easyCBM Iterative Measurement Development: CCSS Math

Behavioral Research & Teaching

Outline

- Original item development
  - Item writing
  - Scaling & test form creation
- Reliability
  - Initial screen
  - Revisions Made
  - Current Reliability
- Criterion Validity Evidence
- Future Directions

BRT
behavioral research & teaching

Item Development

Test Blueprint

- Written to specifically align with CCSS Math Standards
- Three response options
- "Oversampling" of Items (~50%)
- Universal Design
  - Minimal, simple, and direct language
  - Line art
  - White Space

Item Writers

- Master/mentor model
  - 5 teacher leads: intensive in-house training
  - 18 item writers: trained and monitored by teacher leads
- All item writers were middle school mathematics teachers (GenEd & SpEd)
- Master trainers were district math specialists, or had extensive teaching experience

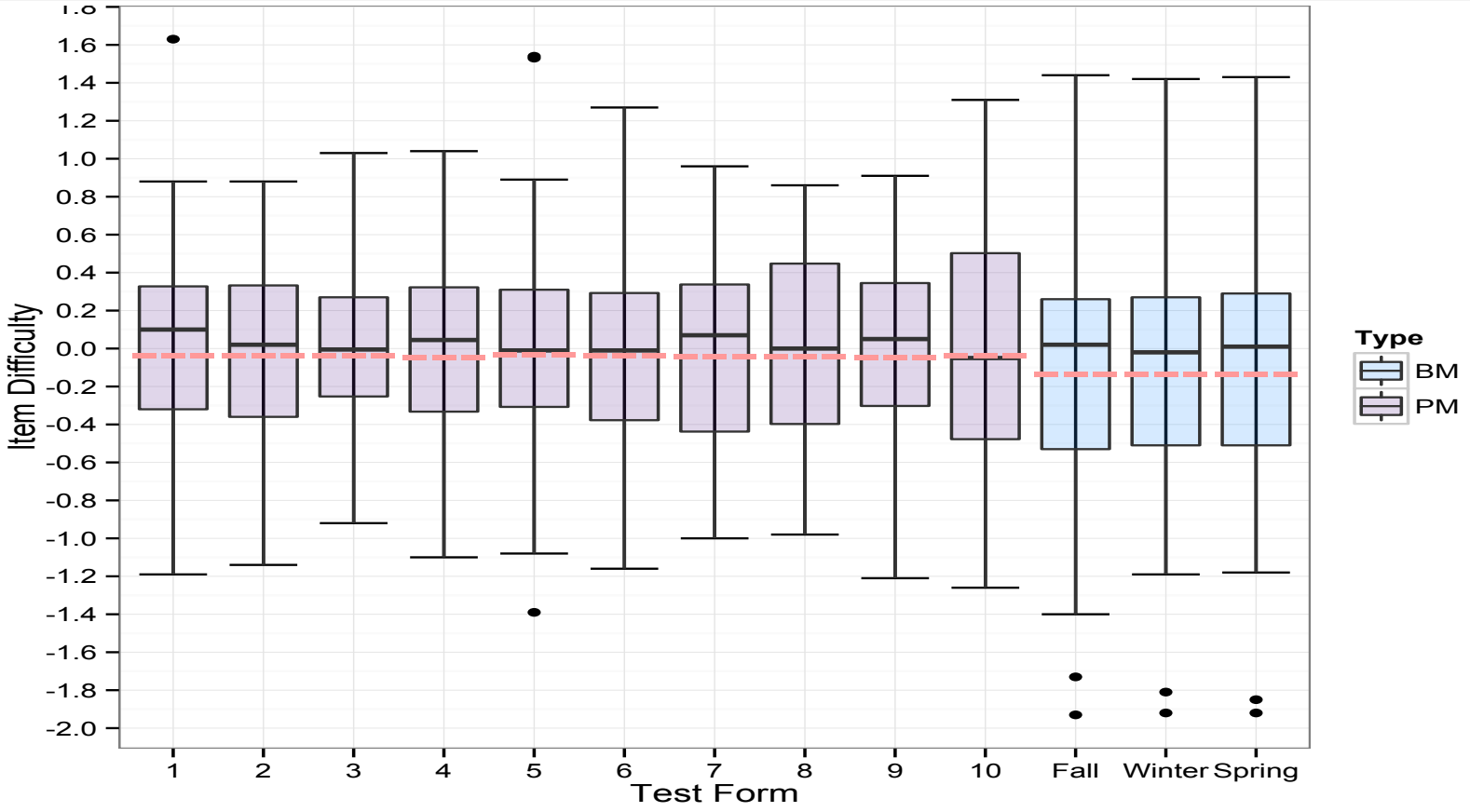
BRT
behavioral research & teaching

Item Screening

- Minimum of 200 students from across the country responded to each pilot item.
- Items calibrated with item response theory
  - Common scale (all item difficulties directly comparable across grades)
- Items removed from consideration if:
  - Pilot data suggested poor functioning
  - Alignment data suggested the item did not measure the intended standard

BRT
behavioral research & teaching

Form Creation



BRT
behavioral research & teaching

Investigating Test Functioning: Reliability

- **Reliability** is

  "concerned solely with how the scores resulting from a measurement procedure would be expected to vary across replications of that procedure" (Haertel, 2006)

- Separate from validity (but is a prerequisite)

Internal Consistency	Alternate Form
Test-retest	Split-half
Generalizability Theory	Etc.

BRT
behavioral research & teaching

Initial Investigations into Reliability

- Sample included ~1,000 students in Oregon, with Five CCSS test forms per grade investigated
