# ARE WE WASTING OUR MONEY: WHAT DO WE KNOW ABOUT THE TECHNICAL QUALITY OF INTERIM ASSESSMENTS?

**Ying Li, University of Maryland**[1]

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

## Abstract

Increasing numbers of schools and districts have expressed interest in interim assessment systems to prepare for summative assessments and to improve teaching and learning. However, schools and districts are struggling to determine which interim assessment is most appropriate to their needs with so many available commercial interim assessments. Unfortunately, there is little work on this to help schools and districts to make their right choice about how to spend their money. Realizing the urgency of developing criteria that can describe or evaluate the quality of the interim assessment, this project tries to build an instrument with detailed criteria that school and district educators could use to analyze the quality and usefulness of the interim assessment.

## Introduction

The standards-based reform movement has resulted in the wide-spread use of assessments designed to measure students' performance at specific points in time—generally at the end of the school year—and to help instantiate the learning targets. In spite of our best hopes and efforts, these end-of-year tests provide very little instructionally useful information for educators particularly for the students who took the particular test. This is not because there is something "wrong" with these summative

accountability tests, rather that they were not designed to meet instructional purposes. Recognizing the inherent limitations of summative assessment, educators are looking for additional assessments to inform and monitor student learning during the year.

Many vendors are now selling what they call “benchmark,” “diagnostic,” “formative,” and/or “predictive” assessments with promises of improving student performance (Burch, in press). These systems often lay claim to the research documenting the powerful effect of formative assessment on student learning. However, the research in this area, including the seminal Black and Wiliam (1998) meta-analysis, evaluated formative assessments of a very different character than essentially all current commercially-available interim assessment programs. While there are some “truth in advertising” concerns about borrowing the research from a very different type of assessment, our concern is even more immediate.

The technical quality of these interim assessments is not well known and most educational leaders purchasing these assessments do not have the background to evaluate the technical materials even if the companies produced the appropriate technical documentation. In addition to the requirements for technical documentation articulated in the Standards for Educational and Psychological Testing, “the joint standards,” (AERA, APA, & NCME, 1999), we argue that developers of interim assessment systems have the additional responsibilities of creating technical summaries that can be understood by practitioner audiences. Recognizing the urgency of developing criteria that can describe or evaluate the quality of interim assessments, this paper reports our attempt to build an instrument with detailed criteria that school and district educators could use to analyze the quality and usefulness of the interim assessment.

## Framework

Perie, Marion, and Gong (in press) distinguished formative assessment, interim assessment and summative assessment in terms of the intended purposes, audience, use of the information, frequency of administration, scope of curricular coverage, and duration of cycle. They defined the interim assessment 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 purposes and intended uses, but the results of any interim assessment must be reported in a manner allowing aggregation across students, occasions, or concepts.*

Perie, et al. (in press) classified the multiple purposes of interim assessments into three major categories: instructional, evaluative and predictive. Instructional purposes involve using the test results to inform classroom teachers about current students' learning so that teachers can adjust their instruction to better meet student needs. Using the test information to analyze the curriculum, pedagogy, or other aspects of the educational program for the benefit of future students falls under evaluative purposes. Predictive purposes involve using the test results to estimate the students' performance on some distal outcome, most typically an end-of-year assessment, so that the students "not on track" to score proficient, for example, could be identified and hopefully be provided effective remediation.

The intent of interim assessments is to provide schools or districts more information than they can get through an end-of-year summative assessment. We are not certain of all of the implementation decisions driving these purchases. We suspect that many leaders either do not understand the distinction between these interim assessments and true formative assessment or see these interim assessments as more practically feasible to implement because (so the thinking might go) interim assessments do not require the same intensive level of professional development as formative assessments. In addition, these assessments provide policymakers with data that can be aggregated at the school or district level, offering a level of bureaucratic control not generally available with formative assessments. For these and other reasons, more and more schools and districts are purchasing interim assessments with the general goal of improving student achievement, at least as measured by end-of-year test. Therefore, having criteria to describe or evaluate the quality of the interim assessment and thus help districts (and schools) make decisions about buying assessments that can best meet the district's needs.

Some states have already made an effort to develop evaluation criteria for interim assessment to help in the decision-making for their districts and schools. South Carolina, for example, developed criteria that tests must meet in order to be placed on a "state approved" list from which districts can use state funds to purchase an assessment. This list has many technical requirements, such as table of specifications, description of field test sample, reliability indices, and standard error for each score point. These criteria are similar to what one might find in the joint standards (AERA, et al, 1999) or in the U.S. Department of Education's peer review criteria. On the other hand, New Mexico also developed an evaluation tool, but their criteria focused on the practical usability of the

assessment, such as the delivery format, frequency and duration of assessment, ease of assessment, and flexibility of administration. We see the merits of the approaches taken by these two states, but we felt like each was only hitting part of the target and there was still part of the evaluation not addressed by either of the two states.

Herman and Baker (2005) described six criteria that can help educators evaluate benchmark assessments. These include (1) aligning standards and benchmark assessments to ensure validity, (2) designing multiple item types to increase diagnostic value for instructional planning, (3) providing fair benchmark assessments for all students including English language learners and students with disabilities, (4) ensuring technical quality of the test reliability and validity, (5) providing user-friendly test results and guidance on interpreting and using the results to improve instruction, and (6) the feasibility and worthiness of the time and money that schools or districts will invest. Herman and Baker confirmed our initial thoughts on having criteria on both technical quality and test utility. In fact, they argued that the utility of the assessments for improving student learning should be the primary criterion, while the more traditional technical criteria are not as important as the utility. The criteria that comprise our evaluation tool are drawn from Herman and Baker (2005) as well as Perie, et al. (in press). The first phase of this project focuses on descriptive criteria; subsequent phases of the project are intended to be more evaluative.

