easyCBM Test Item Development: Merging Researcher and Practitioner Expertise for Student Improvement

P. Shawn Irvin

Behavioral Research & Teaching

College of Education – UO

BRT behavioral research & teaching

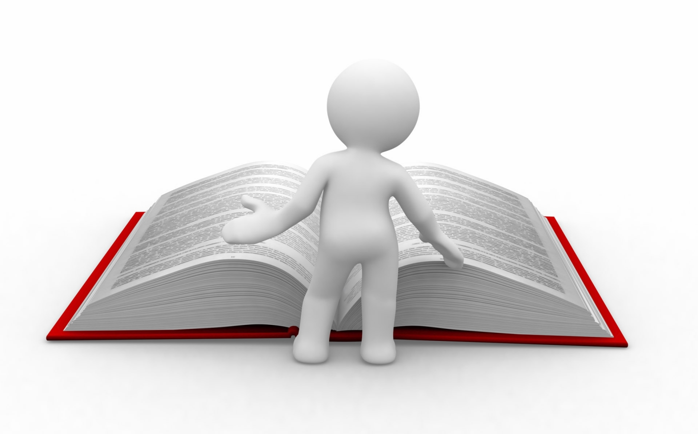
Road Map

- Foundations of Item Development
- Item Development Process
  - Item Writing
  - Editing and Review
  - Graphics/Audio
  - Standards Alignment/Qualit
  - Piloting and Scaling
- Test Form Creation/Equating
- Ongoing Research

BRT
behavioral research & teaching

Foundations

- Accountability
- Standards-based Instructi



- Research
  - English Language Arts and *The Big 5* (NICHD, 2000)
    - phonemic awareness, alphabetic principles, fluency, vocabulary, and comprehension
  - Mathematics
    - numeracy, operations, reasoning skillsets, etc.

BRT behavioral research & teaching

Foundations cont.

- Developing technically adequate interim-formative assessment measures to:
  - Screen for risk, gauge status, and monitor change (McConnell, McEvoy, & Priest, 2002)
  - Establish valid/parsimonious factor structures (Justice, Invernizzi, Geller, Sullivan, & Welsch, 2005)
- easyCBM
  - Reading (early/emergent) and Math
  - RTI framework to improve student learning outcomes through school-wide improvement

Item Development Process

1. Item Writing (P, R)
2. Editing and Review (P, R)
3. Graphics/Audio (P, R)
4. Standards Alignment/Quality (P, R)
5. Piloting and Scaling (P, S, R)

Key stakeholders: Practitioners (P); Students (S); Researchers (R)

1. Item Writing

Recruitment of item writers/reviewers

- Representative sample of practitioner experts
- Experience/expertise (i.e., conte years of experience, position hel education level)
- General/Special educators
- e.g., K-5 CCSS Math: 18 individuals, 16 with Masters, ave of 14 yrs experience (r = 3-32), GenEd/SPED

1. Item Writing cont.

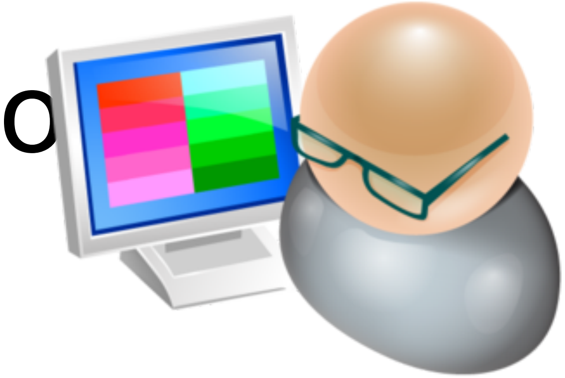
Training of item writers (and reviewers)

- Half-day, webinar/in-person sessions
- High-quality items, according to principles of:
  - Universal Design for Assessment (UDA; precise construct targets, accessible to diverse popns, lack of bias) (Thompson, Johnstone, & Thurlow, 2002)
  - Research-based construction (e.g., Haladyna, 2002; 2004)
  - Logistics (e.g., written >> operational, alignment, style, formatting, templates)
  - Examples/non-examples of quality items
  - Targeted practice

2. Editing and Review

- Multi-stage and iterative
  - Concurrent with item writing
  - Subsequent to item writing, concurrent with graphics/audio
- Employing both in- and out-of-house content and test development experts

3. Graphics and Audio Development



- Professional graphic artists hired to create graphics according to UDA
- In-house audio for most items
  - Students with diverse learning/assessment needs
  - English and Spanish audio created for items/ measures (e.g., NCTM/CCSS)

