Descriptive Criteria for Interim Assessments

Ying Li
University of Maryland College Park

Brian Gong, Scott Marion, & Marianne Perie
National Center for the Improvement of Educational Assessment

AERA, San Diego, April 16, 2009

Outline

- Background
- Purpose
- Theoretical framework
- The criteria and instrument
- An example: applying the instrument
- Discussions and Limitations

Background

- There are a lot of commercially available interim/benchmark assessments.

- Schools and districts are spending thousands of dollars purchasing them, hoping to use them to improve teaching and learning.

- There are a lack of standards on the quality of interim assessments.

Purpose of the study

- Building an instrument with detailed criteria that school and district educators could use to analyze the quality and usefulness of the interim assessments.

Theoretical Framework

Perie, Marion, and Gong (2008) classified

the purposes of interim assessments into three

major categories:

- Instructional
- Evaluative
- Predictive

Theoretical Framework

■ **South Carolina criteria list**

table of specifications, description of field test sample, reliability indices, and standard error for each score point.

■ **New Mexico criteria list**

delivery format, assessment frequency, and flexibility of administration.

Theoretical Framework

Herman and Baker (2005) described six criteria that can help educators evaluate interim assessments.

1. aligning standards and interim assessments to ensure validity,

2. designing multiple item types to increase diagnostic value for instructional planning,

3. providing fair interim 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.

The Criteria and Instrument

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.

An example:
Applying the criteria instrument

- Contacted 7 testing companies for reviewing
- 2 responded quickly

- Apply the criteria to guide the review
- Modify the criteria to be generalizable
- Accommodate the criteria instrument to maximize the information extracted from the review of the test

1. Purpose and Use of the Test

Primary Purpose of the ABC Assessment		
Early Literacy	**Reading**	**Math**
☒ Instructional Planning and Adjustment to Improve Learning ☐ Curriculum Instruction and Pedagogy Evaluation ☐ Statewide Assessment Prediction and Preparation	☒ Instructional Planning and Adjustment to Improve Learning ☐ Curriculum Instruction and Pedagogy Evaluation ☐ Statewide Assessment Prediction and Preparation	☒ Instructional Planning and Adjustment to Improve Learning ☐ Curriculum Instruction and Pedagogy Evaluation ☐ Statewide Assessment Prediction and Preparation
ABC 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, assesses 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.

2. Test Development and Documentation

- Item Level
- Test Level
- Multiple Test Level

Item Documentation

Item Level

Checklist

☒ Item difficulty

☒ Item discrimination

☒ Linked to content standards or learning objectives

☒ Content Fairness: DIF statistics (gender and ethnicity)

☒ Evidence of bias and sensitivity review

☒ Evidence that each item was edited for spelling, grammar and usage conventions, and for cuing and item writing principles

☒ Documentations on item calibration sample

Description

ABC Early Literacy\Reading\Math has Classical Test Theory item difficulties (the percent of correct response out of the total response) and item discriminations (e.g. point-biserial correlation between item score and the total score). Since they use Rarsh model, they also have item difficulties from IRT. Contents in ABC Early Literacy and Math are detailing 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 tried to minimize cultural loading, gender stereotyping, and ethnic bias. Item calibration was conducted using the sample representative to the national population.

Test Documentation

Test Level

Checklist

☐ Test specifications (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 (representative to the target   population)

☒ Reliability and SEM

☒ Documentation on scoring procedures

☒ Standard error for each score point

☒ Information about the interpretation of test scores

☒ Information about score derivation

Multiple Tests Documentation

Multiple Test Level

Checklist

- ☒ Multiple administrations (3 or 4) through out an academic year
- ☒ Description of the relationships of contents and standards among the multiple administrations across a year.
- ☒ Documentations for comparability across forms (Equating procedures)
- ☒ Validity evidence on correlations among internal and external assessments.

Description

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.

3. Administration and Inclusion

Administration and Inclusion

Administration Format (check one)

☐ *Paper and pencil*

☐ *Computer based test (CBT)*

☐ *Paper and pencil or CBT*

☒ *Computer adaptive test (CAT)*

☐ *May be administered in any of the formats above*

☐ *Other – please describe --*

☒ *Frequency of administrations per academic year---------4----------*

Accessibility	
Accommodation	**Special Forms**
☐ Provided to special education students. ☐ Provided to English language learner students.	☐ Provided to special education students. ☐ Provided to English language learner students.

Instructional	**Evaluative**	**Predictive**
☒ Customization of test forms ☒ Flexible date and location of test administration ☒ High speed of results	☐ Customization of test forms ☐ Flexible date and location of test administration ☐ Moderate speed of results	☐ Standardization of test forms ☐ Standardized administration procedures ☐ Moderate speed of results

4. Test Score and Report

Types of Scores

☐**Raw Score**
☒**Scale Score (SS)**
  Criterion-Referenced Score
☒Strand Level Score for Early Literacy and Math
☒Early Literacy Classification: identify Emergent, Transitional, or Probable Readers.
☒Instructional Reading Level (IRL) : Provide an estimate of the most appropriate level of reading material for instruction.
☒Zone of Proximal Development (ZPD): Define the readability range from which students should be selecting books in order to ensure sufficient comprehension
  Norm-Referenced Score
☒Percentile Rank (PR)
☒Grade Equivalent (GE)
☒Normal Curve Equivalent Score (NCE)

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

4. Test Score and Report

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* ☒*Score Distribution Report* ☒*Progress Monitor Report* ☒*Growth Report*	☐*Class Diagnostic Report* ☐*Score Distribution Report*	☐*Progress Monitor Report* ☐*Growth Report*

5. Test Utility

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.

6. Practicality and Logistics

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.
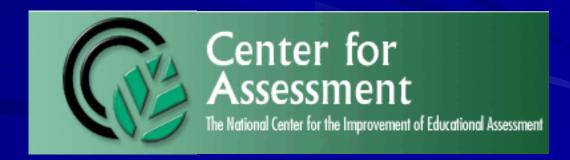
Discussions for educators

- Ask question "which interim assessment is the best for my school/district for these particular purposes?"

- Complete the instrument by indicating what you would like to see in an interim assessment product at the very beginning!

Limitations and subsequent studies

- Review perspective:

     descriptive vs. evaluative

- Inclusion of interim assessments:

     two vs. more

- Research team:

     small vs. large

Acknowledgement

- 2008 summer internship opportunity at National Center for the Improvement of Educational Assessment, Dover, NH

- My great mentors Scott Marian, Brian Gong, and Marianne Perie.

Center for
Assessment
The National Center for the Improvement of Educational Assessment

Questions?

- Ying Li

  yingli@umd.edu

  University of Maryland College Park

  Department of Measurement, Statistics and Evaluation

  Paper is available at http://www.nciea.org/