## The Criteria

We use the following six criteria to describe and begin evaluating interim assessments:

1. Purpose and use of the test;

2. Test development and documentation;
3. Administration and inclusion;
4. Test scores and reports;
5. Test utility, and
6. Practicality and logistics

The first four criteria are consistent with the procedures of test formation, test delivery, and score interpretation and address questions of validity and reliability, while the fifth criterion describes how different stakeholders use the results of the testing program to fulfill specific educational goals. The final criterion describes aspects of the testing program that do not relate directly to the technical quality, but are important to users and include such features as ease of administration, availability of immediate feedback, and manageability of the data format. Selection of these criteria benefited from the work done by South Carolina and New Mexico as well as the earlier criteria developed by the Center for Assessment and CRESST and place an emphasis on both test validity and utility. We explore each of the criteria in more detail below. We have turned these criteria into a set of tables and checklists (where appropriate) for ease of use, but due to space limitations we present the framework and rationale here and not the full tables.

## Test Purpose and Use

Test developers and test users must first clarify the purpose(s) and use(s) of the test. The developers, as put forth in the joint standards below (AERA, et al., 1999), are clearly expected to articulate the purposes and legitimate uses of their assessments. We argue that while it is not explicitly stated in the joint standards, test users must be very clear about their reasons for purchasing an interim assessment system and the uses to

which they intend to put the results. Users should do this as specifically as possible so that they can best find a match between what they want and what the test developer intends with their assessment. Test users must be mindful of one of the truisms in educational measurement that a test promising to fulfill too many purposes tends not to fulfill any of the purposes very well.

- *Standard 3.6. The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers… (AERA, et al., p44)*
- *Standard 1.2. The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described. (AERA, et al., p17)*

While there are a plethora of potential purposes, we suggest that those evaluating interim assessments first use the Perie, et al. (in press) framework to categorize the purpose as instructional, evaluative, and predictive but then to be as specific as possible within these categories. Since purposes and uses are the foundation of a validity argument, we strongly suggest that users articulate the mechanisms and processes by which they think the use of an interim assessment will improve student learning (assuming that is one of the purposes). This abbreviated theory of action will help the users clarify their intended purposes and expected uses and determine whether the proposed interim assessment can help them fulfill their goals.

Similarly, test users need to be very clear about the target population they intend to assess and to ensure that the test was developed and piloted with these students in mind. For example, do the test takers include special education students and English language learners? And what content and grade level are included in the assessment system? These are all important considerations for those interested in using interim assessments.

## Test Development and Documentation

The test development criterion is organized into three levels: item, test, and multiple tests because all are critical for ensuring meaningful assessment experiences.

### Item Level Criteria

A test can be no better than the items comprising the test. This point cannot be overstated. We recommend that test vendors present fairly typical item characteristics such as difficulty, discrimination, differential item functioning (DIF), and alignment. But we strongly recommend that potential users convene committees of reviewers to evaluate the more qualitative item characteristics such as item quality, bias/sensitivity, accessibility, and especially to ensure that multiple item types are included in the assessment. While item quality might be a vague term, reviewers can be instructed—and most should have experience—to ensure that test questions are focused on meaningful content and processes and not on simplistic "gotcha" type questions.

A critical aspect of the item review involves the degree to which each item can be mapped to a content standard or learning objective. Unfortunately, alignment reviews can be subject to a confirmationist bias, therefore, an independent alignment between the test's item bank and the state content standards (or other learning targets) should be

conducted if possible. This type of alignment—ensuring that each item matches an appropriate content target—is only part of the picture. Items must also be presented at the levels of cognitive complexity called for in the standards (Webb, 1999). Further, while it is important to have items aligned to a particular content target, it is just as important to document the degree to which all content targets are represented by the assessment items. If items represent only part of the learning target, the validity of the assessment will likely be threatened by construct under-representation (Messick, 1989).

Our initial examination of current interim assessment systems finds that these systems are based almost exclusively on multiple-choice formats. While this unfortunately mimics the trend in state assessment designs, many have argued that if interim assessments are intended to be used of classroom instructional purposes, multiple item formats, particularly open-ended questions and even performance tasks, should be included to support instructional diagnostic as well as broaden and deepen the understanding of the concepts (Herman & Baker, 2005; Perie, et al., in press; Shepard, 2006).

There should be evidence that the item difficulty statistics were derived from a population similar to the target population. This is especially critical if classical item difficulty statistics (i.e., p-value) are presented, but even if item response theory (IRT) estimates of item difficulty are used, the vendor should document that the populations are similar or provide evidence that the invariance assumption holds. The range of item difficulty values should be appropriate for the intended purposes of the assessment. For example, if the purpose is to spread students out for selection purposes (not a likely purpose of an interim assessment), then a wide range of item difficulty is appropriate, but

if the purpose is to evaluate mastery, it would be more appropriate to include items focused around a particular mastery cutscore. Item discrimination is additional evidence of the appropriateness and effectiveness of the items for the target population.

Test Level Criteria

In addition to the individual item quality checked by the criteria above, the test level criteria is intended to ensure a set of items are selected to cover a certain breath and depth of content standards to form a valid and reliable test. The following characteristics should be documented to help users evaluate the quality of the overall interim tests and not just the individual items.

