge and skills before the end of the school year. Our goal is not to describe one assessment, called an interim assessment, but to focus the discussion on the different uses of these assessments and describe how the best practices for any interim assessment are related directly to the intended use of the assessment.

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 somewhat 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

- ✓ Evaluate how well the student has learned the material taught to date,
- ✓ Predict students' performance on a summative assessment, or
- ✓ Determine whether one pedagogical approach is more effective in teaching the material than another.

These are just three possible purposes. Interim assessments may serve multiple purposes as well, by providing aggregate information on student achievement at a district level, while providing specific feedback on where the gaps in a particular student's knowledge are at the classroom level. Many currently-available interim assessments have been called "early-warning tests" or, more pejoratively, "mini-summative tests." Their purpose is to determine whether the student is on track to succeed on the summative assessment. These tests may also serve formative purposes as they should diagnose and provide corrective feedback to help the student get on track to succeed on the summative assessment and not to simply predict how

[2] One exception to this rule of assessing what has already been taught is if the purpose of the assessment is to determine the starting point for instruction based on the knowledge and skills students already have when the students' level is unknown to the teacher or school.

the student will perform on the end of year test. Interim assessments may serve other purposes as well, including motivating and giving feedback to students about their learning. Many students think they know something but are often surprised and motivated to do something different when they get the results back from a quiz or classroom test, even without specific feedback.[3] Another purpose of an interim assessment could be to provide information to help the instructor better teach the next group of students, by evaluating the instruction, curriculum, and pedagogy.

Exhibit 1 provides a more comprehensive list of possible purposes for interim assessments, as well as information on the types of information that may be assessed and at what level the information will be used. In addition to the purposes discussed in the previous paragraphs, we also considered other purposes, such as ensuring that teachers are staying on track in terms of teaching the curriculum in a timely manner, providing a more thorough analysis of the depth of students' understanding, and determining whether students are prepared to move on to the next instructional unit, to name a few.

Summarizing this large table 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. Further, 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 its purpose.

[3] Although the research (e.g., Sadler, 1989) is quite clear that task-specific feedback is superior for improving student performance.

Exhibit 1: Interim Assessment Chart

		Who the results are primarily intended to inform					When the test is given			Type of item that would be appropriate						
Purposes: How results are used	What is measured	State policy makers	District policy makers	School Administrators	Teachers	Students & parents	Early in the instruc-tional period	Periodically- on a calendar-based interval (e.g., every X weeks)	Periodically - at an instruc-tional break (e.g., end of a unit)	MC	SCR	ECR -- essay	Short tasks	Synthesis task (research report; lab design/ analysis)	Checklist	Oral explanation
Instructional																
To provide aggregate information on how students in a school or district are doing at a given point in the school year	Content standards of material that is supposed to be covered to date	✓	✓	✓	✓	✓		✓		✓	✓	✓	✓			
To provide feedback on individual students for teachers and aggregate data on student progress for schools and districts	Knowledge and skills of material to be covered next		✓	✓	✓				✓	✓	✓	✓	✓			

		Who the results are primarily intended to inform					When the test is given			Type of item that would be appropriate						
Purposes: How results are used	What is measured	State policy makers	District policy makers	School Administrators	Teachers	Students & parents	Early in the instruc-tional period	Periodically - on a calendar-based interval (e.g., every X weeks)	Periodically- at an instruc-tional break (e.g., end of a unit)	MC	SCR	ECR -- essay	Short tasks	Synthesis task (research report; lab design/ analysis)	Checklist	Oral explanation
To diagnose gaps between student knowledge and intended curriculum	Information that has been presented in the classroom to date.			✓	✓	✓	✓		✓	✓	✓	✓	✓			✓
To see where a student's current knowledge and skills are in relation to what is about to be taught. i.e., is the student ready to learn the material in the next instructional unit?	Knowledge and skills required to understand upcoming curriculum as well as samples of upcoming curriculum			✓	✓	✓	✓			✓	✓	✓	✓			✓

		Who the results are primarily intended to inform					When the test is given			Type of item that would be appropriate						
Purposes: How results are used	What is measured	State policy makers	District policy makers	School Administrators	Teachers	Students & parents	Early in the instruc-tional period	Periodically- on a calendar-based interval (e.g., every X weeks)	Periodically - at an instruc-tional break (e.g., end of a unit)	MC	SCR	ECR -- essay	Short tasks	Synthesis task (research report; lab design/ analysis)	Checklist	Oral explanation
To determine students' ability levels in order to group them for instructional purposes	Knowledge and skills required to understand upcoming curriculum as well as samples of upcoming curriculum			✓	✓	✓	✓			✓	✓		✓			✓
To help the teacher determine where to put his/her efforts: reviewing previously taught material, teaching material in more depth, moving on to new unit.	Knowledge and skills of material just taught			✓	✓	✓			✓		✓	✓	✓			✓

		Who the results are primarily intended to inform					When the test is given			Type of item that would be appropriate						
Purposes: How results are used	What is measured	State policy makers	District policy makers	School Administrators	Teachers	Students & parents	Early in the instruc-tional period	Periodically - on a calendar-based interval (e.g., every X weeks)	Periodically- at an instructional break (e.g., end of a unit)	MC	SCR	ECR -- essay	Short tasks	Synthesis task (research report; lab design/ analysis)	Checklist	Oral explanation
To enrich curriculum	Information that has been presented in the classroom, but assessed in a way that requires students to show their depth of understanding			✓	✓	✓			✓			✓	✓	✓		✓
Test knowledge and skills not easily assessed in a time-constrained large-scale assessment	Information that has been presented in the classroom, but assessed using lab analyses or research reports or other such mechanisms				✓	✓			✓					✓		
To allow students to evaluate their knowledge and see areas in which they need to grow.	In depth knowledge and skills of material just taught				✓	✓			✓	✓	✓	✓	✓	✓		✓

		Who the results are primarily intended to inform					When the test is given			Type of item that would be appropriate						
Purposes: How results are used	What is measured	State policy makers	District policy makers	School Administrators	Teachers	Students & parents	Early in the instru c-tional period	Periodica lly- on a calendar-based interval (e.g., every X weeks)	Periodicall y- at an instruc-tional break (e.g., end of a unit)	MC	SCR	ECR -- essay	Short tasks	Synthesis task (research report; lab design/ analysis)	Checklist	Oral explanation
To encourage students to evaluate their own knowledge and discover the areas in which they need to learn more.	Knowledge and skills of material to be covered next				✓	✓	✓		✓	✓	✓	✓	✓			✓
To practice for summative test	A portion of the knowledge and skills that will be measured on the summative assessment			✓	✓	✓		✓		✓	✓	✓				
Evaluative																
To evaluate the effectiveness of various curricular and/or instructional practices	Content standards of material that is supposed to be covered to date		✓	✓	✓				✓							

