Listening. Learning. Leading.®

Research Report
ETS RR–12-24

Statistical Report of 2011 *CBAL*™ Multistate Administration of Reading and Writing Tests

Jianbin Fu

Maxwell Wise

December 2012

ETS Research Report Series

Statistical Report of 2011 *CBAL*™ Multistate Administration of Reading and Writing Tests

Jianbin Fu and Maxwell Wise

ETS, Princeton, New Jersey

December 2012

Associate Editor: James E. Carlson

Reviewers: Shameem Gaj and Richard Schwarz

Abstract

In the Cognitively Based Assessment of, for, and as Learning (*CBAL*™) research initiative, innovative K-12 prototype tests based on cognitive competency models are developed. This report presents the statistical results of the 2 CBAL Grade 8 writing tests and 2 Grade 7 reading tests administered to students in 20 states in spring 2011. Specifically, classical item statistics including rater reliabilities for human-scored items, item *p*+ values, item-total correlations, item response times, item omit and not-reached rates, system error rates, differential item functioning, interscore correlations, and reliabilities of subscores and total scores are reported. In addition, *t*-tests, multiple comparisons, correlations, and mixed models are used to examine the factors influencing test scores, including test form, test order, and various background variables at the student, teacher, and school levels. The results show that these 4 tests performed reasonably well.

Key words: CBAL, writing test, reading tests, item analysis, statistical report

Acknowledgments

Thanks to Randy Bennett, Jim Carlson, Richard Schwarz, Shameem Gaj, and Peter van Rijn for their helpful suggestions and edits on early versions of this report. We are grateful to Ruth Greenwood for her editorial assistance.

Table of Contents

List of Tables

The Cognitively Based Assessment of, for, and as Learning (*CBAL*™) research initiative is intended to create a model for an innovative K-12 assessment system that measures students' achievement (of learning), provides timely feedback for educational intervention (for learning), and is a worthwhile educational experience in and of itself (as learning; Bennett, 2010). To help achieve these goals, CBAL summative tests are intended to be administered multiple times across a school year and are referred to as periodic accountability assessments (PAAs). Aggregate scores across multiple tests are designed for possible uses for accountability purposes; however, in the current stage CBAL is a research project, and CBAL summative tests are used only for experimental purposes.

CBAL tests are developed based on underlying cognitive competency models that incorporate curriculum standards with the results of learning sciences' research. The competency models describe skills that students need to learn, their interrelationships, and hypothesized orderings in which those skills might be taught, often called learning progressions (Deane, 2011; Graf, 2009; O'Reilly & Sheehan, 2009a, 2009b). Tests are administered online and include innovative technology-enhanced items that are typically organized under a common scenario and gauge higher-order critical-thinking abilities.

In spring 2011, two Grade 8 writing PAAs and two Grade 7 reading PAAs were administered as described in the Test and Sampling Designs section below. This report presents the statistical results of the test administration and includes the following content: (a) the test and sampling designs; (b) classic item analyses, including rater reliabilities for human-scored items, item *p*+ values, item-total correlations, item omit and not-reached rates, item response times, and differential item functioning; (c) summary statistics of subscores and total raw scores, including means, standard deviations, interscore correlations, and reliabilities; (d) test performance by demographic groups based on gender, socioeconomic status, English language learner status, test accommodation status, and race/ethnicity; and (e) effects on test theta scores of school and school background variables (percentage free/reduced price lunch, percentage minority, and percentage student-teacher ratio), teacher and teacher background variables (years teaching English, and instruction content), student and student demographic variables, PAA, and test order.

Test and Sampling Designs

Writing PAAs

The 2011 multistate administration included two Grade 8 writing PAAs focused on different writing genres: Ban Ads and Mango Street. Each PAA had both dichotomous and polytomous items, and item types included constructed response (CR), short CR (SCR), selected response (SR), and click and click (C&C; i.e., select and copy text from the passage as the answer and paste into the answer box). An item was either automatically scored by computer or human scored. (See Table 1 for the writing genre, the numbers of CR/SCR and SR/C&C items and subscores, and possible maximum total raw score for each PAA.)

Table 1

CBAL Writing Test Design

PAA	Writing genre	Number of SR/C&C items	Number of CR/SCR items	Number of subscores	Max total raw score[a]
Ban Ads	Persuasive/argumentative writing	21	5	6	63
Mango Street	Writing about literature	10	4	4	41

Note. SR = selected response; C&C = click & click; CR = constructed response; SCR = short CR.
[a] After score weights are applied.

Each PAA was based on a common scenario. Items in each PAA were organized under four tasks based on the nature of the questions. The first three tasks were lead-in tasks measuring critical thinking skills necessary for writing a good essay in a specific genre. The fourth task was writing the essay itself. The first three tasks comprised Section I of the test and the fourth task was Section II of the test. The PAAs were timed at the task level and each section had to be finished in 45 minutes.

Tables 2 and 3 list the information for each item in the two writing PAAs, including item score ID, task, and subscore that an item belongs to, item sequence number, item type, scoring type (computer or human scored), score range after score weights were applied, and score weight. For a description of the test design from the content perspective, see Deane et al. (2009) and Deane, Fowles, Baldwin, and Persky (2011).

Table 2

Ban Ads: Item and Subscore Information

							Subscores					
Task number and name	Item sequence	Item score ID	Type	Scoring type	Score range[a]	Score weight	S1	S2	S3	S4	S5	S6
1. Read and summarize arguments	1	BA_01A_01	SR	A	0–1	1	1					
	2	BA_01A_02	SR	A	0–1	1	1					
	3	BA_01A_03	SR	A	0–1	1	1					
	4	BA_01A_04	SR	A	0–1	1	1					
	5	BA_01A_05	SR	A	0–1	1	1					
	6	BA_01B	CR	H	0–2	1		1				
	7	BA_01C	CR	H	0–2	1		1				
2. Analyze arguments	8	BA_02AX_A	SR	A	0–1	1			1			
	9	BA_02AX_B	SR	A	0–1	1			1			
	10	BA_02AX_C	SR	A	0–1	1			1			
	11	BA_02AX_D	SR	A	0–1	1			1			
	12	BA_02AX_E	SR	A	0–1	1			1			
	13	BA_02AX_F	SR	A	0–1	1			1			
	14	BA_02AX_G	SR	A	0–1	1			1			
	15	BA_02AX_H	SR	A	0–1	1			1			
	16	BA_02AX_I	SR	A	0–1	1			1			
	17	BA_02AX_J	SR	A	0–1	1			1			
	18	BA_02BX_A	SR	A	0–1	1				1		
	19	BA_02BX_B	SR	A	0–1	1				1		
	20	BA_02BX_C	SR	A	0–1	1				1		
	21	BA_02BX_D	SR	A	0–1	1				1		
	22	BA_02BX_E	SR	A	0–1	1				1		
	23	BA_02BX_F	SR	A	0–1	1				1		
3. Critique an argument	24	BA_03	CR	H	0–8	2					1	
4. Write an essay	25	BA_04_I	CR	H	0–15	3						1
	26	BA_04_III	CR	H	0–15	3						1

Note. S1 = Summary Feedback; S2 = CR Summary; S3 = Claims; S4 = Evidence; S5 = Critique; S6 = Essay; SR = selected response; A = automatically scored by computer; CR = constructed response; H = human scored.