- Test specification (e.g. standards being tested, number of items per standards, item types)
- Documentation for Computerized Adaptive Tests (e.g. item selection algorithm, starting and termination conditions, exposure of items)
- Alignment to content standards or learning objectives
- Independent alignment to content standards or learning objectives
- Description of field test or item calibration sample
- Reliability and conditional standard error of measurement (SEM)
- Documentation on scoring procedures
- Information about the interpretation of test scores
- Information about score derivation

Several of these characteristics have been discussed already. We highlight a few of the other characteristics in the following paragraphs.

Test specifications or test blueprints provide an overall plan of the test. The test design identifies the standards or learning objectives being tested and the number of items for each standard. According to the joint standards:

> *The test specifications should define the content of the test, the proposed number of items, the item formats, the desired psychometric properties of the items, and the item and section arrangement. (AERA, et al., 1999, p. 43)*

Any interim assessment should include the test specifications in its documentation and these specifications should match what the users intend. Similarly, the importance of independent "two-way" alignment was discussed above. These two-way alignment studies document the degree to which items are aligned to content targets and intended content targets are measured appropriately.

Reliability refers to the consistency of the tests. It includes internal consistency in a single test, test-retest consistency across time, and alternative form consistency across forms. These reliability indices and associated standard error of measurement should be documented. While it is relatively easy to calculate and report reliability coefficients, it is much less straightforward to determine an appropriate level of reliability for interim assessments. The level of reliability is strongly correlated with the stakes associated with the decisions that the assessment is expected to support. Several have argued that reliability is not very important for formative assessments (e.g., Shepard, 2006) because decisions are very low stakes and can be adjusted on an almost daily basis. However, many interim assessments are used to group students into different instructional tracks for up to several months at a time (the time span between assessment events), which we

argue is at least a moderate stakes use. While we support the cautions against making important decisions on the basis of any single assessment (NRC, 1999), if such moderate stakes decisions are being made, we hope the assessments are at least quite reliable (e.g., r = .90),[1] but if the results are just being used for instructional purposes along with additional information, less reliable assessments (e.g., r=0.75 or 0.80) can still be useful.

Scoring procedures and scoring criteria should be documented. The scoring guidelines (rubrics) should be made explicit to users and examinees if open-ended items are included in the interim assessment (which we encourage). If the open-ended are locally scored, the assessment system should include appropriate training materials and exemplar paper. The following requirement from the joint standards (AERA, et al., 1999) should be applied to interim assessments:

> *Standard 3.22. Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical if tests can be scored locally. (AERA, et al., 1999, p47)*

Several interim assessment companies employ computer adaptive tests (CAT), but the specific item-selection algorithms and scoring routines are often hidden within an assessment "black box." Therefore, there CAT interim assessment vendors are responsible for documenting these algorithms, the scoring procedures, and the stopping

[1] This actually a much larger issue than we have space to discuss here because no matter how reliable the assessment, the assessment and interventions should be validated for this particular use (e.g., grouping).

criteria while controlling the item exposure. Again, the joint standards (AERA, et al., 1999) address the required documentation:

> *Standard 3.12. The rationale and supporting evidence for computerized adaptive tests should be documented. This documentation should include procedures used in selecting subsets of items for administration, in determining the starting point and termination conditions for the test, in scoring the test, and for controlling item exposure. (AERA, et al., p45)*

Multiple Test Level Criteria

The administrative frequency of interim assessments is greater than the frequency of end-of-year summative assessments but less than the frequency of classroom formative assessments. Usually, there are three or four administrations in an academic year. Depending upon the purposes of the assessment, the content relationship among the multiple forms in a year should be based on a clear rationale in order to serve the particular purpose. Unfortunately, we have seen evidence of less than thoughtful designs. For example, some predictive interim assessments actually administer essentially parallel form four times each year, each of which is based on the same blueprint as the end-of-year test. As Bob Linn noted sarcastically, "we used to call that sort of design test-retest reliability" (personal communication, April 2006). We suggest that these sorts of designs—even if the districts/school is interested in prediction—where the teachers first interpretative action is trying to figure out if the particular content had been taught, is not as useful has design that build a coherent coverage of the content through the year. Therefore, it is highly recommended that schools and districts check the content

relationship among multiple assessments to see whether it is consistent with the purposes of the tests and with the curriculum sequence used by the school/district.

If the design calls for the need to make inferences from the results across multiple administrations, the assessment vendor must provide documentation on equating procedures used to establish this comparability.

> *Standard 4.11. When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of equating functions. (AERA, et al., p. 57).*

## Administration and Inclusion

This criterion focuses on some fairly straightforward aspects of the test, but some very important components related to the accessibility of the assessment system. We are concerned that many interim assessments are developed and sold without the attention to accessibility (e.g., Universal Design) that we are starting to see with state end-of-year assessments. Therefore, test users should require information about how the development process attended to administration issues specific to special education and English language learner students. This should include information about design specifications, committee reviews, cognitive laboratory results, and pilot test information. There should also be documentation regarding the type of with accommodations allowed, the rationale for this particular set of accommodations, and empirical information—to the extent possible—about how these accommodations were piloted and/or performed operationally.