		Who the results are primarily intended to inform					When the test is given			Type of item that would be appropriate						
Purposes: How results are used	What is measured	State policy makers	District policy makers	School Administrators	Teachers	Students & parents	Early in the instruc-tional period	Periodically- on a calendar-based interval (e.g., every X weeks)	Periodically- at an instruc-tional break (e.g., end of a unit)	MC	SCR	ECR -- essay	Short tasks	Synthesis task (research report; lab design/ analysis)	Checklist	Oral explanation
To increase teacher knowledge of assessment, content domain, and student learning	Content parallel to what is on summative assessment, although covering only what has been taught to date			✓	✓		✓		✓	✓	✓	✓	✓			
To reinforce curricular pacing	Content standards of material that is supposed to be covered to date		✓	✓	✓	✓		✓		✓	✓	✓	✓		✓	
To provide information on how best to target the curriculum to meet student needs	Knowledge and skills required to understand upcoming curriculum as well as samples of upcoming curriculum		✓	✓	✓		✓		✓	✓	✓	✓	✓			✓

		Who the results are primarily intended to inform					When the test is given			Type of item that would be appropriate						
Purposes: How results are used	What is measured	State policy makers	District policy makers	School Administrators	Teachers	Students & parents	Early in the instruc-tional period	Periodically- on a calendar-based interval (e.g., every X weeks)	Periodically- at an instruc-tional break (e.g., end of a unit)	MC	SCR	ECR -- essay	Short tasks	Synthesis task (research report; lab design/ analysis)	Checklist	Oral explanation
To predict student achievement on summative test	Same content standards as on summative tests, as mini versions of the summative assessment	✓	✓	✓	✓	✓		✓		✓	✓	✓				
To predict student achievement on summative test	Same content standards as on summative tests, assessing appropriate segments that should have been taught to date	✓	✓	✓	✓	✓		✓		✓	✓	✓				
To provide an early warning system for students who are not on track to succeed on X	Knowledge and skills required for success on X (e.g., assessment anchors on summative assessment)	✓	✓	✓	✓	✓		✓		✓	✓					

Instructional

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 better meeting the needs of the students assessed. 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, an assessment that asks a student to explore a concept in greater depth or one that provides tasks that stretch students and teachers to do things at deeper cognitive levels than they might otherwise. These purposes would certainly be considered instructional, and we would argue that these are laudable goals. However, this type of assessment does not neatly fit the earlier definition of a formative assessment that *"information from the assessment is used, during the instructional segment in which the assessment occurred, to adjust instruction with the intent of better meeting the needs of the students assessed"* or that it *"provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes."* Rather, the assessment itself provides the rich instruction. It is worth noting, however, this type of assessment meets an earlier definition of formative provided by Black, et al. (2002) stating that the first priority of a formative assessment is "to serve the purpose of promoting pupils' learning."

As a second example, consider the 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 are 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 fully aligned with state curriculum and provide more in-depth analyses of student misconceptions or lack of understanding along with instructional tools and strategies for improving instruction. However, this type of interim assessment does fall under the instructional category. Also, because the results can be aggregated and used at a level outside of the classroom, it meets the definition for interim.

As a third example, consider an end-of-chapter test that students self-administer when they have completed a unit. The purpose of this assessment may be two-fold—to ask students to assess their understanding of a particular topic and to provide information that can be aggregated at the classroom level to inform the teacher of the readiness of the class to proceed to the next unit.[4] This type of assessment is certainly an interim assessment serving an instructional purpose, although we would argue that there needs to be evidence that this end-

[4] The results of this assessment could also be aggregated across any classroom or school that uses the same textbook.

of-chapter test is fully aligned with state standards if the test is also to serve the purpose of preparing students for the end-of-year assessment.

We also argue that to serve instructional purposes an assessment system must go beyond simply providing data. Test designers must provide strategies for interpreting and using the data to effectively modify classroom instruction. At the very least, the assessment system should include information about possible interpretations for each incorrect answer. 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 in that an assessment system purchased for instructional purposes must include professional development to ensure that educators have the tools to use the results appropriately. We believe that this should be a shared responsibility among the developer and the user.

Evaluative

Another type of purpose an interim assessment might serve is to provide evaluative information about the curricular approach or instructional strategies. 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 that would receive the instruction in the future.

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. Likewise, 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 students understood well and which were less clear to teachers within one or more schools 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 informative about relatively fine-grained curricular units. A global, content area score will not suffice. 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

We suspect that there are few assessment systems where the only purpose for 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. In addition, assessments in this category could be used to predict performance on the high school exit exam or success with post-secondary curriculum. Although our focus will be on the predictive use of these assessments, we expect most users want additional information to help them improve the performance of students for whom failure is predicted.

These 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.

Although there has been some discussion as to the worth of an assessment that only provides information that a student is on track to fail without additional diagnostic information to guide interventions, 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 importance of having all three of these assessment types aligned in 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. That is, it should not be considered a strike against the predictive test if a student predicted to fail the summative test actually passes it, but questions should be raised if too many students predicted to pass the summative test actually fail it.