- Initial screening of data suggested some items weren't working well
- Items were removed, and reliability was adequate, but still less than ideal

BRT
behavioral research & teaching

Initial Investigations into Reliability

- Sample included ~1,000 students in Oregon, with Five CCSS test forms per grade investigated
- Initial screening of data suggested some items weren't working well
- Items were removed, and reliability was adequate, but still less than ideal

BRT
behavioral research & teaching

Brief Dive into Results

Grade 6 Test Form Point-Biserial Correlations

Item	Form				
	6	7	8	9	10
1	.277**	.386**	.473**	.269**	.342**
2	.201**	.382**	.283**	.263**	.461**
3	.534**	.358**	.383**	.404**	0.126
4	.617**	.199**	.343**	.366**	.201**
5	.220**	.415**	.198**	.265**	.343**
6	.508**	.431**	.266**	.231**	.301**
7	.480**	.255**	.467**	.395**	.240**
8	.404**	.156*	.319**	.343**	.237**
9	.313**	-0.003	0.137	.268**	0.124
10	.256**	.188**	0.007	.144*	0.081
11	.241**	.416**	.261**	.266**	.442**
12	.530**	.388**	.396**	.487**	.349**
13	.471**	.373**	.404**	0.063	.377**
14	.409**	.335**	.441**	.410**	.323**
15	.248**	.227**	.512**	.407**	.267**
16	.338**	.405**	.253**	.351**	.282**
17	.402**	.385**	.497**	.463**	.445**
18	.346**	.395**	.315**	.424**	.342**
19	.337**	.219**	.520**	.195**	.386**
20	.478**	.252**	.148*	.284**	.409**
21	.322**	.288**	.290**	.510**	.259**
22	0.042	.472**	.314**	.250**	.420**
23	.400**	.518**	.479**	.174*	.154*
24	.228**	0.138	.156*	.245**	0.048
25	.281**	.258**	.507**	.184*	.154*

Note. Items displayed in red font were removed prior to subsequent analyses.

* $p < .05$

** $p < .01$

Brief Dive into Results

Cronbach's Alpha Reliability Coefficients

		Alpha			
		Day 1		Day 2	
Grade	Form	Full model	Reduced Model	Full model	Reduced Model
6	6	.70	.72	.77	.79
6	7	.66	.69	.67	.72
6	8	.69	.76	.74	.78
6	9	.65	.70	.61	.65
6	10	.57	.63	.59	.69