4. Item Alignment/Quality

Alignment/quality addressed two-fold:

- Before and during writing/review
- Formal alignment research studies using the Distributed Item Review (DIR)
  - Content/instructional experts judge test items as student would see them in the operational measure
  - Address issues of bias, sensitivity, accessibility
  - Feedback for further improvement (i.e., items revised or discarded)

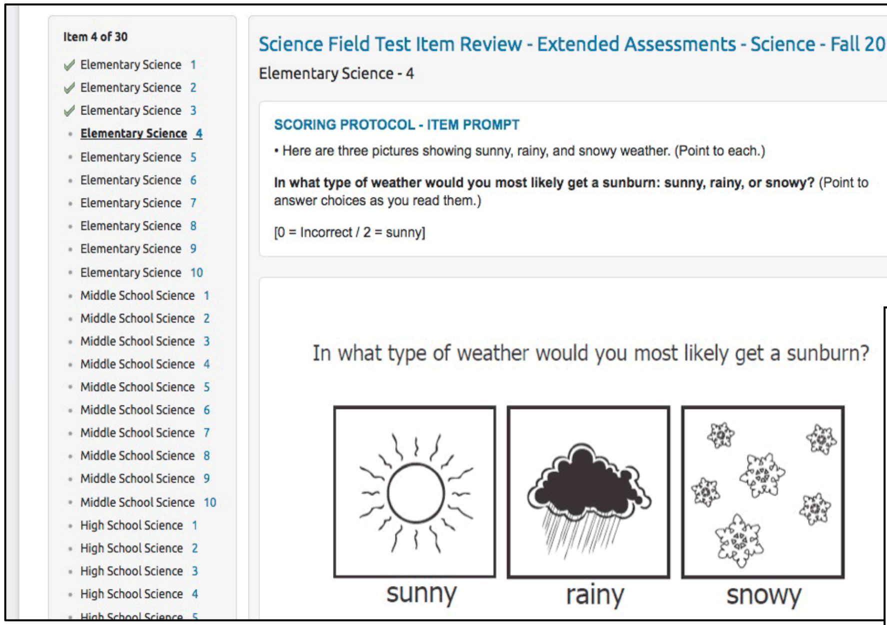
BRT behavioral research & teaching

4. Item Alignment/Quality cont.

Distributed Item Review (DIR; BRT, 2013)

- Distribute test items to expert users across appropriate geography (e.g., national, state)
- Examine dimensions of item quality (e.g., alignment/linkage, bias, sensitivity, accessibility)
- Essential features: diverse item types, pertinent support resources, organized assignment to participants, review contexts (e.g., development, review/improvement).

4. Item Alignment/Quality cont.



Item 4 of 30

- ✓ Elementary Science 1
- ✓ Elementary Science 2
- ✓ Elementary Science 3
- Elementary Science 4
- Elementary Science 5
- Elementary Science 6
- Elementary Science 7
- Elementary Science 8
- Elementary Science 9
- Elementary Science 10
- Middle School Science 1
- Middle School Science 2
- Middle School Science 3
- Middle School Science 4
- Middle School Science 5
- Middle School Science 6
- Middle School Science 7
- Middle School Science 8
- Middle School Science 9
- Middle School Science 10
- High School Science 1
- High School Science 2
- High School Science 3
- High School Science 4
- High School Science 5