Identifying the Goal

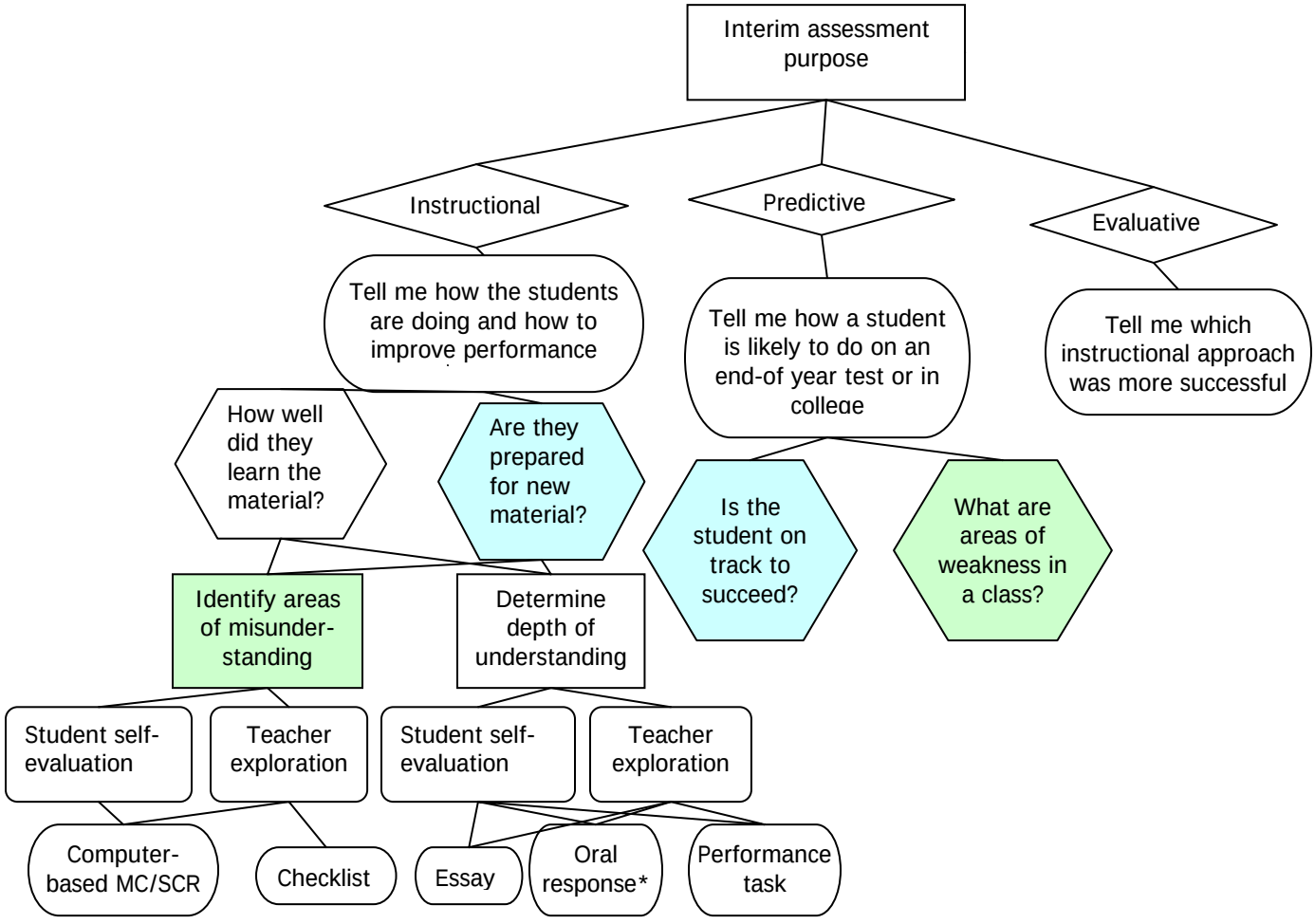
As policymakers decide to bring an interim assessment system to their state/district/school we encourage them to have a theory of action for how the particular assessment system will work in the teaching-learning cycle. As a start, we think it will be helpful for educational leaders to address the following 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?
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 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. We present a partial decision tree as shown in Figure 2, which shows the types of questions policymakers should be able to answer about the type of system they want before developing or purchasing one. 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. Note that it is not complete, as a complete tree would need several pages for the display, but it should provide a clear example for policymakers to use when thinking through the issue of selecting or designing and interim assessment system.

Figure 2: Decision Tree to Use in Creating an Interim Assessment System



*May be better answered using classroom formative assessment techniques.

Answering the questions in this decision tree also may reinforce the idea that it is often appropriate to consider multiple 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 students not understanding key concepts. Likewise, 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. Areas of overlap in the questions are indicated by the coloring schemes. Again, some of the instructional questions may be answered best through the use of classroom formative assessments, but there are interim assessments that may also provide useful information at the classroom level and also answer predictive or evaluative questions.

The next section describes the various characteristics of any effective assessment system. It also provides several possible approaches to developing one. The appropriate approach should be determined based on the answers to the above questions. There is no one-size-fits-all assessment, only a best design for a desired use and the 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 the 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.

Characteristics of an Effective Interim Assessment System

This section is intended to help educational leaders either choose or develop a strong interim assessment system for their schools. 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 both for evaluating currently-available systems and for designing new systems.

Shepard's (2006) requirements for an assessment to have strong formative potential include:

- ✓ Can never be *all* multiple-choice.
- ✓ Must provide *qualitative* insights about understandings and misconceptions not just a numeric score.
- ✓ Should have immediate implications for what to do besides re-teaching every missed item (the 1000 mini-lessons problem).

While we continue to separate formative assessment from interim assessments with formative purposes, we believe that any good interim assessment system will produce results that can be used to inform instruction. We are clear that the current research literature supports the types of formative assessments defined earlier and does not yet offer guidance about the efficacy of certain interim assessments. However, certain reports of best practices (e.g., Marshall, 2006) indicate that effectively using interim assessments can be a lever for powerful educational reform. Characteristics of an effective interim assessment system include:

- ✓ A 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
- ✓ Reliable results that are easy to interpret and clear guidance on how to use the results

Other characteristics may be important depending on the purpose(s) of the assessment. Again, we emphasize that the purpose must be clearly stated before one can truly determine or evaluate the necessary characteristics of the assessment. One element we consider essential to any interim assessment system is validity evidence. The information provided by and uses of interim assessments should be validated, and we will focus on this aspect of the assessment system later in the paper.

One strategy for defining the desired characteristics is to focus on the reporting element. What do we want the tests to tell us? A report is a mechanism for translating the assessment data into action and should be one of the first considerations in designing a new assessment. In the next few pages we discuss important considerations for reporting results from interim assessments, and then we follow the assessment process from design to implementation and scoring.

Reporting Results

We would argue that the score report is one of the most important components of an interim assessment system. It serves to make the results actionable. A good report will indicate not only which questions a student answered incorrectly, but also what the student's incorrect response or set of responses implies about learning gaps or misconceptions, what further probes may be administered to gather more information about the problem, and instructional strategies for improving student learning on that topic. Again, this last point about instructional strategies may be provided in a supplemental professional development package, but we believe it should be considered when designing a reporting system. At the least, there should be research behind the items in the assessment that can be used to inform general instructional approaches that might be considered for improving student learning. That is, assumptions about where a student is in his/her learning progression based on an incorrect response should be grounded in research on learning progressions.

Student-Level Information

In evaluating an interim assessment program, it is important to examine the type of student-level information provided and potential users must evaluate the extent to which that information matches the goals of the program. Many commercially-available systems provide information on which items were answered correctly and incorrectly and how those items map to state content standards. Most systems also aggregate the items into clusters or subscores. That is, the results may be broken out by content strand, so a student may have an overall score for a math test and also have subscores in numbers and operations, geometry, and measurement, for example. Unfortunately, many assessment programs have had little success in validly reporting meaningful subscores, so we must be cautious with promises or assumptions that any interim assessment system can do so. Consider, however, the richness of student-level information that can be gained through performance tasks or essays that are reviewed by the

teachers, or even results of multiple-choice or short-answer tests when the questions have been written to measure specific features in a student's learning progression. These types of student-level information may be displayed qualitatively (in the form of notes) rather than quantitatively (number-correct scores) and could be quite valuable.