BRT
behavioral research & teaching

Brief Dive into Results

Test-Retest Reliability Coefficients

Grade	Form	Test-Retest *r*
6	6	.69
6	7	.69
6	8	.71
6	9	.73
6	10	.61

BRT
behavioral research & teaching

Brief Dive into Results

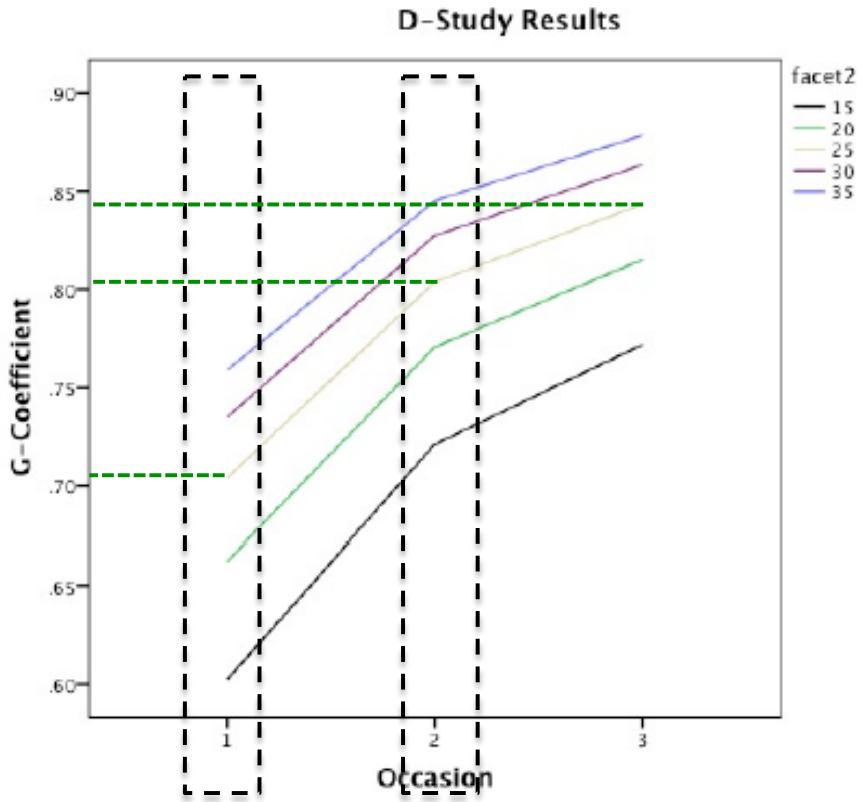
Grade 6: Alternate Form Reliability Coefficients

Test form	6	7	8	9	10	n
6	-	.432	.601	.597	.465	.662
7	.376	-	.819	.641	.760	.572
8	.721	.525	-	.813	.744	.591
9	.492	.720	.426	-	.752	.522
10	.197	.784	.553	.728	-	.549
n	.806	.491	.665	.743	.569	-

Note. Coefficients below the diagonal represent correlations from the first testing occasion, while the coefficients above the diagonal represent correlations from the second testing occasion occurring one week later.

BRT
behavioral research & teaching

Brief Dive into Results



BRT
behavioral research & teaching

Overall Takeaway: Not good enough

- What to do? Revise.
- Items were noticeably more difficult than NCTM
  - Included 5 NCTM items rated as aligning with the CCSS
- **Removed** 5 poorest functioning items from each form
- Conducted additional pilot
- Replaced items with those that pilot data suggest function better.

BRT
behavioral research & teaching

What effect did the changes have?

- Cronbach's alpha now > .9 for all measures investigated.
- Split-half reliability > .8
- Overall takeaway – it looks like it worked!

BRT
behavioral research & teaching

Now they're reliable, are they valid?

- **Validity** is

  "An overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (Messick, 1995)