## Test Scores and Reports

Different types of scores provide educators with different perspectives of the test results. While all tests can typically produce a raw score (i.e., number correct out of number attempted), the way that score is transformed and presented can differ. For example, criterion-referenced scores convey what knowledge and skills have been mastered by a student; while norm-referenced scores describe the relative position of the student compared to the norm group. The appropriate types of scores derived from the test depend on the test purposes. For example, if the district/school is interested in measuring growth over time, some type of scale score that ensures comparability of meaning across administrations must be employed.

After determining the types of scores produced, it is important to examine the manner in which they are displayed. Score reports organize the derived score of the test in a meaningful way to convey the different information to different audiences. Usually, it is preferable to have both individual and aggregated reports for different stakeholders. The types of the score reports provide the information the stakeholders want conveyed by the test, which should also be consistent with the test purposes. For instance, for instructional purposes, an item analysis report, a strand level report and a criterion level report would provide useful information at the individual level as well as averaged across students for class level information for teachers. When the test is meant to serve a predictive purpose, the report should convey information about how results on the interim assessment are related to predictions for statewide end-of-year assessments. This prediction may be displayed as a predicted score on an end-of-year assessment or as a predicted outcome, such as the likelihood that each student would score in each possible performance level.

## Test Utility

Herman and Baker (2005) ranked utility as one of the leading criteria for benchmark assessments: “Utility represents the extent to which intended users find the test results meaningful and are able to use them to improve teaching and learning” (p. 8). This involves helping educators understand and interpret the test results to improve teaching and learning. Most importantly, this means that the interim assessment must fit within the educational system, particularly the curriculum, instructional, and support systems. The data must be provided so that local educators can turn these data into useable information, decisions, and instructional actions. This implies that the use of the assessment must be situated in a theory of action that describes how it fits within the system and how the results of the interim assessment will be used to improve teaching and learning. Evidence for the utility of the interim assessment system is often not gathered and reported by test vendors, in part because of the difficulty associated with conducting the appropriate studies. However, if interim assessment vendors tout the utility (instructional, evaluative, or predictive) or their systems, they should have empirical evidence to back up their claims.

Evaluating the potential utility of an interim assessment system prior to implementing on and the actual utility after it has been implemented is challenging because contexts and situations are always different and quite dynamic. More challenging, though is that most educational leaders are not well trained in critically evaluating educational research studies. Thus, this framework should serve as a basis for which leaders and others can more easily judge the research claims put forth for the various testing systems.

Practicality and Logistics

We have found through conversations with district leaders and others that certain aspects of the testing system that we, as measurement professionals, often take for granted are quite important to users. The ease of use such as the flexibility of the administration, ease of installation and maintenance, ease of use for students and teachers, and degree of ongoing technical support are all important considerations for district leaders especially if they are implementing these systems for large numbers of students. The speed with which results are returned appears to be a very important consideration for large numbers of educators. While the gratification of instant results is attractive, we suggest that users carefully weigh the perceived need for instant results with the uses to which the results will be put. For example, if the interim assessment is used primarily for evaluative purposes, then it is hard to see a need for instant results, at least to the point of outweighing other criteria. Actually, if the purported use is for instruction, especially if the assessment is administered only three or four times each year, it is still hard to argue for instant results considering that up to twelve or more weeks have elapsed since some of the instruction took place. A quick turnaround (e.g., one-two weeks) of results is important, but users should critically consider the need for instant results at the expense of some other design aspects (e.g., including open-ended tasks).

## An Example of Applying the Criteria

The next phase of the project involved testing the criteria against actual interim assessments. We turned the criteria into descriptive checklists, with space for open response comments, for ease of review. The purpose of this initial review was not to

judge the quality of the particular assessments, but to see if the criteria and checklists were complete, too restrictive, or too vague.

## Procedures

We wrote formal letters to seven testing companies, described our project of reviewing interim assessments, and asked if they would like to participate into our project and provide documentation such as technical manuals for our review. The letters were sent both electronically and in post mail. As we noted, only two responded quickly and with documentation: We refer to these two companies as ABC Assessment and QRS Testing, respectively. One of the companies even went so far as to host a two-hour introductory session to their system through WebEx (conducted July 3, 2008). Given the space limitations, we discuss, as an example, the results for ABC Assessments on the following pages. Again, we are illustrating the use of the tool and not evaluating particular assessment systems, at least at this point.

After receiving the documentation and interacting with the two testing companies, we applied the criteria to guide the review of the documents, described the tests as objectively as possible, and then modified the criteria to fit the tests whenever necessary. Typically, these modifications took the form of supplementing the criteria with additional options or categories. The results of this review are described below. If a category has been left blank (i.e., nothing was checked), that typically meant that we could not find the information in the documentation we received.

## Example: ABC Assessment[2]

### Part I: Test Purpose and Use

In reviewing the technical manuals of ABC, we found several differences among the three content areas: Early Literacy, Reading and Math. Therefore, instead of

reviewing and describing the three contents all together, we decided to separate them and evaluate them separately. As you can see below, we customized the table of test purposes for a better presentation of the three content areas still including both checklist and description.

| **Primary Purpose(s) of ABC** | | |
|---|---|---|
| **Early Literacy** | **Reading** | **Math** |
| ☒ Instructional Planning and Adjustment to Improve Learning<br>☐ Curriculum Instruction and Pedagogy Evaluation<br>☐ Statewide Assessment Prediction and Preparation | ☒ Instructional Planning and Adjustment to Improve Learning<br>☐ Curriculum Instruction and Pedagogy Evaluation<br>☐ Statewide Assessment Prediction and Preparation | ☒ Instructional Planning and Adjustment to Improve Learning<br>☐ Curriculum Instruction and Pedagogy Evaluation<br>☐ Statewide Assessment Prediction and Preparation |
| ABC's Early Literacy determines children's mastery of literacy concepts that are required for future success in reading; the results will be used to plan instruction and intervention. | ABC Reading estimates the students' reading comprehension using instructional reading levels, accesses reading achievement relative to national norms, and tracks students' growth at aggregated level. | ABC Math estimates students' instructional math levels relative to national norms, and tracks students growth at aggregated level. |