Aggregate Information

For this to be an interim assessment, it must be possible to aggregate information at the classroom, school, or district level. Given time constraints and limited resources, a teacher may want a more holistic view of the level of understanding of all the students in the classroom, rather than individual diagnostic information for each student. A district policymaker may be concerned about general trends by school or about the proportion of students expected to reach proficiency by the end of the year. It would be important in this situation to have a reporting system in place that could provide aggregate information across any level desired, from classrooms within schools to schools within districts to districts within a state. Following the example in the previous section, rich information about an individual student can be gained from an essay test, but the total score (or even the score on various components of a rubric) could be aggregated across students to provide additional information both at the classroom level and beyond.

Additionally, it may be important to be able to *disaggregate* the data across different groups. That is, it may be important to distinguish performance among different groups of students or across students in different schools or different instructional programs. One result of NCLB is that assessment results are almost always broken out by student groups such as race/ethnicity, gender, socioeconomic status, disability status, and English language proficiency. Many available interim assessment systems also include mechanisms for grouping results into these groups.

Follow-up Information

As discussed in an earlier section, those who wish to use an interim assessment to help inform instruction need to know where the students are in their learning, how that compares to where they are going, and how best to reach that goal. The reports must lend themselves to informing a plan for action intended to further the student's learning in appropriate areas. Simply aggregating the data and identifying the content strands on which the students missed the most items is not sufficient. Ideally the follow-up information will be curriculum-specific. For example, when interim assessments are not tied to specific curriculum sequencing, students might answer items incorrectly because (a) they were instructed on the content but did not understand a concept or (b) they had not yet covered the material at the point the test was administered. Given the money spent on some interim assessment packages, it does not make sense that a teacher's first interpretative action involves determining whether the missed items were taught or not.

Even if the purpose is predictive rather than instructional, the follow up information should include the expected end-of-year achievement for each student along with intervention strategies to improve students' end-of-year outcomes. That is, if a student is currently projected to score at the basic level, it would be helpful to know how to move the student towards scoring at the proficient level. This aspect of reporting is the most frequently overlooked but is arguably the most important. The recommendations included should be grounded in research on student learning and cognition and provide clear guidance to teachers.

Summary of Interim Report Features

Again, it is important to clarify the purpose of the assessment when examining its various features. 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. Further, to judge a reporting system's effectiveness, it must be vetted through those who need to use the information: teachers in most cases but also school leaders.

Developing the Items

Once the purpose and desired form of results are clear, we can begin to think about the types of items that would be appropriate for the interim assessment. Different types of items may be selected for different purposes. For instance, if the test is being used to predict student performance on an end-of-year test, then the item types should match the types of items that will be found on that summative test—typically multiple-choice and short constructed-response items. If, however, teachers also want further information on a student's depth of knowledge or areas of weakness, other types of items may need to be used to supplement the results of the interim assessment, such as performance tasks or essays.

The items must be developed in such a way to represent in-depth coverage of the specific content standards and curricular goals to be assessed. The items should have all the positive elements of any item in a large-scale assessment (reliable, free from bias, etc.) but also be able to provide information on the student's depth of understanding or what a student does not know through an analysis of incorrect responses or a thorough scoring rubric. Items should be linked directly to both the state content standards and to the individual units of learning as determined by the curriculum guides.

Scoring is also an important consideration in item development. Many commercial systems are designed to be scored electronically. Certainly, electronic scoring allows results to be produced quickly and aggregated easily across classrooms and schools. However, one should consider the learning value of students self-scoring or teachers scoring student work. 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. In the following sections, we discuss just a few types of items in terms of their appropriateness for an interim assessment as well as distinguishing characteristics of good items.

Multiple Choice

Shepard (2006) makes it clear that assessments used primarily to inform instruction should not consist solely of multiple-choice (MC) items because MC items do not provide enough information about *how* students understand. However, some 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. Specifically, items built with an understanding of students' learning progressions will provide richer information on gaps in student knowledge. A classic example of a multiple-choice item that provides evidence of student learning is the following:

Consider the four diagrams shown below. In which of the following diagrams, is one quarter of the area shaded?

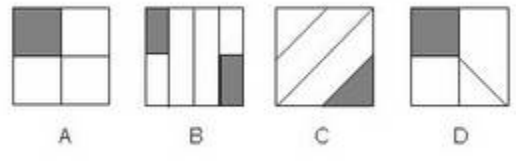


Diagram A is the obvious answer, but B is also correct. However, some students do not believe that one quarter of B is shaded because of a belief that the shaded parts have to be contiguous. Students who believe that one quarter of C is shaded have not understood that one region shaded out of four is 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.[5]

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 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.

Unfortunately, the type of multiple-choice item presented as an example here is rarely found in practice with most commercially available interim assessments. In fact, we would argue the same point regarding end-of-year large-scale assessments. While it is certainly possible to write high quality multiple-choice items, the demand for a large quantity of items produced quickly has led to considerable concerns with quality of multiple-choice items found on commercially-available assessments. Ideally, if the purpose of using an interim assessment requires large numbers of multiple-choice items for quick-turnaround and easily aggregated data, test developers should aim to include as many multiple-choice items as possible that are

[5] This example is from a presentation by Dylan Wiliam delivered at ETS in 2006. The exact origin of the item was not given.

built from learning theory and provide useful information about wrong answers as well as correct ones.

Open Ended

Any type of item that requires students to generate their own answers should provide richer detail about student thinking as well as support the goal of moving students towards deeper thinking about the subject matter. Once again, however, to provide information that is useful to instruction, the items should be constructed in a manner that allows the student to provide information on his/her thinking. Some examples of these types of items include essay prompts that ask students to justify their point of view. An essay may take longer to score and even longer to extract all the information available to inform instruction, but it provides a rich diagnostic tool for teachers. Simply examining the student work as they score the essays can also provide teachers with professional learning opportunities by giving them greater insights into how their students process information.

If the purpose of the assessment is to provide aggregate information quickly, short constructed-response items may be more appropriate. However, 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.[6]

For these types of item sets to be effective, however, strong supporting material must be included as part of the assessment. That is, included in the assessment system should be instruction on the types of errors to look for, how to detect errors, and what corrective steps

[6] An expanded version of this and other task sets can be found in Pellegrino, Chudowsky, and Glaser (eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press

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.

Performance Tasks

Few, if any of the current commercially available interim assessment systems include items other than multiple-choice items. Part of the purpose 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 can be included as part of an interim assessment system to provide opportunities for more in-depth focus on the content area than is often the case with selected-response item types. 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, an assessment with these types of tasks falls under our definition of interim.

The tasks themselves can enrich and deepen the curriculum. That is, a task may require students to learn one particular objective in much greater detail by analyzing it or generalizing from it. These tasks can also serve the purpose of having students self-evaluate their learning and understanding of a concept or to synthesize their understanding across several concepts or standards. Reading about an idea in a book is very different than making it come to life in a task. Students will be able to better evaluate their own depth of knowledge and explore other avenues on their own, which serves a type of formative purpose.