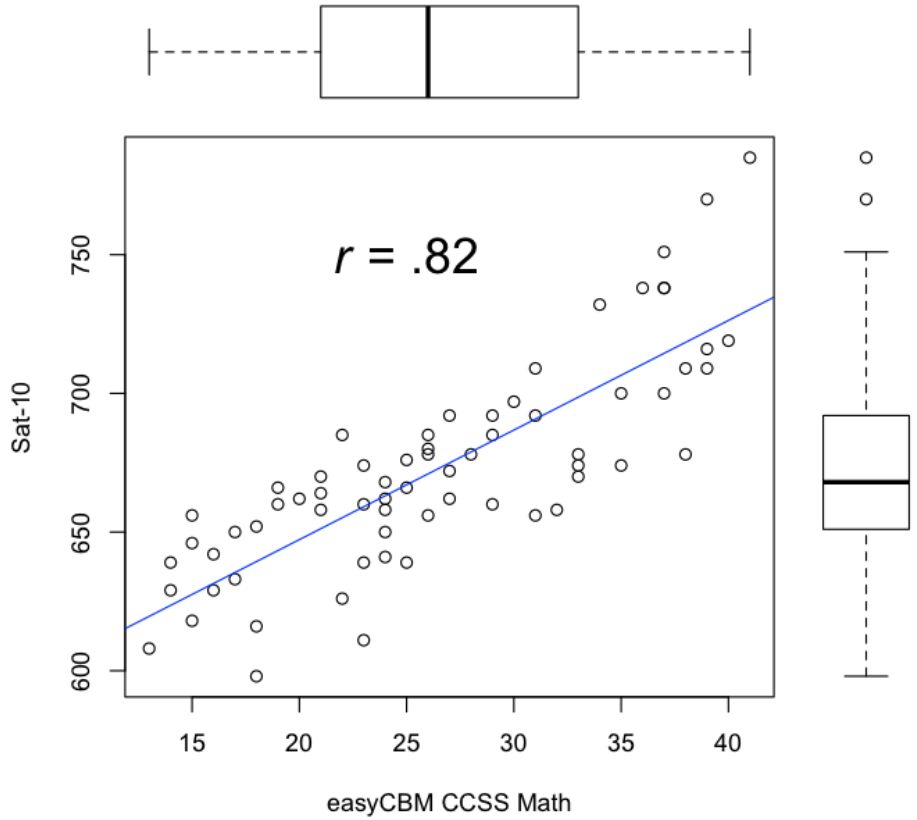
- Basically – does the test ***actually*** measure what it ***says*** it measures
- Not a property of the test

BRT
behavioral research & teaching

Preliminary Investigations

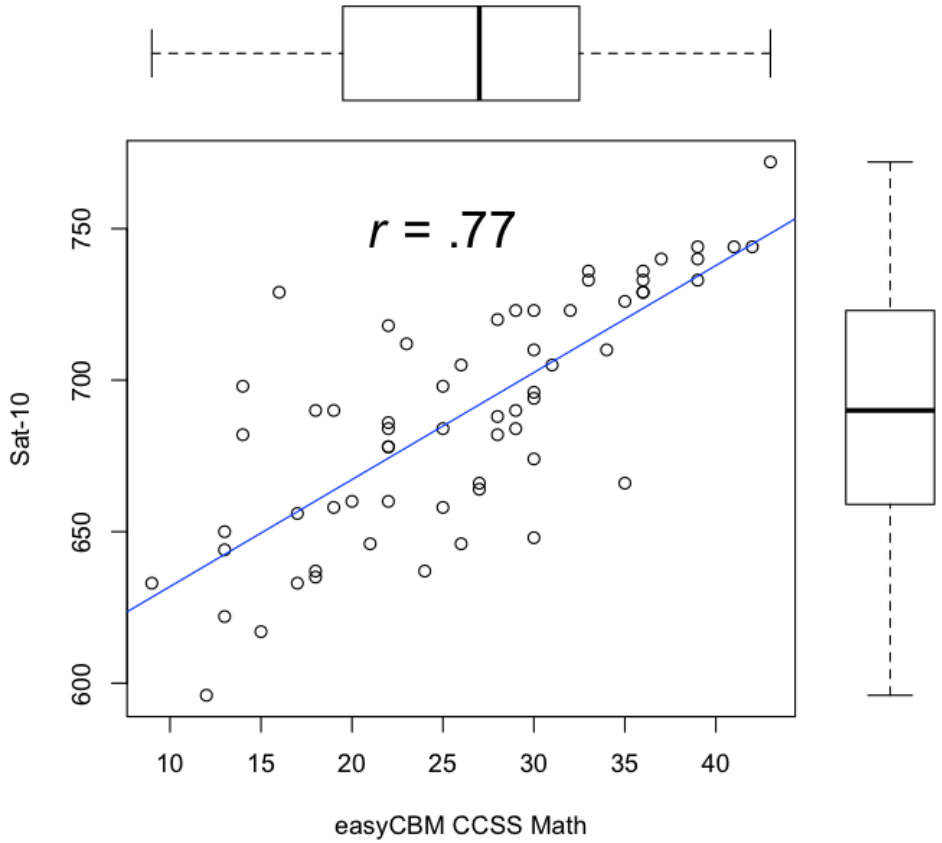
- Criterion validity
  - How well do students' scores on easyCBM "go along" with scores from a criterion measure
  - Note. Measures are not designed to be exactly the same, but scores should at least correlate.
- Sample
  - 65 students in Oregon in each grade.

