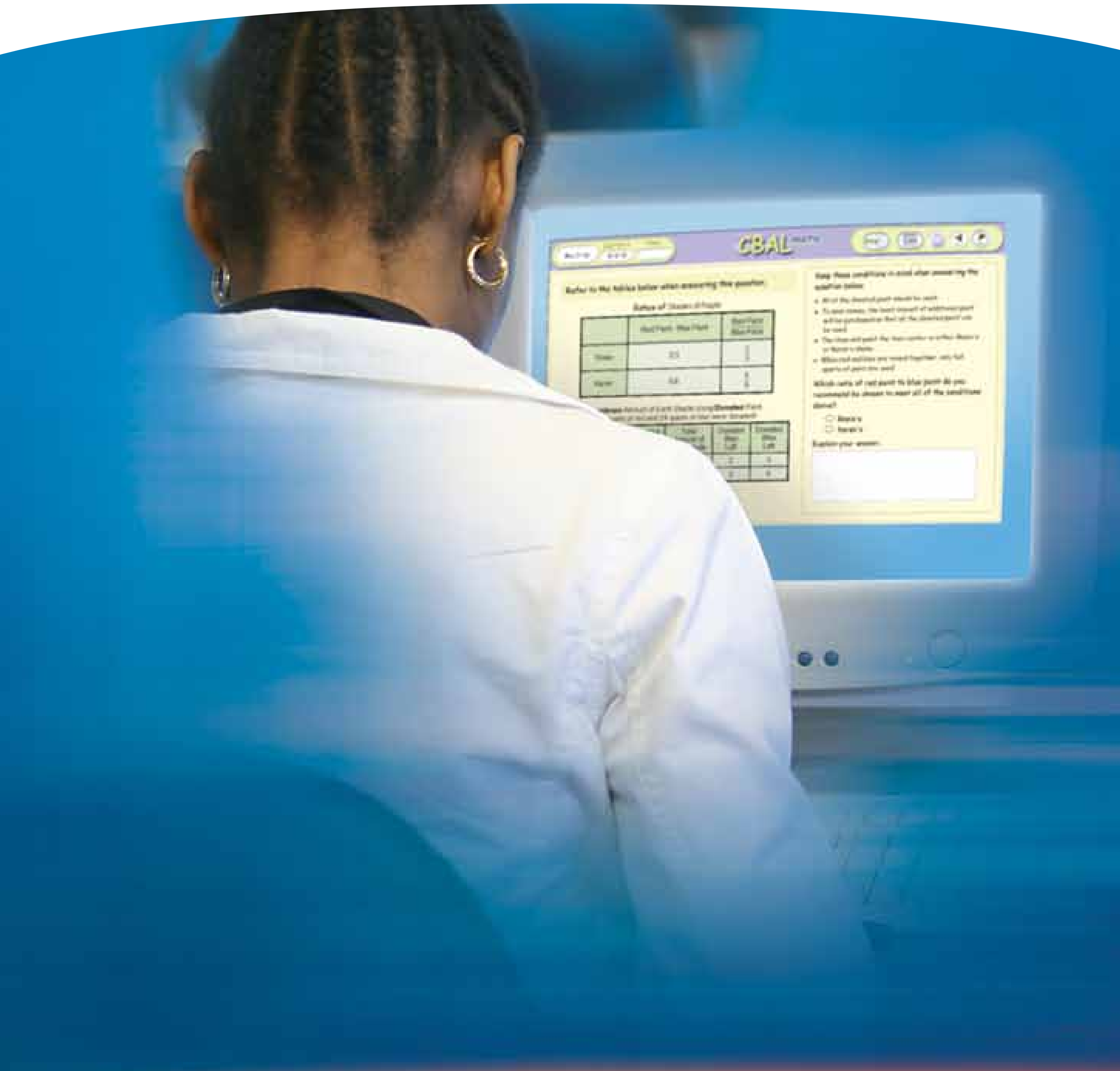
Listening. Learning. Leading.®



ETS Research Spotlight

CBAL™ Results From Piloting Innovative K–12 Assessments

ISSUE 6, MARCH 2012

Foreword

ETS Research Spotlight
ISSUE 6
MARCH 2012

Editor: Jeff Johnson

Copy Editor: Eileen Kerrigan

Layout: Sally Acquaviva

Cover Design: George Barrett

Since 2007, ETS researchers have been conducting a long-term Research & Development initiative called Cognitively Based Assessment *of*, *for*, and *as* Learning (*CBAL*™).