Science Field Test Item Review - Extended Assessments - Science - Fall 2011

Elementary Science - 4

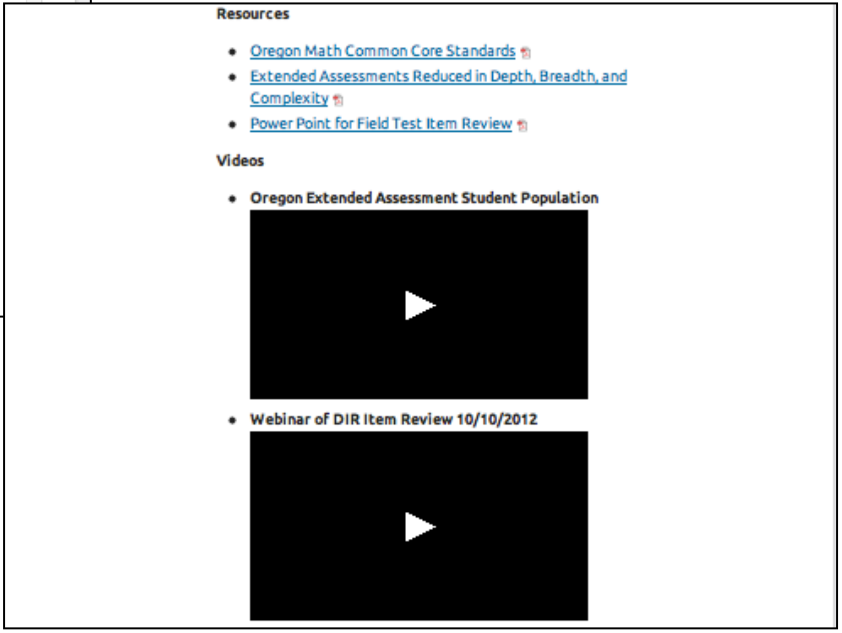
SCORING PROTOCOL - ITEM PROMPT

• Here are three pictures showing sunny, rainy, and snowy weather. (Point to each.)

In what type of weather would you most likely get a sunburn: sunny, rainy, or snowy? (Point to answer choices as you read them.)

[0 = Incorrect / 2 = sunny]

In what type of weather would you most likely get a sunburn?



Resources

- Oregon Math Common Core Standards
- Extended Assessments Reduced in Depth, Breadth, and Complexity
- Power Point for Field Test Item Review

Videos

- **Oregon Extended Assessment Student Population**
- **Webinar of DIR Item Review 10/10/2012**

BRT
behavioral research & teaching

4. Item Alignment/Quality cont.

- 4,245 assessment items
- ELA, Math, Science – easyCBM/OR alternate assessment
- 121 SPEDucators
- 110 GenEducators
- 38 states
- Multi-purpose studies (alignment, b-s-a)
- More on the horizon!!! ☺

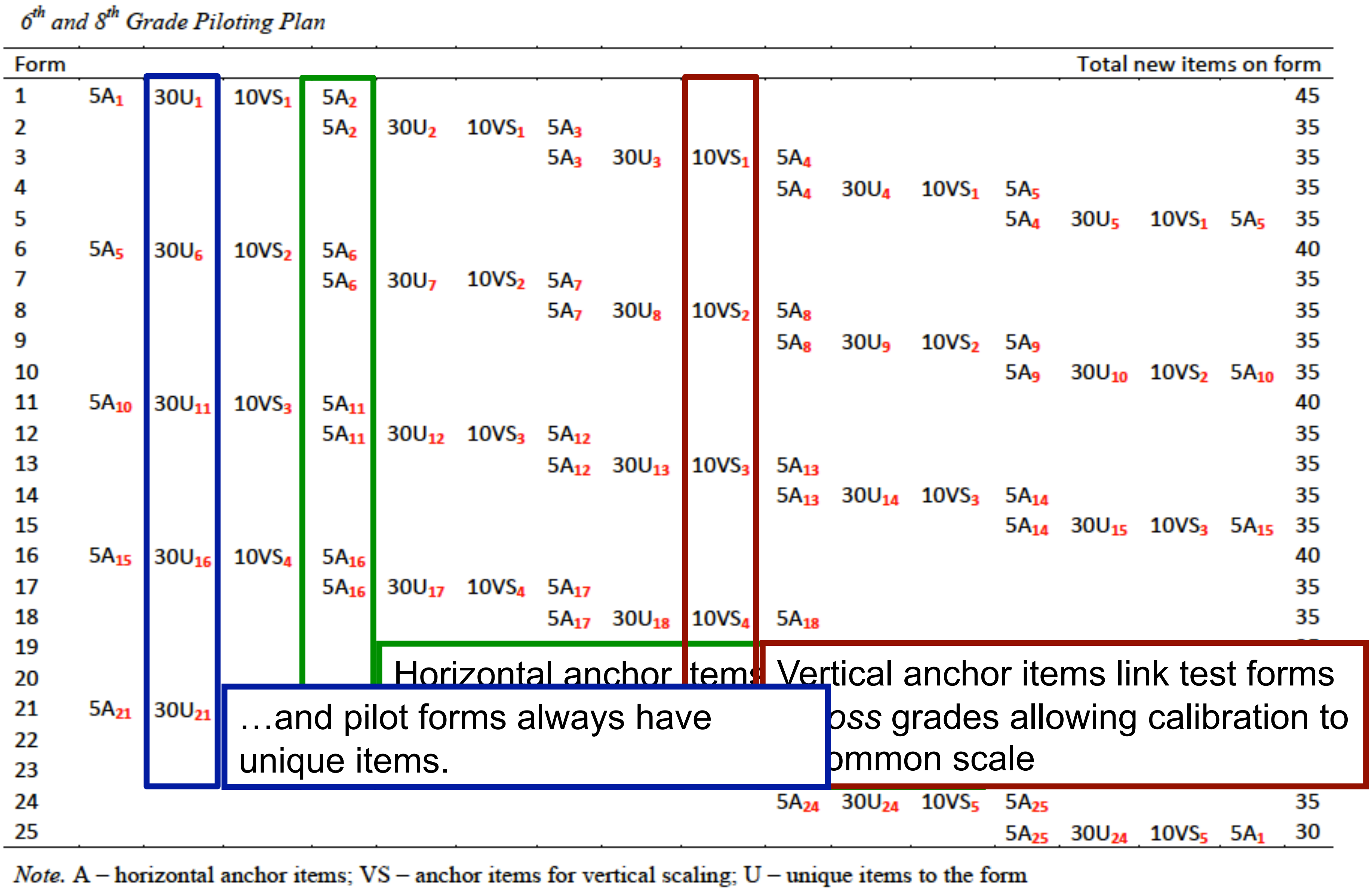
Year	Grade	Subject	Items	SPED Sensitivity	Gen-Ed Content	Total	Purpose
2012	K-8	Rdg/Math	61	6	5	11	SA