Criterion Validity Results: Grade 6



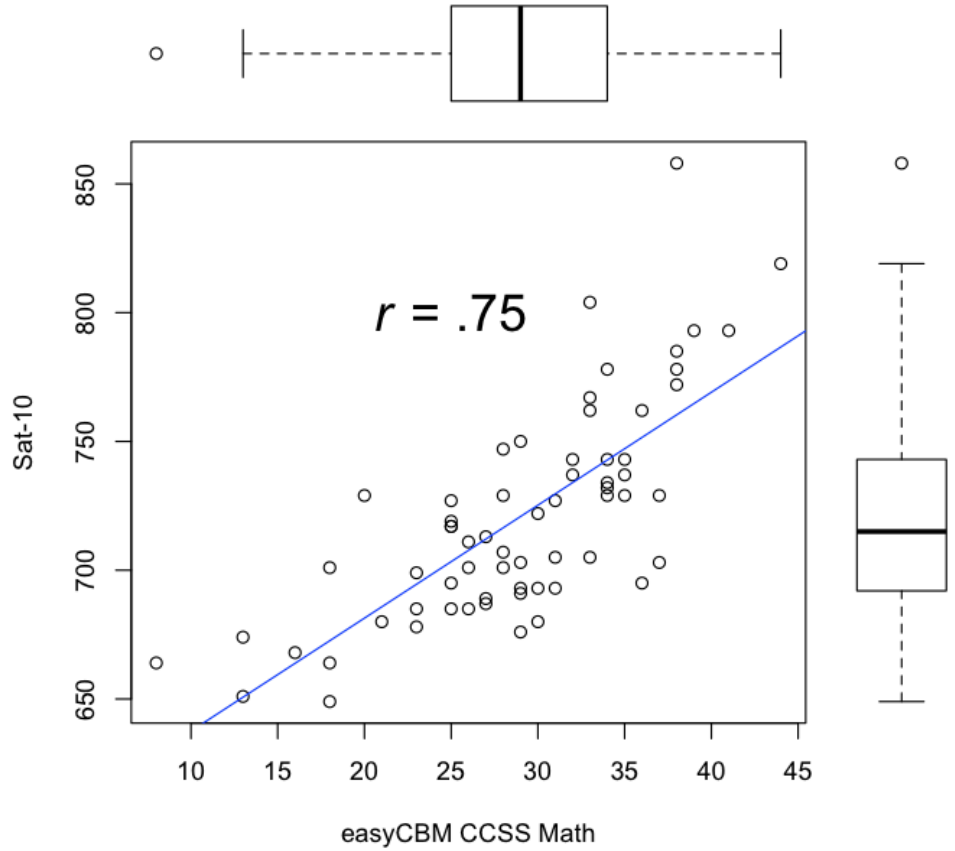
BRT
behavioral research & teaching

Criterion Validity Results: Grade 7



BRT
behavioral research & teaching

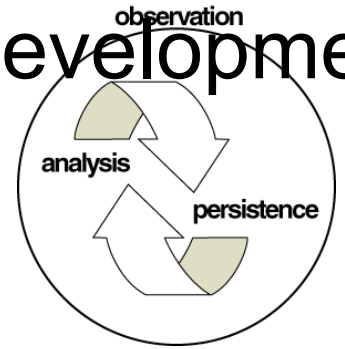
Criterion Validity Results: Grade 8



BRT
behavioral research & teaching

Where to from here?