[a] Score range after score weights are applied.

Table 3

Mango Street: Item and Subscore Information

Task number and name	Item sequence	Item score ID	Type	Scoring type	Score range [a]	Score weight	Subscores			
							S1	S2	S3	S4
1. Support interpretations of the story	1	MG_01_01	C&C	A	0–1	.5	1			
	2	MG_01_02	C&C	A	0–1	.5	1			
	3	MG_01_03	C&C	A	0–1	.5	1			
	4	MG_01_04	C&C	A	0–1	.5	1			
	5	MG_01_05	C&C	A	0–1	.5	1			
2. Explain whether a character's attitude changes	6	MG_02_01	CR	H	0–8	2		1		
3. Help another student interpret the text	7	MG_03_01	SR	A	0–1	1			1	
	8	MG_03_02	SR	A	0–1	1			1	
	9	MG_03_03	SR	A	0–1	1			1	
	10	MG_03_04	SR	A	0–1	1			1	
	11	MG_03_05	SR	A	0–1	1			1	
	12	MG_03_06	SCR	H	0–3	1			1	
4. Write an essay	13	MG_04_I	CR	H	0–10	2				1
	14	MG_04_III	CR	H	0–10	2				1

Note. S1 = Support Interpretation; S2 = Interpretive Discussion; S3 = Choose Interpretation; S4 = Essay. C&C = click & click; A = automatically scored by computer; CR = constructed response; H = human scored; SR = selected response; SCR = short CR.

[a] Score range after score weights are applied.

Reading PAAs

Table 4 shows the test design of the reading forms used in the 2011 multistate administration. These test forms included two primary PAAs (A and B), with two external linking sets (C1 and C2) embedded into each PAA to create four PAA forms: PAA-A1, PAA-A2, PAA-B1 and PAA-B2. The external linking items were not used for scoring.

Each form included two 50-minute sections. Section I was a scenario-based task set including 20 items focused on either information/persuasive reading skills under a common scenario, *Wind Power* (Form A), or literary reading skills under a common scenario, *Seasons* (Form B). Items were organized under five (Form A) or four (Form B) tasks based on the nature of the questions (e.g., community comments and solving problems). Section II contained 28 or 29 discrete vocabulary items in mini-passage sets including 18 items in Block A or B, and 10 and 11 external linking items in Block C1 or C2, respectively.

Table 4

CBAL Reading Test Design

Section	Number of items	Description	
		PAA-A1	PAA-A2
I	20	*Wind Power*: an extended, integrated scenario-based task set, focused on information/persuasive reading skills	Same as A1
II	28/29	Block A (18 items) and external linking Block C1 (10 items): discrete vocabulary items in mini-passage sets focused on literary and information/persuasive reading skills	Block A (18 items) and external linking Block C2 (11 items)
		PAA-B1	PAA-B2
I	20	*Seasons*: an extended, integrated scenario-based task set, focused on literary reading skills	Same as B1
II	28/29	Block B (18 items) and external linking Block C1 (10 items): discrete vocabulary items in mini-passage sets focused on literary and information/persuasive reading skills	Block B (18 items) and external linking Block C2 (11 items)

Note. PAA = periodic accountability assessment.

The items in each form measured the content areas of literary, information/persuasive, and vocabulary skills and were classified into three levels in terms of the complexity of skills as denoted in the CBAL reading competency model (O'Reilly & Sheehan, 2009a, 2009b). Levels 1 and 2 were the two subcategories of model-building skill. Level 1 referred to *identify, retrieve, or infer* when activation was high, and Level 2, which was a more difficult skill, referred to *compare, interpret, or infer* when activation was low. Level 3, the most difficult skill, referred to *applied comprehension* (i.e., *evaluate, integrate, or synthesize*).

Like the writing PAAs, each reading PAA form had both dichotomous and polytomous items that were either automatically scored by computer or human scored, and the item types included C&C, CR, SCR, and SR. Unlike traditional multiple choice items, most of the SR items asked examinees to select more than one correct option.

Each PAA form had six subscores: Model Building (MB), Applied Comprehension (AC), Information Literacy (IL), Vocabulary (V), Informational (I), and Literary (L). Tables 5 through 7 list the item information for each item in the two primary PAAs and the two linking sets, including item score ID, the test section, task, and subscore that an item belongs to, item sequence number in the section, item type, scoring type (computer or human scored), and score range. Note that some items are mapped to two subscores. For a description of the test design from a content perspective, see CBAL ELA Team (2011).

Table 5

Reading PAA-A: Item and Subscore Information

						Subscores					
Task	Item sequence within section	Item score ID	Item type	Scoring type	Score range	S1	S2	S3	S4	S5	S6
Section I (Wind Power)											
How Wind Power Works	1	WP_11	SR	A	0–1	1				1	
	2	WP_12	SR	A	0–1	1				1	
	3	WP_13	SR	A	0–2	1				1	
	4	WP_14	SR	A	0–1		1				
Find Information	5	WP_21	C&C	A	0–1			1			
	6	WP_22	SR	A	0–1			1			
	7	WP_23	SR	A	0–1			1			
	8	WP_24	SR	A	0–1			1			
Possibilities & Challenges	9	WP_31	SCR	H	0–1	1				1	
	10	WP_32	SR	A	0–1			1			
	11	WP_33	SCR	H	0–2	1				1	
	12	WP_34	C&C	A	0–2	1				1	
Community Comments	13	WP_41	C&C	A	0–2			1			
	14	WP_42	C&C	A	0–1		1				
	15	WP_43	SR/CR	H	0–2		1			1	
	16	WP_44	SR/C&C/CR	H	0–2		1			1	
Solving Problems	17	WP_51	SR/C&C	A	0–1	1				1	
	18	WP_52	SR	A	0–1		1			1	
	19	WP_53	C&C	A	0–1		1			1	
	20	WP_54	CR	H	0–2	1				1	
Block A in Section II											
	1	A02	SR	A	0–1	1					1
	2	A03	SR	A	0–2	1					1
	3	A04	SR	A	0–1				1		
	4	A05	C&C	A	0–1	1				1	
	5	A06	SR	A	0–1	1				1	
	6	A07	SR	A	0–1	1				1	
	7	A08	SR	A	0–1				1		
	8	A09	SR	A	0–1	1				1	
	9	A10	SR	A	0–1	1				1	

						Subscores					
Task	Item sequence within section	Item score ID	Item type	Scoring type	Score range	S1	S2	S3	S4	S5	S6
Block A in Section II											
	21	A11	SR	A	0–1				1		
	22	A12	SR	A	0–1	1				1	
	23	A13	C&C	A	0–1	1				1	
	24	A14	SR	A	0–1		1			1	
	25	A15	SR	A	0–1				1		
	26	A16	SR	A	0–2	1				1	
	27	A17	SR	A	0–1	1				1	
	28	A18	SR	A	0–1				1		
	29	A01	SR	A	0–1				1		

Note. S1 = Model Building (MB); S2 = Applied Comprehension (AC); S3 = Information Literacy (IL); S4 = Vocabulary (V); S5 = Informational (I); S6 = Literary (L); CR = constructed response; SR = selected response; SCR = short CR; C&C = click & click; A = automatically scored by computer; H = human scored; PAA = periodic accountability assessment.