Performance tasks, particularly extended tasks, can serve instructional purposes more readily than other interim assessment item types. Having students engage in these tasks provides opportunities for the teacher to observe students as they solve problems or otherwise work on the task. The teacher may stop and probe to understand better why the students are doing poorly or well. Depending on the larger purpose of the interim assessment system, this interaction may not be possible, but teachers can still conduct systematic observations to provide feedback to students soon after the task is completed as well as gather data that can be aggregated across students, concepts, or occurrences.

Another way to conceive of this type of performance item that can be easily aggregated in an interim assessment system is to consider a checklist. Under certain circumstances where a set of discrete tasks is well defined, a teacher may use a checklist to evaluate all students as they conduct these series of tasks. As the students work, the teacher may simply walk around the classroom with the checklist and mark off which students successfully complete each component of the task and note which aspects students are struggling to understand. The checklist can be aggregated across students or tasks and the notes can be used to improve instruction. In addition, the teacher may stop and probe to understand better why the students are doing poorly or well. We would encourage educational leaders to consider the types of tasks being assessed and determine whether a checklist or a rubric would provide more useful information.

Performance tasks, particularly extended tasks, such as a research paper, science laboratory, or historical debate, have the advantage of helping to erase the artificial 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.

Summary on Evaluating Item Types

Again, once the purpose and reporting elements are clearly outlined, the type of items required should become clearer. Assessments serving a predictive purpose might need to be more aligned with the test design of the summative assessments to which they are linked, while assessments serving an instructional or evaluative function will likely be better served by open-ended items and performance tasks. The turnaround time of results should also be considered. That is, if a teacher or educational leader needs results quickly, multiple-choice or short constructed-response items can be scored more rapidly than essays or performance tasks. The audience should also be considered. If a school or district administrator is only interested in obtaining aggregate information on the students, quick-turnaround responses, typically in the form of multiple-choice or short open-ended items, will suffice. If, however, an additional goal is to provide useful information back to the teachers to use in instruction, more performance tasks or longer open-ended items may serve both purposes.

Beyond the specific item types, users should carefully examine the quality of the items they are considering using. An assessment system will be unable to meet any of the goals put forth by stakeholders if the system is built on low-quality items. There has been much written about item quality (e.g., most introductory measurement texts include extended sections on item quality) so we do not want to go into detail about this issue here. However, potential users of commercial systems should be encouraged to conduct regular and structured (even expert led) reviews of the quality of items that will be used on their tests. Similarly, those who intend to develop a customized or other local system need to attend carefully to the quality of the items and tasks used to ensure that the assessments can validly fulfill the intended purposes. That is the items should undergo bias and sensitivity review as well as content review. Item should be field-tested whenever possible. These quality checks might require the use of expert consultation to help with local systems. 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.

Administering the Test

The final consideration in evaluating or designing an interim assessment system is the manner in which the test will be administered. As mentioned early, a key component of many of the commercially-available interim assessments is that they produce results quickly. Most, if not all, commercially available interim systems are computer based, and often web based. The use of technology has several advantages:

- ✓ Allows for flexibility in where the test is taken (anywhere with a computer)
- ✓ Allows for flexibility in when the test is taken
- ✓ Easily creates new forms for different purposes
- ✓ Allows teachers or curriculum leaders to customize forms easily
- ✓ Allows for the possibility of an adaptive test
- ✓ Gives results in real time

- ✓ Aggregates results quickly across a classroom, school, district, state
- ✓ Easily disaggregates results by student subgroup
- ✓ Calculates normative information on the spot

An obvious caveat to the use of the technology is the availability of that technology. If a teacher is to use effectively a web-based system, all students in the classroom need to have access to a computer connected to the internet. If only one such computer exists in the classroom, other provisions may need to be made. Some technology-based systems also include a pencil-and-paper option.

Other potential drawbacks or limitations of being technology-based include:

- ✓ Requires teachers to know how to use the technology effectively
- ✓ Too much customization may reduce the reliability of the assessment
- ✓ May be more difficult to monitor who is actually taking the assessment and under what circumstances
- ✓ Does not allow for interaction between the student and instructor, losing a source of valuable information
- ✓ Current use tends to be limited to selected response items although automated and artificial intelligence approaches to scoring open-ended items are progressing rapidly, particularly in fields such as medicine

Depending on the purpose of the assessment, one should also consider the diagnostic benefits of a teacher administered and scored assessment. Formats such as checklists, performance tasks, and orally administered assessments can provide insights into student thinking that is not readily apparent from a technology-based assessment. Quick turnaround of results is not always essential depending on the purpose. For example, if one purpose is to explore students' thinking and to give teachers better insight into the effectiveness of their instructional strategies, an essay test could provide rich information, even though it may take days, not minutes, to score.

Matching the Administrative Features to the Purpose of the Assessment

The administrative requirements must be considered in conjunction with the test purposes. Depending on the goals of the assessment system, certain features will be necessary, others helpful although not crucial, and other features become irrelevant. Exhibit 2 provides a crosswalk of interim assessment purposes with several administration features.

Exhibit 2: Crosswalk of Assessment Purpose by Administration Requirements

Purpose	Speed of results	Availability of normative information	Flexibility of test administration	Customization of test form	Adaptive
Predict student achievement	Days	Not necessary	Could be given on a flexible schedule or a pre-set times during the year	Not necessary	Nice feature, but not essential
Provide information on strengths and weakness of a particular group of students	Within a few days	Necessary	Needs to be able to be given at the teacher's discretion	Helpful	Nice feature, but not essential
Reinforce curricular pacing	Days-	Helpful	Not necessary	Not necessary	Not necessary
Evaluate the effectiveness of instructional program	Weeks	Necessary	Not necessary	Not necessary	Nice feature, but not essential
Evaluate a student's understanding of a topic	Within a few days	Not necessary	Essential	Helpful	Helpful
Determine a student's preparedness for the next topic	Within a few days	Not necessary	Essential	Helpful	Helpful
Enrich the curriculum	Days or weeks	Not necessary	Essential	Essential	Not necessary
Provide professional development to the teacher	Weeks	Helpful	Not necessary	Not necessary	Not necessary

Summary of Administrative Considerations

In creating this table, we found that we could argue for a different response in some of these cells. For instance, if the purpose of the assessment is to provide exposure to richer curriculum, it could be important to return results immediately, or, if the assessment items themselves provide the instruction, then turnaround time on the results is less important. Likewise, if the purpose of the assessment is to improve a teacher's understanding of the interaction between student learning and performance on the test, it may be important to give the teacher flexibility in determining when to give the assessment and which items to include, or a more standardized assessment could serve the same purpose—particularly if the professional development was intended across a larger scale. However, our purpose in constructing this table was to summarize what was generally the case and to provide an illustration of the thinking an educational leader should do when evaluating or design an interim assessment system.

Evaluative Criteria

To help guide the evaluation of commercially available interim tests, we have provided the following criteria for states and/or district to consider prior to purchasing an assessment system. 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 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 shown 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 to both the knowledge and skills (including depth of knowledge) called for in the target content standards.