Note. Abbreviations are as follows: ELA = English Language Arts; RC = Reading Comprehension; EL = Early Literacy; Rdg = Reading; SPED = Special Education; Gen-Ed = General Education; B-S = Bias-Sensitivity; SA = Standards Alignment.

5. Item Piloting and Scaling

Students of varying ability take multiple test items in carefully designed pilot forms to analyze the quality of item functioning and to calibrate items (from a given measure) to a common scale. ***This makes it so that item difficulty is directly comparable within (and sometimes across) grades.***

Anderson, Irvin, Patarapichayatham, Alonzo, & Tindal, G. (2012)



6th and 8th Grade Piloting Plan

Form																	Total new items on form
1	$5A_{1}$	$30U_{1}$	$10VS_{1}$	$5A_{2}$													45
2				$5A_{2}$	$30U_{2}$	$10VS_{1}$	$5A_{3}$										35
3							$5A_{3}$	$30U_{3}$	$10VS_{1}$	$5A_{4}$							35
4										$5A_{4}$	$30U_{4}$	$10VS_{1}$	$5A_{5}$				35
5													$5A_{4}$	$30U_{5}$	$10VS_{1}$	$5A_{5}$	35
6	$5A_{5}$	$30U_{6}$	$10VS_{2}$	$5A_{6}$													40
7				$5A_{6}$	$30U_{7}$	$10VS_{2}$	$5A_{7}$										35
8							$5A_{7}$	$30U_{8}$	$10VS_{2}$	$5A_{8}$							35
9										$5A_{8}$	$30U_{9}$	$10VS_{2}$	$5A_{9}$				35
10													$5A_{9}$	$30U_{10}$	$10VS_{2}$	$5A_{10}$	35
11	$5A_{10}$	$30U_{11}$	$10VS_{3}$	$5A_{11}$													40
12				$5A_{11}$	$30U_{12}$	$10VS_{3}$	$5A_{12}$										35
13							$5A_{12}$	$30U_{13}$	$10VS_{3}$	$5A_{13}$							35
14										$5A_{13}$	$30U_{14}$	$10VS_{3}$	$5A_{14}$				35
15													$5A_{14}$	$30U_{15}$	$10VS_{3}$	$5A_{15}$	35
16	$5A_{15}$	$30U_{16}$	$10VS_{4}$	$5A_{16}$													40
17				$5A_{16}$	$30U_{17}$	$10VS_{4}$	$5A_{17}$										35
18							$5A_{17}$	$30U_{18}$	$10VS_{4}$	$5A_{18}$							35
19																	
20																	
21	$5A_{21}$	$30U_{21}$															
22																	
23																	
24										$5A_{24}$	$30U_{24}$	$10VS_{5}$	$5A_{25}$				35
25													$5A_{25}$	$30U_{24}$	$10VS_{5}$	$5A_{1}$	30

Note. A – horizontal anchor items; VS – anchor items for vertical scaling; U – unique items to the form

Horizontal anchor items

…and pilot forms always have unique items.