We are engaging in this complex initiative because we believe that existing approaches to K–12 accountability assessment could be markedly improved. The CBAL initiative aims to provide a model for such improvements by, among other things:

- incorporating into the CBAL initiative findings from learning-sciences research on what it means to be proficient in a domain;
- designing tasks that model effective teaching and learning practice;
- creating mechanisms for returning information about student performance in a rapid-enough fashion to be of use to teachers and students; and
- developing a system of testing on multiple occasions so that highly consequential decisions have a stronger evidential basis.

The Featured Research Synopsis in this issue of *ETS Research Spotlight* focuses on a recent ETS Research Report by Randy Bennett, ETS's Norman O. Frederiksen Chair in Assessment Innovation. In this report, Bennett summarizes the results of research from more than 10,000 online CBAL pilot administrations conducted between 2007 and 2010.

This research demonstrates the complex technical issues involved in assessment innovation — issues we must consider if we are to achieve meaningful changes in consequential educational assessment.

Pages 8–11 of this issue contain a brief list of online resources and publications that our staff members have created in support of the CBAL initiative.

If you'd like to learn more about how the ETS Research & Development scientists work to improve educational assessment, visit us on the web at **www.ets.org/research**.

Ida L

Ida Lawrence
Senior Vice President
Research & Development

The views expressed in this report are those of the authors and do not necessarily reflect the views of the officers and trustees of Educational Testing Service.

Copies can be downloaded from:
www.ets.org/research
March 2012
Research & Development Division
Educational Testing Service

To comment on any of the articles in this issue, write to ETS Research Communications via email at:
RDWeb@ets.org

FEATURED RESEARCH SYNOPSIS

CBAL: Results from Piloting Innovative K–12 Assessments

Editor's note: *The full reference list and technical details regarding this research appeared in the original work:*

Bennett, R. E. (2011). CBAL: Results from piloting innovative K–12 assessments (ETS Research Report No. RR-11-23). Princeton, NJ: Educational Testing Service. Retrieved from: ***http://www.ets.org/Media/Research/pdf/RR-11-23.pdf***

Introduction

CBAL stands for *Cognitively Based Assessment* of, for, *and* as *Learning*. It is an ETS research initiative intended to create new knowledge and capability for improving educational assessment. One focus of the initiative is a model for an innovative K–12 assessment system that documents what students have achieved (*of* learning), facilitates instructional planning (*for* learning), and is considered by students and teachers to be a worthwhile experience in and of itself (*as* learning).[1]

In the technical report summarized in this issue of *ETS Research Spotlight*, the author synthesizes empirical results from CBAL prototype summative assessments. As of December 2010, nearly 10,000 CBAL assessments had been piloted online in middle school grades in more than a dozen states. One purpose of these pilot administrations was to try out various assessment designs and tasks so that the ideas could be improved. A related purpose was to gather data necessary to address scientific questions.

Approach

The author summarized findings that CBAL researchers reported between 2007 and 2011 across three areas — reading, writing, and mathematics.[2]

Several types of findings were summarized, as described below:

Basic psychometric functioning (reported for reading, writing, and mathematics) – Empirical data were reviewed from a number of studies that provided evidence of how the test items and test forms behaved. Some common indicators of psychometric functioning include:

- item difficulty;
- item discrimination, or how well the items served to differentiate skilled performers from less-skilled performers;

[1] See also Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York: Springer.

[2] Since the time that the reported pilot studies were conducted, the CBAL initiative's competency models have evolved so that the reading and writing assessments have been combined into a series of English language arts (ELA) assessments. To learn about the CBAL ELA competency model and provisional learning progressions, visit **http://www.ets.org/research/topics/cbal/learning_progressions/**.

- internal consistency, which is one measure of a test's reliability; and
- the number of omits (items that students skipped) and not-reached items (items at the end of a test that students did not complete).

Internal structure (reported for reading, writing, and mathematics) – Some of the studies summarized also evaluated whether the relationships observed among test items were consistent with the tests' substantive organization. That is, these analyses sought to either support or refute the working assumptions and theories about the knowledge and skills that CBAL assessment prototypes aim to cover.

External relations (reported for reading, writing, and mathematics) – Some of the studies summarized in the original report considered how students' performance on CBAL assessments relates to their performance on other measures. In one case, for example, data were available that allowed researchers to compare students' CBAL reading scores to their scores from the Maine Educational Assessment and from the Northwest Evaluation Association's Measures of Academic Progress (MAP).