## Part II: Test Development and Documentation

### A. Item Level Criteria

As discussed earlier, a criteria table with checklist and description is applied below, encompassing both general requirements and a specified description of evidence provided by the ABC developers. A checkmark implies that the required criterion was provided in a technical manual supplied by the vendor. It makes no judgment, however, as to the quality of the measure.

The description below is a brief introduction of the evidence that ABC provides to meet the criteria. Reviewing the documents on ABC (or other testing products)

demonstrates the differential emphasis on various criteria. However, just including a paragraph describing all the criteria in the checklist does not reflect the amount and quality of the evidence provided for each criterion. Therefore, the different quality of information provided by the testing company should be reflected in a more qualitative/judgmental review process. This could be accomplished by creating rubrics to evaluate the information presented.

| Item Level |
| --- |
| Checklist<br>☒ Item difficulty and item discrimination<br>☒ Item aligned to content standards or learning objectives<br>☒ Item content fairness: DIF statistics (gender and ethnicity)<br>☒ Item bias and sensitivity review<br>☒ Item edited for spelling, grammar and usage conventions, and for cuing and item writing principles<br>☒ Multiple item types such as multiple choice, and open-ended.<br>**Description**<br>ABC Early Literacy\Reading\Math has Classical Test Theory item difficulties (p-values) and item discriminations (e.g. point-biserial correlations). Since they use Rasch model, they also have IRT item difficulties. Content in ABC Early Literacy and Math are divided into several domains or strands with clustered skills or objectives within the domains and strands; items are written according to the domains/strands and clustered skills/objectives within the domains/strands. ABC item writing and editing appeared to try to minimize cultural loading, gender stereotyping, and ethnic bias. |

The alignment criterion presented above is a good example regarding the quality of the gathered information. The ABC technical manuals provide a lot of information on alignment. It divides Early Literacy into seven sub-domains and divides Math into eight strands. Existing items are categorized according to the sub-domains or strands, and new items are written based upon the sub-domains or strands. While independent alignment reports are almost always preferably to studies conducted by the vendor, this fine-grained

blueprint makes it possible for educational leaders to check the believability of the alignment reports with their own content experts without having to conduct a full alignment study.

We argue that any tool should expand on the information regarding multiple item types for each of the assessment. This information can be presented in a table or in text with examples, but should be as descriptive as possible (e.g., 4-choice, multiple-choice items, three-step short constructed response tasks).

B. Test Level Criteria

While the quality of the individual items is critical, how the items are comprised into a test form is an essential criterion. Again, the criteria table with checklist and description at the test level was completed after reviewing the ABC documents. It is worth mentioning that the first two criteria are essentially mutually exclusive: test specifications are more applicable for a non-CAT design, while item selection algorithm is more applicable for CAT. Therefore, having at least one of the two criteria checked is appropriate.

| Test Level |
| --- |
| Checklist<br>☐ Test specifications (e.g. standards being tested, number of items per standards, item types)<br>☒ Documentation for Computerized Adaptive Tests (e.g. item selection algorithm, starting and termination conditions, exposure of items)<br>☒ Alignment to content standards or learning objectives<br>☐ Independent alignment to content standards or learning objectives<br>☒ Description of field test or item calibration sample ( representative to the target population)<br>☒ Reliability and SEM<br>☒ Documentation on scoring procedures<br>☒ Standard error for each score point<br>☒ Information about the interpretation of test scores<br>☒ Information about score derivation<br>**Description**<br>ABC Early Literacy, Reading and Math are computerized adaptive tests. Items are selected to match the student's current ability level and grade level as well as to represent the broad coverage in content. Alignment studies were conducted with several state content standards. Several reliability indices (e.g. generic reliability, split-half reliability, and test-retest reliability) were calculated. A proprietary Maximum-Likelihood IRT estimation is used for scoring scale score as soon as the student has at least one item correct and one item incorrect, and associated conditional SEM is also calculated. Other scores such as Percentile Ranks and Grade Equivalents are derived from Scale Scores. |

As mentioned previously, beyond the criteria table with checklist and description, additional information, such as what is provided in the tables below, should be added to provide additional information on some criteria in the checklist. As an example, using the documents provided by ABC Assessment, additional information on the CAT design, item calibration sample, and reliability indices are provided in the tables below.