3. 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 uses for which there is no validity evidence.

4. Tasks should be applicable to a wide-range of student populations, including English language learners and students with disabilities.

5. 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, more intensive and sustained professional development is required for assessments serving instructional purposes compared with other purposes.

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

Instructional

1. To the extent possible, interim assessments for instructional purposes should be as seamless with instruction as possible and represent an opportunity for student learning during the assessment experience.

2. 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.

3. Clear guidelines should be provided explaining how the results of the assessment, including the results of particular tasks and items, should be used to help inform instructional decisions.

4. Each particular assessment in the system by must be closely linked 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 to determine readiness for some learning in the near future.

5. To best serve instructional purposes, each interim assessment should assess only a limited number of important curricular goals to make it more likely that remediation can be timely and targeted appropriately.

6. In general, interim assessments to serve instructional purposes should be comprised more from high quality open-ended tasks than selected-response items because of the greater probability for correctly diagnosing students' understandings and misconceptions. Multiple-choice items should be developed with an understanding of learning progressions, such that useful information can be gleaned from specific incorrect answers.

7. Instructional interim assessments should measure instructional and curricular goals not easily assessed on the states large scale assessment such as extended tasks or synthesis works.

8. 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.

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. Again, this should be examined during the alignment study.

2. The assessments should be comprised of items and tasks with a mix of formats to allow for users to gain a deep understanding of the 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. Considering the potential high stakes associated with such assessments (e.g., determining whether or not to spend scarce resources), the reports must accurately portray the error associated with the scores and subscores.

5. The assessment should be related to other measures of the intended constructs and less related to tests of domains other than the intended constructs.

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 without adding another assessment to the system.

2. The predictive assessment should be comprised of 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, but each test should include only content on which the students have been instructed up to that point.

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. The assessment should contain enough diagnostic information so that remediation can be targeted for students predicted to score below the cut on the criterion measure. If the assessment is unable to provide such information, additional guidance should be included in the system to help with remediation.

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 suggest that educational leaders consider the criteria when evaluating which, if any, system to purchase or when evaluating a proposal to create a customized system. Additionally, any vendor should be required to provide evidence of the validity of the system for the intended purposes. Once the system has been implemented, districts and/or states should 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:

- ✓ 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 for current students.

- ✓ 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 are supported.

- ✓ If the test is used for predictive purposes, do a follow up study to determine that the predictive link is reasonably accurate, more than things such as grades and teacher judgments, and that the use of the test contributes to improving criterion (e.g., end of year) scores.

In addition to these suggestions, interim assessment systems should be evaluated for the effects on important aspects of the teaching and learning process, such as:

- ✓ Student learning, especially in terms of generalizability and transfer
- ✓ Student motivation as a result of engaging with these tasks
- ✓ Curricular quality as a result of incorporating tasks
- ✓ Increases in teacher knowledge of content, pedagogy, and student learning
- ✓ Manageability, including the quality of implementation

Current Commercially Available Interim Assessment Systems

As mentioned earlier, many test publishing companies offer interim assessment products, often labeled "formative" or "benchmark" assessment products. Before writing this paper, we searched the internet for companies that offered these types of assessments by entering the terms "formative assessment" "predictive assessment" and "benchmark assessment" into search engines. After reviewing over a dozen websites of various vendors marketing these products[7], we found many commonalities across the systems.

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, 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[8], 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. That is, the student score on the interim assessment is used to predict performance on the large-scale summative assessment. Some of these systems have quantitative data showing the statistical link between the interim and summative assessments

[7] And sometimes even the same vendor selling multiple versions of their product using all of these labels!

[8] 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.

and the correlations between the scores as evidence of the interim assessment's predictive ability. They usually include statistical data on 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).

In addition, policymakers should consider the validity of these assessments as part of their overall state comprehensive assessment system. The alignment of these assessments to the state content standards should be evidenced through an external alignment study. Furthermore, these assessments should be considered in the wider context of a comprehensive, balanced assessment system. Policymakers should be aware of the potential for several assessments to provide conflicting information. This is particularly problematic when a teacher is using a short-cycle formative assessment in her classroom and receiving different information from a district-wide benchmark assessment. If the short-cycle and medium-cycle assessments are providing conflicting information on the content students have mastered, they are not likely to be useful to that teacher.

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
- ✓ 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
- ✓ Which students are at risk of scoring below Proficient on the end-of-year NCLB tests?
- ✓ Which content standards are the students' performing relatively best (or worst) on[9] (for a student, classroom, school, district, state)?
- ✓ How does this student's performance compare to the performance of other students in the class?

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

Although most systems meet some of the requirements for an effective interim assessment, few, if 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.

Thus, we find that most commercially-available interim assessment systems currently do not:

- ✓ Provide rich detail about the curriculum assessed
- ✓ Provide a qualitative understanding of a student's misconception(s)
- ✓ Provide detailed information on the student's depth of knowledge on a particular topic
- ✓ Further a student's understand through the type of assessment task
- ✓ Give teachers the information on how to implement an instructional remedy

Furthermore, these systems cannot 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?

Matching the Purpose with the Assessment

The main driver of this paper 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. To make this idea more concrete, we have created some hypothetical case studies.

We have assumed different assessment goals for each scenario and mapped possible assessment types using examples of what currently exists along with what we would like to see developed. Each case study explores the desired reporting features, item types, and administrative aspects best suited to meet the specific purpose of the fictional district.

Case Study 1: Exploring a Student's Level of Understanding and Areas for Further Instruction

In this first case study, we have a district that is interested in developing an interim assessment system that can be used in the classroom to help a teacher gauge her students' understanding of what has just been taught and identify areas that need further instruction, either for remediation or enrichment. The goal is not to provide this information to the district policymakers, but results must be aggregated across classrooms to provide periodic reports to the school administrator. Ultimately, the district is not interested in predicting future achievement as its philosophy is that if students are taught the curriculum sufficiently well, they will perform successfully on the end-year-test. Their goal is to develop a system that is fully aligned with the state content standards, will evaluate students' understanding, help the school administrators monitor the students' learning, and provide professional development to teachers on how to administer the system, interpret the results, and apply instructional strategies to remedy any problems found in student learning.

From the reporting perspective, the most important level of reporting is the student level. Teachers need concrete information on what each student knows and does not know, and specifically where the lack of understanding or misunderstanding is occurring. Moreover, the reporting system needs to include information on further probes to explore the student's thinking and suggestions for teaching strategies to effectively counter any existing misconceptions or hindrances to the student's learning. Corrective feedback needs to be provided to students as well as suggestions for further exercises the help the student gain a rich understanding of the concepts covered.