Vertical anchor items link test forms across grades allowing calibration to common scale

5. Item Piloting and Scaling cont.

- Items analyzed using *item response theory* (IRT)
- Item-level stats, pre-defined criteria (e.g., Wright and Linacre, 1994)
  - *Mean square outfit* – indicator of item performance given item difficulty and student ability
  - *Discrimination* – indicator of relation b/t item and test success, i.e., Does the item yield unique info? Does the item distinguish b/t students with higher-lower performance?
- Poorly functioning items edited/discarded

Test Form Construction/ Equating

- Standard (domain) representation
- Range of difficulty – sensitivity at “lower” end of the performance spectrum
- Alternate forms of appx equivalent difficulty (status *and* growth, teacher/ school DM)
- Nuances to reduce construct-irrelevant variance (e.g., domain clustering, ramping difficulty)

Ongoing Research and Collaboration

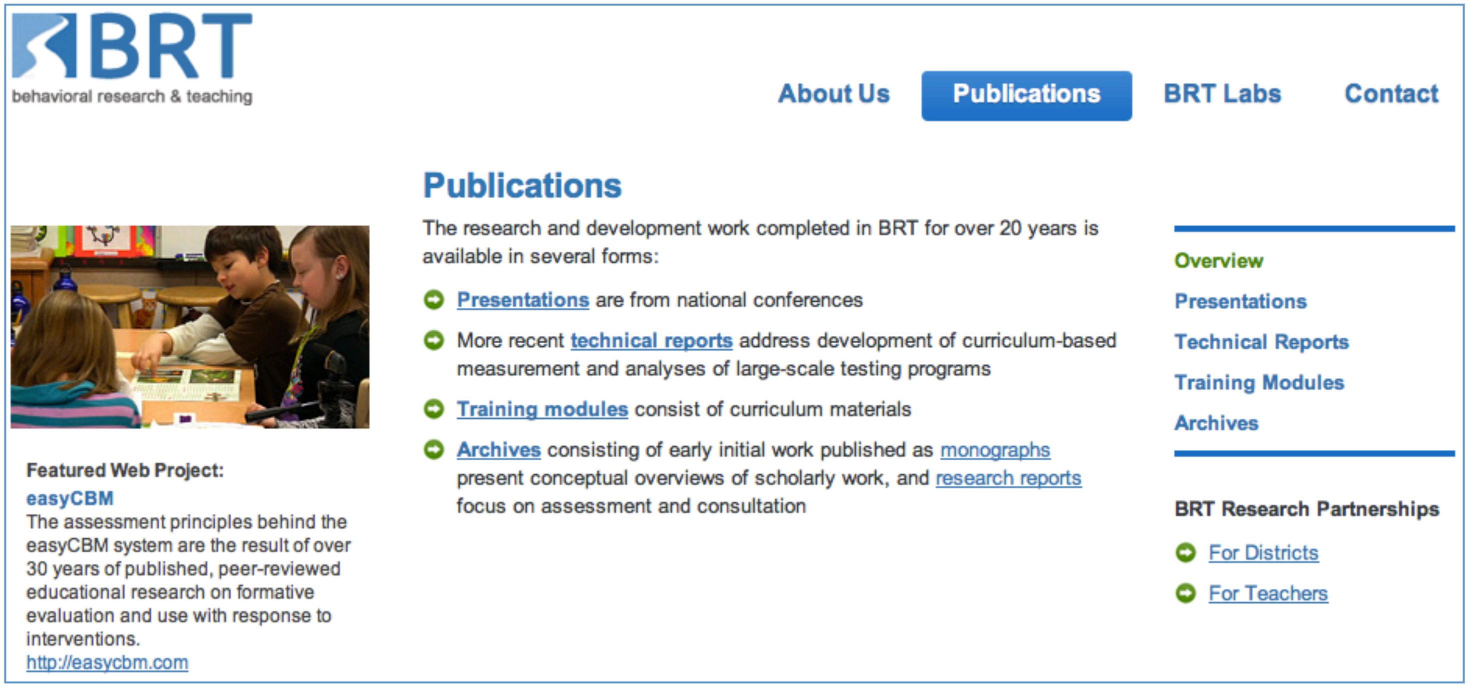
- Reliability
- Validity
- Cross-validation and Diagnostic Efficiency
- National and Regional Norms
- Test Use and Associated Teacher Decision-making

BRT
behavioral research & teaching

Thank you! Questions?

http://www.brtprojects.org

http://easyCBM.com



Publications

The research and development work completed in BRT for over 20 years is available in several forms:

- Presentations are from national conferences
- More recent technical reports address development of curriculum-based measurement and analyses of large-scale testing programs
- Training modules consist of curriculum materials
- Archives consisting of early initial work published as monographs present conceptual overviews of scholarly work, and research reports focus on assessment and consultation

BRT Research Partnerships

- For Districts
- For Teachers

Featured Web Project:

easyCBM

The assessment principles behind the easyCBM system are the result of over 30 years of published, peer-reviewed educational research on formative evaluation and use with response to interventions.

http://easycbm.com