Participants

The CBAL PAAs were administered online to a convenience sample of 3,576 Grade 8 students from 35 schools in 20 states. (See Table 8 for the sample's distribution by various demographic indicators.) The students took two PAAs out of the four in one of the 14 orders (see Table 9). These test sequences also took into account the balance of linking sets in the reading tests. For security reasons, a first PAA could not be used as a second PAA in the same school. To accommodate this restriction, these test sequences were grouped into four clusters each including four test sequences. A school was randomly assigned to one of the four clusters, and the students in a school were randomly assigned to one of the four sequences in the cluster to which the school was assigned.

Table 10 shows the sample sizes for each test sequence. Students completed both PAAs within 68 days on average (with a standard deviation of 14 days). Note that all the sample sizes were reported after the test dataset was cleaned (see Appendix A for the data cleaning process).

Table 6

Reading PAA-B: Item and Subscore Information

						Subscores					
Task	Item sequence within section	Item score ID	Item type	Scoring type	Score range	S1	S2	S3	S4	S5	S6
Section I (Seasons)											
Sound of Summer Running	1	SS_11	SCR	H	0–1	1					1
	2	SS_12	SR	A	0–1	1					1
	3	SS_13	C&C	A	0–2	1					1
	4	SS_14	CR	H	0–1	1					1
	5	SS_15	SR	A	0–1	1					1
	6	SS_16	C&C	A	0–1	1					1
	7	SS_17	SR	A	0–1	1					1
	8	SS_18	SR	A	0–1	1					1
	9	SS_19	SR	A	0–2		1				1
Berkshires in April	10	SS_21	SR	A	0–2	1					1
	11	SS_22	SR	A	0–2	1					1
	12	SS_23	C&C	A	0–2	1					1
	13	SS_24	C&C	A	0–2	1					1
	14	SS_25	SCR	A	0–1	1					1
Combined	15	SS_31	SR	A	0–1		1				1
	16	SS_32	SR	A	0–1		1				1
Using rubric	17	SS_41	SR	A	0–2		1				1
	18	SS_42	SR	A	0–2		1				1
	19	SS_43	SR	A	0–2		1				1
	20	SS_44	SR	A	0–2		1				1
Block B in Section II											
	1	B02	C&C	A	0–1	1					1
	2	B03	SR	A	0–1	1					1
	3	B04	SR	A	0–1				1		
	4	B05	SR	A	0–1	1				1	
	5	B06	SR	A	0–2	1				1	
	6	B07	SR	A	0–2		1			1	
	7	B08	SR	A	0–1				1		
	8	B09	SR	A	0–1	1				1	
	9	B10	SR	A	0–1	1				1	
	21	B11	SR	A	0–1				1		
	22	B12	SR	A	0–1	1				1	
	23	B13	C&C	A	0–1	1				1	
	24	B14	SR	A	0–1	1				1	
	25	B15	SR	A	0–1				1		
	26	B16	SR	A	0–1	1				1	
	27	B17	C&C	A	0–1	1				1	
	28	B18	SR	A	0–1				1		
	29	B01	SR	A	0–1				1		

Note. S1 = Model Building (MB); S2 = Applied Comprehension (AC); S3 = Information Literacy (IL); S4 = Vocabulary (V); S5 = Informational (I); S6 = Literary (L); CR = constructed response; SCR = short constructed response; H = human scored; SR = selected response; A = automatically scored by computer; C&C = click & click; PAA = periodic accountability assessment.

Table 7

Reading Linking Blocks C1 and C2: Item and Subscore Information

						Subscores					
Section	Item sequence within section	Item score ID	Item type	Scoring type	Score range	S1	S2	S3	S4	S5	S6
Linking block C1 in Section II	10	C05	C&C	A	0–1	1				1	
	11	C06	SR	A	0–1	1				1	
	12	C10	C&C	A	0–2	1				1	
	13	C08	SR	A	0–2		1			1	
	14	C01	SR	A	0–1				1		
	15	C13	C&C	A	0–2	1					1
	16	C15	C&C	A	0–2	1					1
	17	C16	C&C	A	0–2	1					1
	18	C22	SR	A	0–1		1				1
	19	C23	SR	A	0–1		1				1
Linking block C2 in Section II	10	C02	SR	A	0–1	1				1	
	11	C03	SR	A	0–1	1				1	
	12	C04	C&C	A	0–2	1				1	
	13	C07	SR	A	0–1	1		1		1	
	14	C11	C&C	A	0–1		1			1	
	15	C12	SR	A	0–1				1		
	16	C14	C&C	A	0–2	1					1
	17	C18	C&C	A	0–1	1					1
	18	C19	C&C	A	0–2	1					1
	19	C20	C&C	A	0–2		1				1
	20	C21	SR	A	0–2		1				1

Note. S1 = Model Building (MB); S2 = Applied Comprehension (AC); S3 = Information Literacy (IL); S4 = Vocabulary (V); S5 = Informational (I); S6 = Literary (L); CR = constructed response; SR = selected response; SCR = short CR; C&C = click & click; A = automatically scored by computer; H = human scored.