| **ABC Design of Assessment** | | |
|---|---|---|
| With computerized adaptive tests, ABC assessments select items at levels of difficulty that most match the student's current ability level. Students are motivated because the items are neither too difficult nor too easy for them. If the student has taken a test in previous six moths, the appropriate starting point is based on his or her previous test score information. | | |
| **Early Literacy** | **Reading** | **Math** |
| The adaptive test is fixed length with 25 items with two or more items from each of either five or seven domains. There are two parts in the 25-item test. The first part includes 16 shorter items in terms of their audio time and students' response time. The second part includes 9 longer items in terms of audio time and response time. Items in both parts are subject to content constraints to ensure the broad content coverage. | The adaptive test has a fixed length of 25 items. At grade levels 3 and above, there are 20 vocabulary-in-context items and five authentic test passage items. At grade levels K-2, there are only vocabulary-in-context items. | The adaptive test is fixed length with 24 items. The first 16 items are selected from the Numeration Concepts and Computation Processes strands evenly, which are recognized as foundation and basics. The rest items are selected from the other six strands balancing the strand coverage and the students' grade level. |

| **Item Calibration Sample** | | |
|---|---|---|
| To ensure the sample used for calibration study is representative of the target population, the sampling consisted of all US schools, stratified on three variables: geographic region, school size, and socioeconomic status. The comparison of the distributions between the sample and the population at various levels of the three variables was documented. | | |
| **Early Literacy** | **Reading** | **Math** |
| ABC Early Literacy calibration sample in Fall 2000 included 32,493 students from 308 schools. | ABC Reading 2.0 calibration sample in Spring 1998 included 27,807 students from 287 schools. | ABC Math 2.0 calibration sample in Spring 2001 included 44,939 students from 261 schools from 45 out of 50 states. |

| ABC Reliability | | |
|---|---|---|
| **Early Literacy** | **Reading** | **Math** |
| With total sample of 9146, the generic reliability is 0.92, the split-half reliability is 0.91, and the retest reliability is 0.86. | With norming sample of 29169 in Spring 1999, both generic reliability and split-half reliability is 0.96; test-retest reliability with sample of 2095 is 0.94; alternative forms reliability is 0.95 with the alternate form sample of 4551. | With norming sample of 29228, generic reliability is 0.947 and the split-half reliability is 0.944; alternative reliability with sample of 7389 is 0.908. |

C. Multiple Test Level Criteria

Interim assessments, as we noted in the beginning of the paper, are designed to be administered multiple times each year. Therefore, the nature of the system of multiple tests is crucial to the evaluation of interim assessments. We begin with the standard tables on multiple test level criteria, and then provide an extra table on correlational evidence to show more completely the evidence provided by ABC Assessment.

| Multiple Test Level |
|---|
| Checklist<br>☒ Multiple administrations (usually 3 or 4) through out an academic year<br>☒ Description of the relationships of contents and standards among the multiple administrations across a year.<br>☒ Documentations for comparability across forms (Equating procedures)<br>☒ Validity evidence on correlations among internal and external assessments.<br><br>**Description**<br>Since ABC Early Literacy, Reading and Math are computerized adaptive tests, items for tests in the same content are selected from the same item bank with calibrated items. Since items are in the same scale, the test consists of the items are also in a common scale. With the instruction and multiple administrations through out the year, we are expecting students' ability is increasing and they are able to get more difficulty items correct in later administrations. Validity evidence is provided as the correlations between the ABC Early Literacy\Reading\Math and other external assessments. |

| **Correlational evidence** | | |
|---|---|---|
| **Early Literacy** | **Reading** | **Math** |
| Concurrent validity is calculated as the correlation between the ABC Early Literacy/ Reading/Math and other external tests administered within a two-month time period. Predictive validity is calculated as the correlation between ABC Early Literacy/ Reading/Math and the criterion test administered more than two months later. | | |
| In Spring 2001, within grade concurrent validity coefficients were 0.64, 0.68, 0.52 and 0.57 for grades K-3 respectively.<br>The within grade average predictive validity coefficient for pre-K-3 were 0.57, 0.52, 0.62, 0.67 and 0.77 respectively. | In Spring 1999, the within grade average concurrent validity coefficient varied from 0.71 to 0.81 for grade 1-6 and from 0.64 to 0.75 for grade 7-12.<br>The within grade average predictive validity varied from 0.68 to 0.82 for grade 1-6 and varied from 0.81 to 0.86 for grad 7-12. | In Spring 2002, the within grade average concurrent validity coefficient varied from 0.63 to 0.71 for grade 1-6 and from 0.47 to 0.73 for grade 7-12.<br>The within average predictive validity coefficient varied from 0.55 to 0.73 for grade 1-6 and from 0.75 to 0.80 for grade 7-12. |

## Part III: Test Format and Administration

Testing experts tend not to consider things like administration issues at the same level of importance as things such as item quality and form design. However, our conservations with local districts leaders suggests that administration and practical issues are just as important as any psychometric criteria measurement experts might consider. As measurement specialists, we, not surprisingly, would rather see district and state leaders focus on the measurement criteria, but we are convinced that reporting and evaluating administration considerations is very important to our intended audience.

| Test Format and Administration | | |
|---|---|---|
| Administration Format<br>☐ Paper and pencil<br>☐ Consumable<br>☐ Non-consumable<br>☐ Computer based test (CBT)<br>☐ Paper and pencil or CBT<br>☒ Computer adaptive test (CAT)<br>☐ May be administered in any of the formats above<br>☐ Other – please describe<br>---------- | | |
| Test Accessibility | | |
| Accommodation | Special Forms | |
| ☐ Provided to special education students<br>☐ Provided to English language learners | ☐ Provided to special education students<br>☐ Provided to English language learners | |
| Instructional | Evaluative | Predictive |
| ☒ Customized forms<br>☒ Flexible date and location for administration<br>☒ High speed of results | ☐ Customized forms<br>☐ Flexible date and location for administration<br>☐ Moderate speed of results | ☐ Standardized forms<br>☐ Standardized administration procedures<br>☐ Moderate speed of results |

## Part IV Score Reports

Score reports should be designed to translate the assessment results into actionable information. Educational leaders should have a well-articulated sense of how they intend to use the assessment results and therefore, in what form the results should be presented. The information presented below for the ABC Assessment system describes the types of scores and information from the assessment. Of course, there is a difference between simply reporting scores and establishing the validity of the inferences from such scores, but this information presented below is intended to at least provide an initial look.