Where possible, such analyses considered both correlations between CBAL scores and external measures of the *same* subject (i.e., comparing CBAL reading scores to reading scores on another assessment) as well as correlations between CBAL scores and external measures of a *different* subject (i.e., comparing CBAL math scores to external measures of reading).

Population group performance and differential item functioning (reported for reading and writing) – As is common in assessment research, the CBAL initiative has included studies that aim to check for significant differences in the performance of population groups delineated by demographic characteristics such as gender, race, or socioeconomic class.

When such differences are found, it is important to check whether they are likely to be related to unfair test content or other irrelevant sources. One common method of studying such differences is known as differential item functioning (DIF) analysis. When using DIF analysis, researchers match population groups by skill level — for example, comparing high-performing male students with high-performing female students, mid-range male students with mid-range female students, etc. — and look for individual test items that were unusually easy or hard for one population group or the other.

If researchers find such items, which are then said to "show DIF," content experts evaluate whether those items might present a substantial obstacle (or advantage) for one group due to reasons that are unrelated to the target construct. For example, if a reading test item unintentionally requires familiarity with a type of television program that boys are more likely than girls to watch, then female students with otherwise good reading skills might have a harder time answering the question than male students with similar reading skills. Such a test item might be harder for an entire population group for the wrong reasons (i.e., television viewing habits as opposed to reading skills), which would be unfair.

If researchers find differences between two population groups on the test as a whole without identifying any items that showed significant DIF, then those differences unfortunately may have more systemic causes that are unrelated to the test.

Diagnostic utility (reported for reading) – Inherent in CBAL's full name is the intention of modeling a system that includes not only assessments *of* learning, but also assessments *for* learning. Therefore, research is necessary to confirm whether the results of CBAL assessments are in fact useful for guiding and planning instruction. For example, some of the prototype CBAL assessments incorporate learning progressions that posit a theory of the typical sequence for which critical domain knowledge and skills might be acquired. Determining the level in a developmental sequence for a learner might be helpful in identifying what to teach that individual next. Among other things, test items written to measure student standing should follow the expected difficulty ordering.

Automated scoring (reported for reading, writing, and mathematics) – CBAL assessments make extensive use of "constructed response" items. Such tasks call upon a test taker to *construct* an answer — for example, by filling in the blank or writing a passage such as an essay or an email message — rather than choosing from a set of possible answers as in a multiple-choice question (which is one kind of "*selected*-response" test item). Constructed-response items are often thought to elicit more "authentic" evidence of what a test taker knows or can do, but without the help of computer technology to score them, it may be sometimes too costly to use them on a large scale. Therefore, CBAL research also considers whether the items lend themselves to reliable scoring through applications such as ETS's *c-rater*^{SM} scoring engine, which is used to evaluate short, free-text responses.

Findings

Tables 1–3 on the following pages synthesize findings from the studies in CBAL reading, writing, and mathematics, across the six categories defined above. The original report (Bennett, 2011) and its references provide data detailing these findings and a comprehensive discussion of their implications.

Table 1: Synthesis of CBAL Reading Findings

Category	Findings
Basic Psychometric Functioning	• In five CBAL summative reading pilots, researchers observed: – a reasonable level of internal consistency, which is one common indicator of reliability; – item difficulty levels that were appropriate for the samples of students tested (middle school); and – low missing-response rates, meaning that there were few items that students skipped or did not reach.
Internal Structure	• Data from most analyses suggest that CBAL reading comprehension items measured a single dimension, even across different test forms that were constructed to measure different skills. • Reading items designed to measure two key aspects of the CBAL reading competency model appeared to differentiate themselves in expected ways based on difficulty.
Population Performance and DIF	• Investigations of population-group performance found differences that were similar to those that are found on national reading assessments. • Differential item functioning (DIF) was largely absent from the CBAL reading forms studied.
External Relations	• As expected, CBAL reading forms appeared to be more highly correlated with several other measures of reading than with measures of math. • The CBAL reading comprehension section was correlated more highly with another measure of comprehension than it was with a measure of oral reading; similarly, the CBAL spoken section had the expected, opposite pattern.
Diagnostic Utility	• Data suggested that CBAL reading summative tests may provide initial formative information useful for informing classroom follow-up. • Evidence also suggested that the comprehension and spoken sections might be able to jointly identify student groups with distinct skill patterns and that items written to learning progressions might be used to place students tentatively in instructional levels.
Automated Scoring	• An analysis of the machine scoring of student answers to constructed-response questions showed that, although human judges generally agreed more highly with themselves than with an automated system, the automated system's scores were, on average, not dramatically different from judges' ratings.