Table 8

Test Sample Distribution by Demographic Characteristic

	All data		PAA A [a]		PAA B [a]		Ban Ads		Mango	
Demographics	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
State										
AL	244	7	103	7	95	6	96	7	94	7
AR	31	1	12	1	12	1	14	1	11	1
AZ	33	1	15	1	13	1	11	1	13	1
CA	383	11	151	11	154	10	143	10	143	10
CO	112	3	33	2	46	3	43	3	34	2
GA	134	4	48	3	55	4	48	3	56	4
KY	110	3	46	3	45	3	52	4	53	4
MA	51	1	22	2	23	2	25	2	22	2
MI	299	8	139	10	133	9	135	9	145	10
MN	107	3	49	3	53	4	45	3	47	3
MS	92	3	0	0	24	2	26	2	0	0
NJ	336	9	137	10	130	9	141	10	142	10
NY	254	7	117	8	85	6	101	7	95	7
OH	300	8	125	9	134	9	114	8	122	9
PA	98	3	43	3	43	3	42	3	40	3
SC	277	8	104	7	120	8	125	9	104	7
SD	81	2	37	3	34	2	39	3	39	3
TN	63	2	25	2	28	2	29	2	29	2
TX	523	15	198	14	222	15	197	14	185	13
WI	48	1	17	1	23	2	20	1	23	2
Region										
East	739	21	319	22	281	19	309	21	299	21
Midwest	976	27	425	30	434	29	419	29	440	32
South	1,333	37	478	34	544	37	521	36	468	34
West	528	15	199	14	213	14	197	14	190	14
Locale										
Rural	1,535	43	589	41	654	44	622	43	591	42
Suburban	1,166	33	473	33	471	32	483	33	468	34
Urban	875	24	359	25	347	24	341	24	338	24
Title 1										
Yes	2,413	67	945	67	1,010	69	974	67	903	65
Unreported	1,163	33	476	34	462	31	472	33	494	35
Charter										
Yes	193	5	65	5	82	6	74	5	70	5
Unreported	3,383	95	1,356	95	1,390	94	1,372	95	1,327	95
Gender										
Male	1,674	47	656	46	687	47	673	47	631	45
Female	1,707	48	724	51	734	50	718	50	720	52

	All data		PAA A [a]		PAA B [a]		Ban Ads		Mango	
Demographics	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
Unreported	195	5	41	3	51	3	55	4	46	3
Race										
African American	398	11	141	10	165	11	165	11	164	12
Asian/Pacific Islander	262	7	107	8	101	7	116	8	108	8
Hispanic	559	16	218	15	235	16	212	15	199	14
Native American	10	0	2	0	4	0	2	0	4	0
White	2,149	60	909	64	915	62	896	62	875	63
Unreported	198	6	44	3	52	4	55	4	47	3
Low socioeconomic status (SES)[b]										
No	1,819	51	777	55	784	53	784	54	747	53
Yes	1,265	35	502	35	507	34	499	35	491	35
Unreported	492	14	142	10	181	12	163	11	159	11
English language learner (ELL)										
Current ELL (Yes)	75	2	24	2	33	2	31	2	22	2
Former ELL (No)	65	2	33	2	31	2	28	2	18	1
English proficient (No)	2,764	77	1,156	81	1,157	79	1,152	80	1,136	81
Unreported	672	19	208	15	251	17	235	16	221	16
Test accommodations										
No	2,661	74	1,126	79	1,125	76	1,121	78	1,094	78
Yes	233	7	82	6	95	6	81	6	76	5
Unreported	682	19	213	15	252	17	244	17	227	16

Note. Many participant schools failed to fill in the background questionnaire; thus, a lot of demographic information was missing. PAA = periodic accountability assessment; No = no, not participating; Yes = yes, participating.

[a] PAA-A includes PAA-A1 and PAA-A2 forms, and PAA-B includes PAA-B1 and PAA-B2 forms. [b] Low socioeconomic status, based on participation in free or reduced-price lunch program.

Table 9

Test Sequence and Cluster

Sequence no.	Time 1	Time 2
	Cluster A	
1	Ban Ads	Mango Street
2	Wind Power (A1)	Seasons (B2)
3	Ban Ads	Seasons (B2)
4	Wind Power (A1)	Mango Street
	Cluster B	
1	Ban Ads	Mango Street
5	Seasons (B2)	Wind Power (A1)
6	Ban Ads	Wind Power (A1)
7	Seasons (B2)	Mango Street
	Cluster C	
8	Mango Street	Ban Ads
9	Seasons (B1)	Wind Power (A2)
10	Mango Street	Wind Power (A2)
11	Seasons (B1)	Ban Ads
	Cluster D	
8	Mango Street	Ban Ads
12	Wind Power (A2)	Seasons (B1)
13	Mango Street	Seasons (B1)
14	Wind Power (A2)	Ban Ads

Table 10

Sample Sizes of Test Sequences

Test sequence	Total	PAA-A Section I (Wind Power)	PAA-A Section I (Seasons)	PAA-A1 Section II	PAA-A2 Section II	PAA-B1 Section II	PAA-B2 Section II	Mango lead-in (Tasks 1–3)	Mango essay (Task 4)	Ban Ads lead-in (Tasks1–3)	Ban Ads essay (Task 4)
1	514							460	383	492	437
2	246	238	224	218			212				
3	238		221				211			224	204
4	253	244		229				230	193		
5	254	218	240	191			225				
6	252	218		192						244	221
7	249		241				227	210	170		
8	409							399	367	392	338
9	212	195	206		189	203					
10	215	204			186			210	197		
11	217		213			202				210	185
12	169	164	161		158	143					
13	186		179			150		184	176		
14	162	159			151					160	129
Total	3,576	1,640	1,685	830	684	698	875	1,693	1,486	1,722	1,514

Note. PAA = periodic accountability assessment.

Classical Item Analyses

Rater Agreement for Human-Scored Items

According to Tables 2, 3, 5, and 6, each PAA had two to five human-scored items with a total of 16 items for the four PAAs. For each human-scored item, about 90% of the total responses were scored by only one rater, and about 10% of the total responses were scored by at least two raters (for purposes of evaluating rater agreement). For those responses scored by multiple raters, a third rater scored a reading item if the first two raters' scores were not the same, or a writing item if the difference between the first two raters' scores (before score weights were applied) was larger than one point. All the raters were familiar with the CBAL tests. For purposes of evaluating rater agreement, only the first two raters' scores were used. Omit scores were treated as 0 and not-reached as missing. Students receiving any missing rater score on a human-scored item were excluded from the analysis on that item.

Table 11 shows the weighted kappa coefficient for each human-scored item as a measure of interrater agreement between the first two raters, the sample size used in each kappa calculation, the asymptotic standard error estimate (ASE) of each weighted kappa coefficient, and the percentage of exact rater agreement. The weights used for the kappa calculations were the Fleiss-Cohen weights (commonly known as quadratic weights; Fleiss & Cohen, 1973). The quadratic weight for a pair of raters with score difference d was $1-d^2/k^2$, where k was the score difference between the highest score category and the lowest score category of an item. The quadratic weighting gives smaller weight to raters' scores having larger differences, ranging between weight 1 for the same scores and weight 0 for scores having the maximum possible difference, to represent the severity of disagreement. For dichotomous items, the weighted kappa coefficients were the same as the unweighted kappa coefficients. The weighted kappa coefficient in this case is equivalent to the intraclass correlation coefficient as demonstrated in Fleiss and Cohen (1973). The weighted kappa coefficients were in the range of .62 to .89. One possible interpretation of kappa is as follows (Altman, 1991, p. 404): poor agreement = less than .20, fair agreement = .20 through.40, moderate agreement = .40 through .60, good agreement = .60 through .80, and very good agreement = .80 through 1.00.