**Types of Scores**

**Raw Score**

☐ Providing a summary of student mastery of the items on the test

**Scale Score (SS)**

☒ Providing equivalent scores to make all tests comparable.

**Criterion-Referenced Score**

☒ Strand Level Score for Early Literacy and Math
Providing scores for each sub-domain or strand level

☒ Early Literacy Classification
Providing cut scale scores to identify different levels of literacy: Emergent Reader, Transitional Reader, and Probable Readers.

☒ Instructional Reading Level (IRL)
Providing an estimate of the most appropriate level of reading material for instruction.

☒ Zone of Proximal Development (ZPD)
Defining the readability range from which students should be selecting books in order to ensure sufficient comprehension and therefore achieve optimal growth in reading skills without experiencing frustration.

**Norm-Referenced Score**

☒ Percentile Rank (PR)
Providing the percentage of scores in the norm group at or below a particular score

☒ Grade Equivalent (GE)
Indicating the grade placement of students for whom a particular score is typical.

☒ Normal Curve Equivalent Score (NCE)
Providing the ability scale with mean of 50 and standard deviation of 21.06 resulting in having a set of equal interval scores ranging from 0 to 99.

| Instructional | Evaluative | Predictive |
|---|---|---|
| ☐ Raw Score<br>☒ Scale Score<br>☒ Criterion Score<br>☒ Grade Equivalent Score<br>☒ *Instructional Reading Level*<br>☒ *Zone of Proximal Development* | ☐ Criterion-referenced score | ☐ Scale Score<br>☐ Performance Level |

The typical raw scores, scale scores, criterion-referenced scores and norm-referenced scores, other derived scores from the ABC assessments are presented within the appropriate categories. Additionally, ABC derived its own Instructional Reading Level and Zone of Proximal Development scores that are intended to directly inform instruction. While these types of scores sound very attractive, especially if the district's main purpose is informing instruction, district leaders and/or other evaluators need to determine the meaningfulness of such scores. ABC also reports sub-domain or strand level scores for Early Literacy and Math, which are also classified into criterion-referenced score. The criterion table can be adjusted according to the need for information presentation.

**Types of Reports**

☒ **Criterion-Referenced Report**

Reporting the performance objectives that have been mastered and not yet mastered at individual and aggregated level.

- *Student Diagnostic Report*
- *Class Diagnostic Report*

☒ **Norm-Referenced Report**

Reporting the relative position of an individual, a class or school in the norm group.

- *Score Distribution Report*

☒ **Multi-Test Report**

Reporting multiple results from previous assessment, monitor progress of students' achievement, and identify in risk students on statewide assessment.

- *Growth Report*
- *Progress Monitor Report*

| **Instructional** | **Evaluative** | **Predictive** |
|---|---|---|
| ☒ *Student Diagnostic Report* | ☐ *Class Diagnostic Report* | ☐ *Progress Monitor Report* |
| ☒ *Class Diagnostic Report* | ☐ *Score Distribution Report* | ☐ *Growth Report* |
| ☒ *Score Distribution Report* | | |
| ☒ *Progress Monitor Report* | | |
| ☒ *Growth Report* | | |

Extra tables below detail the reports by content areas and highlight descriptive information about the multiplicity of reports available through the ABC assessment system. The *Student Diagnostic*, *Growth* and *Progress Monitor Reports* are designed to improve instruction planning, but might also be used for evaluative purposes. The information about the available reports appears to confirm that the types of report are consistent with the test purpose we identified at the first stage of the evaluation.

| **Types of Reports in Detail** | | |
|---|---|---|
| **Early Literacy** | **Reading** | **Math** |
| *Class Diagnostic Report* provides the eight domain scores for and skill scores under its domain for a class.<br><br>*Score Distribution Report* provides the domain score distribution and skill sets within each domain score distribution for a class.<br><br>*Growth Report* provides each student's GP, SS, domain score, literacy classification, as well as the average scores for a class across tests.<br><br>*Progress Monitor Report* provides the averaged GP, SS and domain scores for all the test results of a class over a school year.<br><br>*Student Diagnostic Report* provides the eight domain scores for and skill scores under its domain for a student. | *Student Diagnostic Report* provides SS, GE, PR, PR range, IRL and ZPD for a student.<br><br>*Growth Report* provides each student's GP, SS, GE, PR, PR range, NCE, IRL as well as the average scores for a class across tests.<br><br>*Progress Monitor Report* provides the averaged SS, GE, PR, PR range, NCE, IRL, and ZPD for all the test results of a class over a school year. | *Student Diagnostic Report* provides SS, GE, PR, PR range, NCE and recommended accelerated Math library and skill levels under its strand for a student.<br><br>*Growth Report* provides each student's GP, SS, GE, PR, PR range, NCE as well as the average scores for a class across tests.<br><br>*Progress Monitor Report* provides the averaged SS, GE, PR, PR range, and NCE for previous test results of a class over a school year. |

## Part V. Test Utility

The utility criteria are intended to get at how the assessment system supports subsequent decisions about instructional or programmatic plans. We outlined two main criteria related to utility. ABC addressed the first criterion in this section by including instructional suggestions on the Diagnostic reports, but did not appear to meeting the

second criterion because there was no professional development includes as part of the assessment system.