Table 2: Synthesis of CBAL Writing Findings

Category	Findings
Basic Psychometric Functioning	• Compared with CBAL reading pilot results, the CBAL writing test forms were more variable in terms of their internal consistency reliability, difficulty, and missing-response rates. This variation may, in part, reflect the range of design changes that have been explored since 2007.
Internal Structure	• Factor analyses generally suggested a single factor both within forms and across forms, though other analytical approaches indicated the possibility of more complex dimensional structures.
Population Performance and DIF	• An analysis of population-group differences showed patterns similar to those observed on the U.S. Department of Education's National Assessment of Educational Progress (NAEP), often known as The Nation's Report Card.
External Relations	• As expected, CBAL writing scores appeared to be more highly correlated with reading scores than with math scores on other standardized assessments.
Automated Scoring	• Automated scores were highly correlated with human scores and reasonably correlated with both the same form's total test score and with total scores on another CBAL writing form taken by the same students.

Table 3: Synthesis of CBAL Mathematics Findings

Category	Findings
Basic Psychometric Functioning	• As seen in the CBAL writing assessments, there was considerable variation across pilot test forms with respect to internal consistency, difficulty, and missing-response rates. This variation may, in part, reflect the range of design changes that have been explored since 2007.
Internal Structure	• Factor-analytic results most often suggested a single dimension within forms.
External Relations	• As expected, CBAL math scores appeared to be more correlated with math scores from other assessments than with reading scores from those assessments (but only by relatively modest amounts).
Automated Scoring	• For the types of textual-response items studied, agreement of automated scores with human raters was noticeably lower than that found between human raters.

Conclusion

The primary goal of CBAL research is to create new knowledge and capability for improving educational assessment, including in the K–12 arena. In the conclusion of the original report, the author not only summarized the initial data on the technical quality of CBAL scores, but also proposed a list of questions that CBAL research must answer if the CBAL initiative is to demonstrate improvement over the status quo. While the initiative is making progress, the author noted that the results to date highlight the considerable investment and time horizon required to successfully execute meaningful innovation in consequential educational assessment.

References

See the original paper, cited in the Editor's note on page 2 of this issue of *Spotlight*, for a list of references that the author used when conducting and reporting on the research described in this synopsis.

About the Author

Randy Bennett is ETS's Norman O. Frederiksen Chair in Assessment Innovation.

You May Also Be Interested In . . .

Below is a select list of other ETS resources and publications related to the CBAL research initiative.

RESOURCES

Visit the CBAL initiative pages on the ETS.org website for examples of CBAL tasks, recent research publications, information on becoming a pilot site, and an opportunity to interact with other researchers and educators to discuss innovation in K–12 assessment.

ETS recently added to the site a page that focuses on the CBAL English language arts (ELA) competency model and provisional learning progressions. The new page also includes access to a site that encourages educators to share their thoughts on the CBAL ELA model. Access the CBAL initiative homepage at **http://www.ets.org/research/topics/cbal/initiative**.

PUBLICATIONS

Where available, the full text for these recent publications related to the CBAL initiative can be accessed at **http://www.ets.org/research/topics/cbal/publications/**.

2011

Constructed-Response Mathematics Tasks Study

J. H. Fife, E. A. Graf, & S. Ohls (2011)
ETS Research Report No. RR-11-35

This report describes potential ways in which selected CBAL constructed-response mathematics tasks might be revised to reduce construct-irrelevant variance.

Automated Scoring of CBAL Mathematics Tasks with *m-Rater*

J. H. Fife (2011)
ETS Research Memorandum No. RM-11-12

This paper presents the automated scoring work done in CBAL Mathematics in 2009 using *m-rater*, a technology developed at ETS for automated scoring of mathematics items.

CBAL: Results From Piloting Innovative K–12 Assessments

R. E. Bennett (2011)
ETS Research Report No. RR-11-23

This report summarizes empirical results from almost 10,000 online administrations of CBAL summative assessments conducted from 2007 to 2010.

The CBAL Reading Assessment: An Approach for Balancing Measurement and Learning Goals

K. M. Sheehan & T. O'Reilly (2011)
ETS Research Report No. RR-11-21

This paper presents a framework for developing new types of reading comprehension assessments that provide evidence about what students know and can do and that help to move learning forward.