Therefore, all the human-scored items showed good to very good agreement between the first two raters. The actual percentages of exact rater agreement ranged from 48% to 97%.

Table 11

Weighted Kappa Coefficient and Percentage of Exact Agreement

Human-scored item	Number of score categories	Sample size	Weighted kappa[a]	ASE of kappa	Pct. exact agreement
BA_01B	3	346	.75	.02	67
BA_01C	3	333	.68	.03	72
BA_03	5	346	.84	.02	67
BA_04_I	6	316	.87	.01	72
BA_04_III	6	316	.79	.03	66
MG_02_01	5	344	.61	.03	48
MG_03_06	4	353	.68	.03	53
MG_04_I	6	318	.88	.02	80
MG_04_III	6	315	.75	.03	64
WP_31	2	337	.89	.03	94
WP_33	3	337	.96	.01	95
WP_43	3	324	.96	.01	94
WP_44	3	340	.84	.02	82
WP_54	3	333	.81	.03	78
SS_11	2	339	.89	.04	97
SS_14	2	356	.69	.04	85

Note. ASE = asymptotic standard error; pct = percent.

[a] Quadratic weights (Fleiss & Cohen, 1973).

Item Summary Statistics

To be consistent among all responses of human-scored items, the first rater's score was treated as the final score of a human-scored item. Tables B1 through B5 in Appendix B list the item-score frequencies including the frequencies of omit and not-reached items, as well as system errors (i.e., the online testing system failed to capture a student's response) for the four PAAs and reading linking sets, respectively. Tables 13 through 18 contain item summary statistics for the four PAAs and reading linking sets, respectively, including the following statistics: sample size (*N*), mean, standard deviation, maximum possible score point, *p*+ value, item-total polyserial correlation, item-total Pearson correlation, mean and standard deviation of item response time, percentage of omit, percentage of not reached, percentage of system error, and percentage of nonresponses (sum of percentages of omit, not reached, and system error), as well as item flags, which, as defined in Table 12, single out items with extreme item statistics to be reviewed. At the bottom of Tables 13 through 18, summary statistics across items, mean, standard deviation, minimum, and maximum, are also provided. Note that omit was treated as zero across the analyses in this study, while not reached and system error were treated as missing. A composite score including any missing item score was designated as missing.

Table 12

Item Flag Definition

Flag value	Reasons for flagging	Criterion: Dichotomous	Criterion: Polytomous
A	Low average item score	*p*+ < .25	*p*+ < .30
H	High average item score	*p*+ > .95	*p*+ > .70
R	Low item-total polyserial or Pearson correlation	Item-total polyserial correlation < .30	Item-total polyserial correlation < .60
		Item-total Pearson correlation < .20	
O	High percentage of omits	Percentage of omits > 5%	
N	High percentage of not reached	Percentage of not reached > 5%	
P	High percentage of nonresponses	Percentage of nonresponses > 5%	

The correlation between an item score and the total score is used to indicate the association between an item and the construct (represented by total score) that it measures; this index is closely related to test reliability. In this case, the polyserial correlation is preferred to the ordinary Pearson correlation because the polyserial correlation more closely reflects the actual relationship between an ordinal variable and a continuous underlying variable, while the Pearson correlation tends to underestimate this relationship (Garson, 2012). The polyserial correlation assumes that the ordinal variable has an underlying normal distribution and that the two variables follow a bivariate normal distribution.

Tables 13 through 18 provide both polyserial and Pearson item-total correlations because convergence was not reached for some polyserials during estimation. For each reading linking item, the polyserial and Pearson item-total correlations were calculated with the respective total scores of PAA-A, PAA-B, and the linking set (C1 or C2) in which this item was located. One can see that all polyserials were higher than their Pearson correlation counterparts. Most items had adequate item-total correlations; a few items with low item-total correlations were indicated by the flag of R in the column Flag. One item BA_01A_02 had a polyserial correlation of -.17

and was excluded from all the subsequent analyses and reports of summary item statistics. The mean item-total polyserial correlations for Ban Ads, Mango Street, PAA-A, and PAA-B were .48, .63, .60, and .54, respectively. For the reading linking items, the mean item-total polyserial correlations with PAA-A and PAA-B were .58 and .57, respectively.

For a dichotomous item the *p*+ value refers to the proportion of correct responses and is the same as the mean, whereas for a polytomous item the *p*+ statistic is calculated as the ratio of the mean to the maximum possible score. The *p*+ values for Ban Ads and Mango Street were between .13 and .87 with averages of .59 and .60, respectively; thus, the two writing tests had similar difficulties. However, in Ban Ads the standard deviation of item *p*+ values (0.22) was slightly larger than that in Mango Street (0.14). The item *p*+ values for PAA-A, PAA-B, and linking Blocks C1 and C2 were between .13 and .87 with the averages .60, .58, .55, and .55, respectively, which indicates that PAA-A, PAA-B, and the linking sets had similar difficulties. In addition, the standard deviations of the item *p*+ values in the four groups of items were similar in the range from 0.17 to 0.13.

Tables 13 to 18 show that the nonresponse rates were small (no more than 3.72%), which indicates that test speededness was not an issue. Students spent 15 and 13 minutes on average out of the time limit of 45 minutes on the essays of Ban Ads and Mango Street, respectively.

Table 19 shows the summary statistics of correlations between item scores and item response times for all items, as well as separately for selected response items (including SR and C&C) and constructed response items (including CR and SCR) within each PAA and the reading linking sets. One can see that the correlations varied across items, and on average they were quite small and close to 0 except for the CR and SCR items in the two writing tests, which had mean correlations of .41 and .38 for Ban Ads and Mango Street. Note that in this case item response time was just a rough estimate of how much time a student spent on an item because, for example, the computer could not separate the time a student spent in reading a passage from that the student actually answered the item. Another study using eye tracking techniques is underway and will provide more insights regarding the relationship between item response time and item score, which is valuable information for test developers evaluating items.