- Measures appear reliable and to be measuring what we intend them to measure.
- Are we done? **NO!**
- Measurement development is **iterative**



BRT
behavioral research & teaching

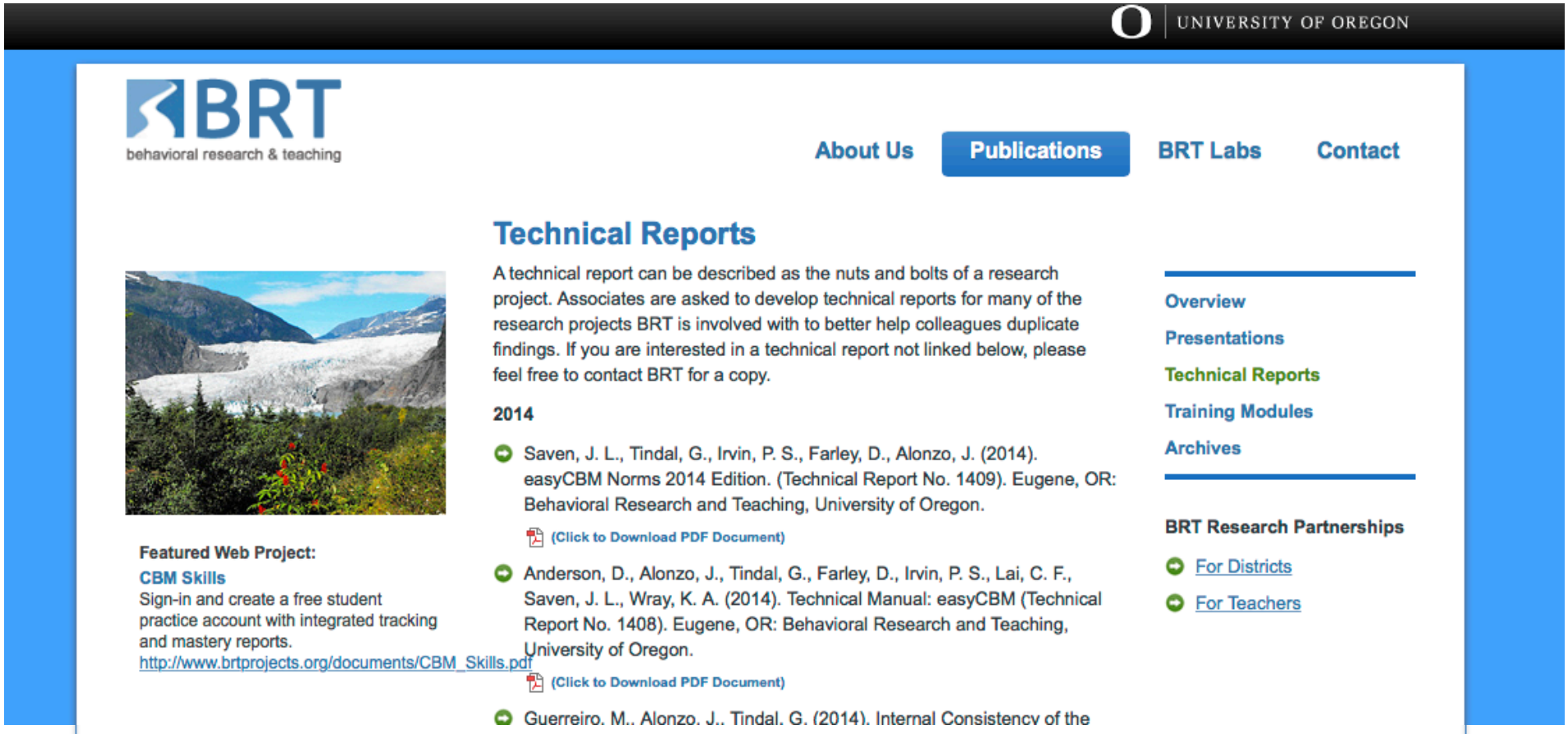
Continued Investigations

Current

- Item functioning (annual evaluation)
- Vertical scale creation
- Dimensionality
  - Does the test only measure one thing? Multiple things?
- Average growth

Planned

- Item fairness
- More investigations into reliability & validity



Thanks!

- Daniel Anderson: Behavioral Research and Teaching
  - daniela@uoregon.edu
  - http://www.brtprojects.org/publications/technical-reports

BRT
behavioral research & teaching