**Test Utility**

**Instructional Strategies/Implications**

☒ Provided based on student performance at individual or class level from *ABC Diagnostic Report*.

**Professional training**

☐ Included to help teachers understand and interpret the data

☐ Provided to help teachers diagnosis students' strengths and weakness to plan instruction accordingly.

## Part VI. Practicality and Logistics

As noted earlier, the practical and logistical issues appear to be very important to test consumers. As seen from the checklist below, ABC Assessment appears to be quite easy to use and addresses many of the logistical concerns that a district leader may have.

**Practicality and Logistics**

**Flexibility of Administration**

☒ Can be administered at the group or individual level with computers.

**Ease of Administration**

☒ Can be achieved by minimal training of administrators and standardizing the administration procedures.

**Technical assistance**

☒ Provided online or by telephone in a timely manner to support the use by teachers, school and district administrators.

**Accessibility**

☐ Available to all students including English language learners and students with disabilities.

**Manageable Data format**

☒ Can be easily aggregated or disaggregated based upon the needs of teachers, school or district administrators.

**Immediate feedback**

☒ Can be provided via computer or other means as soon as the test being administered.

**Periodic assessments**
☒ Provided with multiple assessments through out the academic year.

## Discussion

We expect many schools and districts will still ask "which interim assessment is the best" thinking the answer will allow them to buy the best test. However, a better question that should be asked as "which interim assessment is the best **for my school or district for these particular purposes**?" Educational leaders need to ensure that the purposes for which they want to use an interim assessment should be the same as the purposes for which the interim assessment was developed. One suggestion for using this tool is for educational leaders to first complete the tool by indicating what they would like to see in an interim assessment product. These ideal set of responses then can be compared to the description—based on the tool—for each interim assessment system the district is considering. Of course, the first priority should be to ensure that the purpose of the interim assessment matches the purpose for which the school or district wants to use the test.

After confirming the purpose consistency, the school or district should check other properties of the test using the criteria such as *Test Development and Documentation*, *Test Form and Administration*, and *Test Score and Report* to see if those properties served consistently with the test purposes. For example, if district wants to use the interim assessment as an indicator of statewide assessment, the district should pick an interim assessment with predictive purpose; items and the whole test should be aligned with state content standards; correlations between interim assessments and statewide

assessments should be relatively high; the scores report should focus on predicted proficiency levels on statewide assessment to identify in risk students, and provide strategies for intervention to help students meet the state requirement.

## Limitations and Subsequent Studies

This study was not trying to provide a rank-ordering or a "Consumer Reports" rating of interim assessment quality, because we recognized that such a simplistic result would obfuscate the complexity of the interim assessment process. In reality such a ranking system would have to be so conditional as to make it unwieldy. For example, such a rating system might lead to results such as, "If your primary purpose is for instructional purposes, you want standard-referenced reports, your curriculum emphasizes reform-based mathematics, you have ample computers, and you require professional development for your teachers, then assessment XYZ might be the best for your district." In spite of this limitation, we still think that we or anyone else attempting to review interim assessment products needs to be significantly more evaluative than the descriptive information provided with the current version our tool. Most district and school assessment leaders do not have the assessment knowledge to critically evaluate the claims from the various test publishers and would welcome the type of evaluative information we envision including with the next version of this tool. Our challenge, in the next phase of this project, will be to be fair to the test publishers, yet judgmental enough to help district leaders at least narrow the field of potential suitable interim assessment products.

Another limitation of this project as that we were able to review only two interim assessments systems. The main reason for using the tests was to try out and modify the criteria in order to make the criteria more generalizable. Therefore, the more tests

included the more useful we could make the criteria. Another reason for including actual assessments at this stage of the project was to show how the criteria could be flexibly applied to describe or evaluate different commercial tests. Therefore, subsequent projects will include more tests to allow us to better refine the criteria.

The small research team reviewing the tests was another limitation of this study. To ensure accuracy and correct interpretation, the criteria should be applied by more than one researcher or small research teams and perhaps even reviewed by the test developer. Ideally, the reviews of the assessments should be conducted by a larger group of people to ensure the inter-subjective agreement of the review or evaluation.

## References

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

Burch, P. (in press) Benchmark assessments: Who is selling them? Who is buying them? And why? Peabody Journal of Education.

Herman, J.L., &Baker, E.L. (2005). Making Benchmark Testing Work. *Assessment to Promote Learning, 63,* 48-54.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education, Macmillan Publishing

National Research Council (1999). *High stakes: testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

New Mexico Public Education Department Division of Assessment and Accountability. (2006). *Consumer Guide to Formative Assessments.*

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

Perie, M., Marion, S.F., & Gong, B. (2008, in press). Moving Towards a Comprehensive Assessment System: A Framework for Considering Interim Assessments. *Educational Measurement: Issues and Practice.*

Shepard, L.A.. (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.

South Carolina State Board of Education. (2006). *2007-08 Formative Assessment List for South Carolina: Stage One Submission Cover Sheet.*

Webb, N. L. (1999, August). *Alignment of science and mathematics standards and assessments in four states:* Research Monograph No. 18. Washington, DC: Council of Chief State School Officers.

[1] The first author would like to thank the National Center for the Improvement of Educational Assessment for funding this project in summer, 2008.

[2] All information about the ABC assessments was obtained from technical manuals developed by ABC.