Table 13

Ban Ads: Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,445)	Pearson correlation (*N* = 1,445)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
BA_01A_01	1,718	.15	.36	1	.15	.17	.11	79	66	.06	.00	.06	.12	AR
BA_01A_02	1,718	.52	.50	1	.52	-.17	-.18	34	36	.00	.00	.06	.06	R
BA_01A_03	1,719	.43	.50	1	.43	.31	.24	27	33	.00	.00	.00	.00	
BA_01A_04	1,718	.13	.34	1	.13	.46	.31	17	23	.00	.06	.00	.06	A
BA_01A_05	1,718	.87	.34	1	.87	.36	.22	17	28	.00	.06	.00	.06	
BA_01B	1,716	.94	.95	3	.31	.69	.64	271	148	.47	.17	.00	.64	
BA_01C	1,685	.78	.77	3	.26	[a]	.56	249	120	1.28	1.98	.00	3.26	A
BA_02AX_A	1,717	.77	.42	1	.77	.45	.31	106	62	.00	.00	.06	.06	
BA_02AX_B	1,717	.84	.37	1	.84	.38	.24	105	58	.00	.00	.06	.06	
BA_02AX_C	1,717	.59	.49	1	.59	.21	.17	105	58	.00	.00	.06	.06	R
BA_02AX_D	1,717	.85	.36	1	.85	.57	.34	105	58	.00	.00	.06	.06	
BA_02AX_E	1,717	.81	.39	1	.81	.55	.36	105	58	.00	.00	.06	.06	
BA_02AX_F	1,717	.84	.37	1	.84	.48	.30	105	58	.00	.00	.06	.06	
BA_02AX_G	1,717	.72	.45	1	.72	.37	.28	105	58	.00	.00	.06	.06	
BA_02AX_H	1,717	.87	.33	1	.87	.59	.33	105	58	.00	.00	.06	.06	
BA_02AX_I	1,717	.76	.43	1	.76	.50	.34	105	58	.00	.00	.06	.06	
BA_02AX_J	1,717	.60	.49	1	.60	.32	.25	105	58	.00	.00	.06	.06	
BA_02BX_A	1,717	.61	.49	1	.61	.21	.16	26	20	.00	.06	.00	.06	R
BA_02BX_B	1,715	.57	.49	1	.57	.43	.34	22	20	.06	.17	.00	.23	
BA_02BX_C	1,714	.74	.44	1	.74	.53	.37	21	25	.00	.23	.00	.23	
BA_02BX_D	1,713	.63	.48	1	.63	.56	.43	18	20	.00	.29	.00	.29	
BA_02BX_E	1,713	.46	.50	1	.46	.40	.32	24	24	.00	.29	.00	.29	
BA_02BX_F	1,713	.76	.43	1	.76	.51	.36	20	26	.00	.29	.00	.29	
BA_03	1,706	2.58	2.36	8	.32	.79	.76	319	152	.00	.00	.00	.00	
BA_04_I	1,485	6.82	3.22	15	.45	.89	.86	919	573	.07	.00	.00	.07	
BA_04_III	1,485	7.45	3.28	15	.50	.89	.86	919	573	.00	.00	.00	.00	
Mean [b]		1.26	.76	2.56	.59	.48	.38	160	97	.08	.14	.03	.25	
BA_04_I	1,485	6.82	3.22	15	.45	.89	.86	919	573	.07	.00	.00	.07	

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,445)	Pearson correlation (*N* = 1,445)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
BA_04_III	1,485	7.45	3.28	15	.50	.89	.86	919	573	.00	.00	.00	.00	
Mean [b]		1.26	.76	2.56	.59	.48	.38	160	97	.08	.14	.03	.25	
BA_04_I	1,485	6.82	3.22	15	.45	.89	.86	919	573	.07	.00	.00	.07	
BA_04_III	1,485	7.45	3.28	15	.50	.89	.86	919	573	.00	.00	.00	.00	
Mean [b]		1.26	.76	2.56	.59	.48	.38	160	97	.08	.14	.03	.25	
SD [b]		1.78	.83	3.94	.22	.19	.20	237	145	.26	.39	.03	.63	
Min [b]		.13	.33	1	.13	.17	.11	17	20	.00	.00	.00	.00	
Max [b]		7.45	3.28	15	.87	.89	.86	919	573	1.28	1.98	.06	3.26	

Note. A = low average score; R = low item-total polyserial or Pearson correlation; Pct. = percentage.

[a] Item-total polyserial correlation did not converge. [b] Excluded BA_01A_02.

Table 14

Mango Street: Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,397)	Pearson correlation (*N* = 1,397)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
MG_01_01	1,688	.44	.48	1	.44	.48	.39	142	122	.00	.00	.30	.30	
MG_01_02	1,660	.67	.45	1	.67	.59	.49	45	54	.00	1.71	.24	1.95	
MG_01_03	1,651	.68	.46	1	.68	.37	.28	47	52	.00	2.30	.18	2.48	
MG_01_04	1,645	.78	.42	1	.78	.62	.43	27	46	.00	2.84	.00	2.84	
MG_01_05	1,630	.49	.48	1	.49	.63	.52	62	55	.00	3.72	.00	3.72	
MG_02_01	1,687	3.90	1.62	8	.49	.75	.70	245	147	.00	.00	.00	.00	
MG_03_01	1,681	.72	.45	1	.72	.53	.39	37	31	.00	.00	.18	.18	
MG_03_02	1,682	.79	.41	1	.79	.72	.49	32	26	.06	.00	.12	.18	
MG_03_03	1,681	.62	.49	1	.62	.58	.45	33	30	.06	.06	.12	.24	
MG_03_04	1,682	.79	.41	1	.79	.66	.46	27	25	.00	.12	.00	.12	
MG_03_05	1,679	.62	.49	1	.62	.50	.39	45	33	.06	.12	.18	.36	
MG_03_06	1,681	1.40	.92	3	.47	.63	.59	116	82	.24	.18	.00	.42	

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,397)	Pearson correlation (*N* = 1,397)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
MG_04_I	1,467	4.52	1.90	10	.45	.85	.80	789	576	.00	.00	.00	.00	
MG_04_III	1,467	4.17	1.86	10	.42	.85	.79	789	576	.00	.00	.00	.00	
Mean		1.47	.77	2.93	.60	.63	.51	174	132	.03	.79	.09	.91	
SD		1.50	.57	3.54	.14	.14	.16	267	191	.06	1.29	.11	1.26	
Min		.44	.41	1	.42	.37	.28	27	25	.00	.00	.00	.00	
Max		4.52	1.90	10	.79	.85	.80	789	576	.24	3.72	.30	3.72	

Note. Pct. = percentage.