In addition, teachers should have the tools to explore overarching patterns across the classroom in terms of which concepts were clear and which need further instruction. So, classroom-level results should be available. That is, there should be a mechanism in place of quickly aggregating individual student reports to a classroom report. Finally, there needs to be a mechanism for providing this classroom report to a school administrator, along with a summary of any further probes used and instruction given.

For the assessment design, the assessment should be integrated directly into an instructional unit with items linked to specific concepts within a unit. Care should be taken to ensure these concepts (and ultimately the items) are linked directly to the state content standards that will be assessed at the end of the year. Preferred item types are open-ended items that include performance tasks or essays. These items should require critical thinking and student self-reflection. For quick results that allow for easy aggregation, we could consider a system of multiple-choice items, checklists, or short constructed-response items that are followed by probes of why the student thinks there answer is the right answer. These data could be supplemented (or replaced) with a performance task such as asking the students to design and conduct a scientific experiment and chart and analyze the results (science) or developing an argument and writing an essay on why a historical event was inevitable or could have been prevented (history or language arts). Each component of this type of task could be analyzed to see where any misunderstanding or incorrect procedure occurs. These results could be used to give corrective feedback to an individual student and aggregated in a report that shows the percentage of students who can correctly articulate a hypothesis, develop a scientific design with all the appropriate components, record data, graph data, interpret data, and draw conclusions (science) or who demonstrate a rich understanding of a historical event, can develop a persuasive argument, and develop an essay appropriately combining opinion with fact (history or language arts).

Administration of this type of assessment requires much flexibility. The timing may be linked to a date on the calendar or a point in the curriculum. The administration may be computer-based or consist of paper-and-pencil tasks. Even if a teacher intends to read the students' essays or hypotheses, the student can type them into a computer. Results should be available relatively quickly to provide corrective feedback during the course of instruction although certain components, such as essays, may require more time to score. Some further probes may take additional time, but, again, this assessment is intended to work within instruction, not as a separate component.

This type of assessment requires a strong professional development component to be effective. Teachers will need to be trained to evaluate the information and use these types of tasks meaningfully, including modeling how to learn from student work. Professional development supports should also include next steps that reflect the information provided by the assessment

results. Any type of professional development on using assessment results to improve instruction should be continuous, implemented throughout the school year, and not a one-time course.

Case Study 2: Evaluating the Effectiveness of a Curriculum or Instructional Program to Identify Areas for Professional Development

Consider a district with multiple schools, with varying demographics and instructional approaches. The district leaders and policymakers are interested in having information to monitor that schools are providing fair opportunities to learn the required content and skills and to provide information to help evaluate the effectiveness of various district initiatives, including curricular reform activities. The district leaders are interested in receiving more in-depth information about student mastery of specific strands than can be obtained from the end-of-year assessment.

Or, to focus on a more specific simplistic (although less realistic) example with similar requirements, think of comparing two groups of schools, each using a different approach to teach mathematics. One is using a basic skills approach, while another uses an approach that requires integrating mathematical concepts into the everyday world. The first approach has a long list of knowledge and skills to be taught while the second approach has fewer knowledge and skills but requires greater depth of understanding for each. The two schools are similar in terms of student demographics and have performed similarly on previous assessments. The district superintendent wants information throughout the school year on which program appears to be most effective so that information can be provided to the other schools about this instructional approach.

The type of report that will best serve these purposes is a school-level report. The purpose is not to intervene at the student level but to evaluate the approach for the students as a whole. Therefore, the results will need to be aggregated across the school, and then disaggregated by student demographic group to look for any discrepancies in the effectiveness across students. In addition, student subscores would be important, as the policymakers would want to examine the results by the different content strands or instructional units.

The requirements of the assessment design could be similar to those for the end-of-year tests if a policymaker or educational leader wishes to analyze performance towards the goal at several points throughout the year. In this case, the items would need to map directly to the content standards and be similar in type to the items on the end-of-year test to provide information during the year on how prepared students are for the summative assessment. However, additional probes should be added to provide further understanding regarding a student's thinking and solution-strategy for each item. Another approach would be to assess similar domains using different formats to gather more in-depth information to complement the summative data resulting in a more robust evaluation. In this case, the policymakers might consider an open-ended supplement given periodically to provide greater insight into the students' level of understanding.

Finally, the administration requirements should be fairly standardized across the schools. Because these assessments will be used to inform instruction in the next school year, the results do not need to be turned around quickly. Either computer-based or paper-and-pencil tests would be appropriate for this situation. Students can be matrix sampled, too, to provide a

similar level of information with less of a burden on instruction. Measuring achievement periodically of a sample of students from each school will allow policymakers to evaluate the students' learning and growth over the year in the two programs and make appropriate comparisons.

In this case, the designers of an evaluative assessment system must take care to avoid privileging one type of assessment format over another so as not to unintentionally favor one curriculum approach more than the other. By linking carefully to the end-of-year criterion measure, the designers can document that their evaluative assessment is as fair as possible. In the case of this example, the system would likely include only two or three tests spaced widely through the year. This approach is to avoid favoring one curricular approach over another as a result of differential pacing relative to the assessment targets.

Case Study 3: Predicting Success on the End-of-Year Tests and Identifying Gaps in Knowledge

For the final case study, consider a district that wishes to implement an early-warning system to identify which students, classrooms, and schools are on track to perform well on the end-of-year assessment and which might need intervention to meet the annual measurable objective. Furthermore, for those students who are not on track, the district wants to be able to identify areas of weakness at the student level and aggregated to the classroom and school level. Ideally, they would like a system that provides additional tools for improving performance on those areas identified as weak. Finally, they would like to administer this test 3–4 times over the year and track student progress toward the goal.

First, let's examine the necessary components of the reporting system. The first criterion is that the assessment reports "on-track to succeed" as well as any areas of weakness. The results need to be aggregated across classrooms, schools, and the district. Scores should also be disaggregated by the same reporting categories used in the end-of-year reports, such as student racial/ethnic group, gender, disability status, economic status, and LEP. Each subsequent report should illustrate progress, providing feedback on where the student is, how they have developed over time, and how the progress relates to where they should be by the end of the year. That is, the report should show each student's current and project trajectory.

A key component to this reporting system is that areas of weakness for a student, classroom, and school must be clearly identified. Ideally, this information would be deeper than simply identifying content strands or indicators where students were less likely to succeed, but would also focus on specific concepts and provide follow up activities for each potential area of weakness. It should include instructional strategies based on research for improving student performance.

Next, let's look at the requirements of the assessment design. The items need to map directly to the state content standards and be similar in type to the items on the end-of-year test to provide a solid statistical link for predictive purposes. However, each test should only assess what's been taught to date. That is, this system should not be designed to give a series of parallel mini-summative assessments. The first assessment should cover the material that was scheduled to be taught in the first quarter, for example. The second assessment may provide some overlap with the material taught toward the end of the first quarter and cover the material taught in the second quarter. It is also important that the items link not only to the

state content standards but to teaching units and text books specific to that district. Using items that link directly to instructional materials will help provide the connection between any weaknesses found and instructional interventions.