Automated Scoring Within a Developmental, Cognitive Model of Writing Proficiency

P. Deane, T. Quinlan, & I. Kostin (2011)
ETS Research Report No. RR-11-16

This paper focuses on the potential for using automated scoring techniques to support learning effectively within CBAL assessments.

Four Years of Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL): Learning About Through-Course Assessment (TCA)

J. P. Sabatini, R. E. Bennett, & P. Deane (2011)

Proceedings of the Invitational Research Symposium on Through-Course Summative Assessments

Center for K–12 Assessment & Performance Management at ETS

This paper describes the lessons learned about through-course summative assessment and the reasoning behind some of the design decisions that underlie such assessment in CBAL.

Writing Assessment and Cognition

P. Deane (2011)
ETS Research Report No. RR-11-14

This paper reviews a model that places a strong emphasis on writing as an integrated, socially situated skill.

The CBAL Summative Writing Assessment: A Draft Eighth-Grade Design

P. Deane, M. Fowles, D. Baldwin, & H. Persky (2011)
ETS Research Memorandum No. RM-11-01

This paper describes the process and results of developing draft summative writing assessments within the CBAL Initiative. It outlines and reviews four designs, and briefly discusses initial results from preliminary pilots.

Formative Assessment: A Critical Review

R. E. Bennett (2011)
Assessment in Education: Principles, Policy and Practice, Vol. 18, No. 1, pp. 5–25

This paper covers six interrelated issues in formative assessment, several of which motivated the approach taken in the CBAL initiative.

2010

Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL): A Preliminary Theory of Action for Summative and Formative Assessment

R. E. Bennett (2010)
Measurement: Interdisciplinary Research & Perspectives, Vol. 8, No. 2–3, pp. 70–91

This paper describes the notion of a theory of action, offers a preliminary version of such a theory for the CBAL initiative, and outlines research necessary to evaluate that theory.

An Evidence-Centered Approach to Using Assessment Data for Policymakers

J. S. Underwood, D. Zapata-Rivera, &
W. VanWinkle (2010)
ETS Research Report No. RR-10-03

District-level policymakers receive reports of student achievement data that are complex, difficult to read, and even harder to interpret. In this report, the authors propose an evidence-centered reporting framework in order to design reports that will help policymakers make sense of data.

Highlights from the Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL) Project in Mathematics

E. A. Graf, K. Harris, E. Marquez, J. H. Fife, &
M. Redman (2010)
ETS Research Spotlight, No. 3, pp. 19–30

This article describes the early design and development stages of the mathematics strand of the CBAL project.

2009

c-rater^{SM}: Automatic Content Scoring for Short Constructed Responses

J. Z. Sukkarieh & J. Blackmore (2009)

Paper in the proceedings of the 22nd Florida Artificial Intelligence Research Society (FLAIRS) Conference

The *c-rater* automated scoring engine, developed at ETS, is a technology for automatic content scoring for short, free-text responses. This paper describes recent developments in this technology.

Defining Mathematics Competency in the Service of Cognitively Based Assessment for Grades 6 Through 8

E. A. Graf (2009)
ETS Research Report No. RR-09-42

This report makes recommendations for the development of middle-school mathematics assessment. It discusses how to model mathematical competency at the middle school level, the kinds of evidence that reflect student competency and support future learning, and how to design tasks that elicit evidence.

Cognitively Based Assessment *of*, *for*, and *as* Learning: A Framework for Assessing Reading Competency

T. O'Reilly & K. M. Sheehan (2009)
ETS Research Report No. RR-09-26

This paper presents the rationale and research base for a reading competency model designed to guide the development of cognitively based assessment of reading comprehension.

Cognitively Based Assessment *of*, *for*, and *as* Learning: A 21st Century Approach for Assessing Reading Competency

T. O'Reilly & K. M. Sheehan (2009)
ETS Research Memorandum No. RM-09-04

This paper describes the CBAL system's approach for assessing reading comprehension in an accountability setting. The approach uses evidence-centered design to develop a competency model that drives the development of summative, formative, and professional support aspects of the assessment.

Horizontal and Vertical Linking in a Longitudinal Design

F. Rijmen (2009)
ETS Research Memorandum No. RM-09-03

This paper describes two longitudinal data collection designs that may result in substantial reduction of costs.

ETS®

Listening. Learning. Leading.®

www.ets.org