Table 15

Reading PAA-A (Wind Power): Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,421)	Pearson correlation (*N* = 1,421)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
WP_11	1,640	.55	.50	1	.55	.63	.50	144	88	.00	.00	.00	.00	
WP_12	1,637	.66	.47	1	.66	.71	.54	118	69	.00	.00	.18	.18	
WP_13	1,637	1.28	.90	2	.64	.60	.50	82	59	.00	.00	.18	.18	
WP_14	1,636	.45	.50	1	.45	.44	.35	87	51	.00	.00	.24	.24	
WP_21	1,637	.50	.50	1	.50	.46	.37	72	55	.00	.06	.12	.18	
WP_22	1,638	.68	.46	1	.68	.61	.47	36	27	.00	.06	.06	.12	
WP_23	1,633	.34	.47	1	.34	.13	.10	32	28	.00	.06	.37	.43	R
WP_24	1,639	.58	.49	1	.58	.38	.31	22	26	.00	.06	.00	.06	
WP_31	1,635	.53	.50	1	.53	.61	.49	134	78	.18	.30	.00	.49	
WP_32	1,633	.73	.45	1	.73	.67	.48	41	32	.00	.30	.12	.43	
WP_33	1,635	1.22	.76	2	.61	.69	.63	156	95	.12	.30	.00	.43	
WP_34	1,634	1.37	.68	2	.69	.67	.59	88	50	.00	.37	.00	.37	
WP_41	1,630	1.19	.70	2	.59	.60	.54	80	47	.00	.55	.06	.61	
WP_42	1,629	.26	.44	1	.26	.40	.30	57	42	.00	.67	.00	.67	
WP_43	1,627	1.18	.86	2	.59	.64	.57	91	60	.00	.79	.00	.79	
WP_44	1,626	.66	.78	2	.33	.70	.61	142	77	.30	.85	.00	1.16	
WP_51	1,620	.32	.47	1	.32	.66	.50	67	51	.00	1.22	.00	1.22	
WP_52	1,611	.13	.34	1	.13	.68	.41	60	46	.00	1.52	.24	1.77	A

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,421)	Pearson correlation (*N* = 1,421)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
WP_53	1,605	.59	.49	1	.59	.60	.47	49	29	.00	2.01	.12	2.13	
WP_54	1,600	.85	.86	2	.43	.78	.70	159	94	.30	2.44	.00	2.74	
A02	1,500	.70	.46	1	.70	.66	.49	63	45	.00	.00	.13	.13	
A03	1,501	1.34	.79	2	.67	.72	.64	34	29	.00	.00	.07	.07	
A04	1,502	.73	.44	1	.73	.61	.45	20	20	.00	.00	.00	.00	
A05	1,501	.40	.49	1	.40	.49	.38	53	36	.00	.00	.07	.07	
A06	1,502	.50	.50	1	.50	.52	.42	37	31	.00	.00	.00	.00	
A07	1,502	.80	.40	1	.80	.81	.54	19	23	.00	.00	.00	.00	
A08	1,502	.76	.43	1	.76	.51	.38	28	21	.00	.00	.00	.00	
A09	1,502	.43	.49	1	.43	.37	.30	67	48	.00	.00	.00	.00	
A10	1,502	.63	.48	1	.63	.77	.60	24	29	.00	.00	.00	.00	
A11	1,497	.48	.50	1	.48	.73	.58	20	18	.00	.33	.00	.33	
A12	1,496	.14	.35	1	.14	.52	.33	35	26	.00	.33	.07	.40	A
A13	1,497	.67	.47	1	.67	.71	.54	15	14	.00	.33	.00	.33	
A14	1,497	.40	.49	1	.40	.43	.34	29	25	.00	.33	.00	.33	
A15	1,497	.68	.47	1	.68	.73	.55	15	12	.07	.33	.00	.40	
A16	1,495	1.07	.85	2	.53	.71	.65	44	36	.00	.40	.07	.47	
A17	1,495	.54	.50	1	.54	[a]	.57	22	21	.00	.40	.07	.47	
A18	1,496	.85	.36	1	.85	.74	.46	11	13	.00	.40	.00	.40	
A01	1,495	.76	.43	1	.76	.40	.29	15	22	.00	.47	.00	.47	
Mean		.68	.54	1.24	.55	.60	.47	60	41	.03	.39	.06	.48	
SD		.32	.16	.43	.17	.15	.13	43	23	.08	.56	.09	.60	
Min		.13	.34	1	.13	.13	.10	11	12	.00	.00	.00	.00	
Max		1.37	.90	2	.85	.81	.70	159	95	.30	2.44	.37	2.74	

Note. A = low average score; R = low item-total polyserial or Pearson correlation; Pct. = percentage.

[a] Missing cell: Item-total polyserial correlation did not converge.

Table 16

Reading PAA-B (Seasons): Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Poly-serial (*N* = 1,472)	Pearson correlation (*N* = 1472)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
SS_11	1,684	.84	.37	1	.84	.68	.44	424	299	.12	.00	.00	.12	
SS_12	1,683	.50	.50	1	.50	.61	.49	53	74	.00	.00	.06	.06	
SS_13	1,682	1.14	.83	2	.57	[a]	.55	122	94	.00	.06	.06	.12	
SS_14	1,682	.54	.50	1	.54	.45	.36	132	90	.18	.12	.00	.30	
SS_15	1,679	.35	.48	1	.35	.55	.43	59	48	.00	.18	.12	.30	
SS_16	1,673	.43	.50	1	.43	.55	.44	66	51	.00	.30	.36	.65	
SS_17	1,676	.51	.50	1	.51	.55	.44	51	41	.00	.36	.12	.48	
SS_18	1,675	.38	.49	1	.38	.54	.42	50	38	.00	.42	.12	.53	
SS_19	1,677	1.22	.88	2	.61	.63	.55	43	30	.00	.42	.00	.42	
SS_21	1,675	1.51	.78	2	.76	.62	.50	84	55	.00	.53	.00	.53	
SS_22	1,674	.77	.70	2	.39	.36	.32	29	22	.00	.53	.06	.59	
SS_23	1,672	.98	.64	2	.49	.36	.32	96	55	.00	.71	.00	.71	
SS_24	1,664	1.46	.61	2	.73	.35	.30	38	28	.00	1.13	.06	1.19	
SS_25	1,664	.41	.49	1	.41	.64	.50	59	41	.00	1.19	.00	1.19	
SS_31	1,660	.53	.50	1	.53	.40	.32	33	29	.00	1.31	.12	1.43	
SS_32	1,651	.60	.49	1	.60	[a]	.57	37	26	.00	1.48	.48	1.96	
SS_41	1,655	1.17	.91	2	.59	.34	.29	83	58	.00	1.66	.06	1.72	
SS_42	1,650	.70	.77	2	.35	.21	.19	27	23	.00	2.02	.00	2.02	R
SS_43	1,644	1.08	.92	2	.54	.64	.56	22	24	.00	2.14	.24	2.38	
SS_44	1,646	.73	.87	2	.36	.48	.41	24	31	.00	2.26	.00	2.26	
B02	1,563	.73	.44	1	.73	.69	.52	92	60	.00	.00	.26	.26	
B03	1,567	.64	.48	1	.64	.65	.51	15	19	.00	.00	.00	.00	
B04	1,566	.87	.34	1	.87	.74	.46	18	18	.00	.00	.06	.06	
B05	1,566	.66	.47	1	.66	.54	.42	57	42	.00	.00	.06	.06	
B06	1,566	.99	.88	2	.50	.64	.56	33	30	.00	.06	.00	.06	
B07	1,564	1.34	.79	2	.67	.59	.51	47	33	.00	.13	.06	.19	
B08	1,564	.62	.49	1	.62	.29	.23	29	26	.00	.13	.06	.19	R
B09	1,564	.71	.45	1	.71	.53	.40	53	41	.00	.19	.00	.19	
B10	1,561	.65	.48	1	.65	.60	.47	24	24	.00	.19	.19	.38	
B11	1,544	.76	.43	1	.76	.70	.51	16	15	.00	1.34	.13	1.47	
B12	1,545	.57	.50	1	.57	.66	.53	37	27	.00	1.40	.00	1.40	
B13	1,542	.77	.42	1	.77	.67	.48	18	16	.00	1.53	.06	1.60	