And although it is important to use similar item types as used in the end-of-year assessment for statistical reasons, consideration should be given to adding a few open-ended probes to help diagnose any weakness found in student performance and to allow these less constrained assessments to measure student performance in ways unable to be assessed on the single end-of-year tests.

Finally, let's consider administration requirements. Because these assessments are not designed to integrate seamlessly with instruction but rather as a periodic check, it is not as important that the results be turned around within a day. However, the results should be available within a week or so to allow time for intervention. Particularly assuming that one unit builds on the next, it will be important to inform teachers quickly if students are not on track to succeed.

In terms of the actual administration, either computer-based testing or a pencil-and-paper test would serve this district's purpose. Also, standardization in the items administered would be necessary to aggregate results across the district. Therefore, flexibility is not a strong requirement for this system. Furthermore, since the goal is to ensure all students are on track to meet a certain criterion, normative data are not as important to provide as criterion references.

A critical assumption justifying this predictive use is that the results provide useful information beyond what teachers already have. If the results of predicting which students had mastered a unit sufficiently to be successful on the end-of-year assessment were quite different from the teacher's predictions or grades for the student, then there would be concerns about the degree of mismatch between the instruction and the end-of-year assessment. Likewise, if the assessment provided no additional information beyond what the teacher already knew from her daily interactions with the class, then the assessment would not be worthwhile.

Summary of the Case Studies

We recognize that most actual instantiations of interim assessment systems do not map perfectly onto one of these cases. In almost all cases, 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 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, depending on the level for which the primary purpose is intended to address. 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 and then consider approaches to aggregate the results to more general levels in the educational system.

Discussion

Throughout this paper, we have attempted to maintain a neutral tone while providing an overview of interim assessment systems. Now, however, we wish to use this discussion section to express some of our concerns with the current use of interim assessments and our hopes for the future direction of this work.

We first approached this paper from the perspective of promoting formative assessment. However, as we examined what is now in the field 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. When asked why we chose to focus on interim assessments rather than the purer formative assessment, our answer was simple: because states and districts are spending considerable resources to implement such systems. We recognize that to develop a strong formative assessment system as advocated by Black, Wiliam, Shepard, and others is difficult to do at a state level. 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 DOE personnel) is several steps removed from the classroom. While states can support professional development programs that teach our educators how to develop and use such tools, it would be helpful if states and districts could purchase a pre-existing system that supports formative and professional learning needs. In addition, states may have other requirements for an assessment program, such as to develop an early-warning system to identify students who are not on track to succeed and give them additional supports. Or, the states may wish to use these interim assessments as evaluation tools for different schools, instructional programs, or pedagogy. That is why we chose to define interim assessments as tools to evaluate students' knowledge and skills relative to a specific set of academic goals that are designed to inform decisions at a level above the classroom and to focus our discussion on the purposes and uses of such assessments.

That said, we are concerned that many of the commercially-available systems have moved far a field from what the research currently supports and those selling such system promise more than they can deliver. For example, these systems often lay claims to the research documenting the powerful effect of formative assessment on student learning when it is clear that the Black and Wiliam meta-analysis evaluated studies with formative assessments of very different character than essentially all current commercially-available interim assessment programs.

We believe that an interim assessment system that simply administers a series of mini-summative assessments is ineffective. We have seen several systems where shorter versions of the end-of-year assessment are given periodically. 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 towards the goal. One system we saw gave essentially the same form up to four times a year so the students actually saw the same or very similar items multiple times. We believe these systems are not the best use of money or instructional time. Testing students on material they have not yet been exposed to is useful only when one is unaware of what they have and have not been taught. Why test students in October on the material scheduled to be taught to them in February? What have you learned when you see they do better on the same items in March?

A good interim assessment can be an integral part of a state'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. 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. And, most importantly, they need to provide the next step—an intervention for correcting a misconception or teaching a missed lesson. 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. Is the purpose of the assessment simply to identify students not on track in order to place them in remedial instruction or provide additional supports, such as tutoring? Or is it the teachers' responsibility to interpret the results of the tests, determine areas of weakness, and correct them within the classroom? The requirements of the assessment would be different depending on the desired next steps.

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 is not available on the state large scale assessment or in the classroom through daily instructional activities. This additional information would include concepts such as allowing for the assessment of deeper learning that is challenging to address on the large scale assessment and more information about particular content strands than can be accomplished on typically short teacher-created tests. 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 them to tests in general? These interim assessments are an integral part of any comprehensive assessment system and should be considered as a piece of a whole and evaluated as such.

There are organizations trying to develop the types of thoughtful probes we discussed, but there are others that are simply trying to sell item banks and reporting systems. Our goal here was not to condemn all currently available products, but rather to provide a framework for the consumer to use in evaluating them. An additional purpose of this paper is to promote interest in further research in this area, and to that end, we conclude by describing our vision for this research.

Future Areas of Research Needed

This field is rich for further research. New studies funded by the U.S. Department of Education's Institute for Education Sciences (IES) are exploring areas that may serve to 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 was related 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 published.

We feel it is important to continue to examine how the use of interim assessments can help further student learning. In particular we see the need for research in the following areas:

1. Create a validity argument for how interim assessments lead to success on summative assessment and gather evidence to evaluate this argument. Choose several types of interim assessments and validate their uses.

    a. Is a predictive assessment truly predicting student performance on an end-of-year assessment 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. Is the use of instructional assessments improving instruction? Are there any unintended consequences?

    c. Are evaluative interim assessments effectively identifying differences in various pedagogies or instructional approaches?

2. Examine differential effects of interim assessments on students' intrinsic motivation to learn. Consider the theory that frequent assessments can diminish intrinsic motivation by shifting the effort and purpose from learning "to know" to learning so as "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. Examine the types of feedback that are most effective for improving student performance. Kluger and DiNisi (1996) 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. 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 types of professional development are linked to effective use of interim assessments? What is the best delivery system for this professional development?

References

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). Working inside the black box: assessment for learning in the classroom. London, UK: NFER-Nelson.

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.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy of Sciences.

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

Shepard, Lorrie. (June 2006). *Can Benchmark Assessments Be Formative?: Distinguishing Formative Assessment from Formative Program Evaluation*. Presented at the CCSSO Large Scale Assessment Conference, San Francisco, CA.

Scriven, Michael. (1967). The methodology of evaluation in perspectives of curriculum evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation,* 39-83. Chicago, IL: Rand McNally

Viadero, Debra. (2006). Cognition studies offer insights on academic tactics: U.S.-funded projects eye ways of helping students remember more material. *Education Week* August 30.

Wiliam, D. (2006). Assessment for learning: why, what and how. *Orbit: OISE/UT's magazine for schools, 36*(3).