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Poly-serial (*N* = 1,472)	Pearson correlation (*N* = 1472)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
B14	1,542	.38	.48	1	.38	.21	.17	17	14	.06	1.53	.06	1.66	R
B15	1,541	.69	.46	1	.69	.65	.49	17	13	.00	1.60	.06	1.66	
B16	1,542	.45	.50	1	.45	.59	.47	35	28	.00	1.60	.00	1.60	
B17	1,541	.56	.50	1	.56	.78	.62	17	19	.00	1.66	.00	1.66	
B18	1,538	.78	.42	1	.78	.63	.45	14	14	.00	1.79	.06	1.85	
B01	1,537	.56	.50	1	.56	.32	.26	14	21	.00	1.91	.00	1.91	
Mean		.75	.57	1.32	.58	.54	.43	55	43	.01	.84	.08	.93	
SD		.30	.17	.47	.14	.15	.11	68	47	.04	.76	.11	.76	
Min		.35	.34	1	.35	.21	.17	14	13	.00	.00	.00	.00	
Max		1.51	.92	2	.87	.78	.62	424	299	.18	2.26	.48	2.38	

Note. R = low item-total polyserial or Pearson correlation; Pct. = percentage.

[a]Missing cell: Item-total polyserial correlation did not converge.

Table 17

Linking Block C1: Item Statistics

Item score ID	*N*	Mean	SD	Max. possible score	*p*+	Poly-serial[a] (*N* = 1,502)	Poly-serial A[b] (*N* = 786)	Poly-serial B[c] (*N* = 635)	Pearson correlation[a] (*N* = 1,502)	Pearson correlation A[c] (*N* = 786)	Pearson correlation B[c] (*N* = 635)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
C05	1,520	.73	.45	1	.73	.71	.65	.57	.53	.48	.43	43	38	.00	.07	.07	.13	
C06	1,518	.48	.50	1	.48	.46	.36	.32	.37	.29	.26	33	24	.00	.13	.13	.26	
C10	1,518	1.67	.60	2	.84	.77	[d]	[d]	.62	.52	.52	55	37	.00	.20	.07	.26	H
C08	1,516	1.10	.85	2	.55	.69	.59	.47	.62	.52	.42	61	43	.00	.33	.07	.39	R
C01	1,515	.62	.49	1	.62	.65	.59	.55	.52	.46	.44	19	17	.00	.33	.13	.46	
C13	1,516	.97	.61	2	.49	.53	.35	.39	.47	.31	.33	52	34	.00	.33	.07	.39	R
C15	1,515	.66	.91	2	.33	.78	.60	.52	.61	.47	.41	86	53	.00	.33	.13	.46	R
C16	1,516	1.34	.86	2	.67	.84	.73	.77	.75	.62	.66	65	39	.00	.39	.00	.39	
C22	1,514	.43	.49	1	.43	[d]	.68	.58	.56	.54	.46	49	41	.00	.53	.00	.53	
C23	1,512	.40	.49	1	.40	[d]	.61	.60	.58	.48	.47	22	21	.00	.59	.07	.66	

Item score ID	*N*	Mean	SD	Max. possible score	*p*+	Poly-serial[a] (*N* = 1,502)	Poly-serial A[b] (*N* = 786)	Poly-serial B[c] (*N* = 635)	Pearson correlation[a] (*N* = 1,502)	Pearson correlation A[c] (*N* = 786)	Pearson correlation B[c] (*N* = 635)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
Mean		.84	.63	1.5	.55	.68	.57	.53	.56	.47	.44	48	34	.00	.32	.07	.39	
SD		.42	.18	.53	.16	.13	.13	.13	.10	.10	.11	20	11	.00	.16	.05	.15	
Min		.40	.45	1	.33	.46	.35	.32	.37	.29	.26	19	17	.00	.07	.00	.13	
Max		1.67	.91	2	.84	.84	.73	.77	.75	.62	.66	86	53	.00	.59	.13	.66	

Note. H = high average item score; R = low item-total polyserial or Pearson correlation; Pct. = percentage.

[a] Polyserial or Pearson item correlation with the total score of Block C1. [b] Polyserial or Pearson item correlation with the total score of PAA-A. [c] Polyserial or Pearson item correlation with the total score of PAA-B. [d] Polyserial item-total correlation did not converge.

Table 18

Linking Block C2: Item Statistics

Item score ID	*N*	Mean	SD	Max. possible score	*p*+	Poly-serial[a] (*N* = 1,502)	Poly-serial A[b] (*N* = 786)	Poly-serial B[c] (*N* =635)	Pearson correlation[a] (*N* = 1,502)	Pearson correlation A[c] (*N* = 786)	Pearson correlation B[c] (*N* = 635)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
C02	1,545	.84	.36	1	.84	.89	.77	.83	.55	.46	.53	36	34	.00	.13	.00	.13	
C03	1,545	.58	.49	1	.58	.74	.65	.59	.58	.51	.47	38	31	.00	.13	.00	.13	
C04	1,540	.95	.96	2	.47	.82	.74	.66	.70	.61	.56	60	41	.00	.39	.06	.45	
C07	1,541	.75	.43	1	.75	.67	.47	.62	.49	.34	.46	33	27	.00	.39	.00	.39	
C11	1,540	.39	.49	1	.39	.53	.40	.43	.42	.32	.34	58	39	.00	.45	.00	.45	
C12	1,539	.35	.48	1	.35	.55	.48	.47	.43	.37	.36	19	19	.00	.52	.00	.52	
C14	1,538	1.14	.72	2	.57	.55	.39	.45	.50	.35	.40	53	35	.00	.58	.00	.58	R
C18	1,538	.58	.49	1	.58	.73	.55	.62	.58	.44	.49	37	30	.00	.58	.00	.58	
C19	1,535	1.17	.92	2	.59	.83	.69	.74	.74	.60	.65	75	51	.00	.78	.00	.78	
C20	1,533	1.18	.79	2	.59	.75	.62	.64	.69	.56	.58	40	28	.00	.90	.00	.90	
C21	1,532	.77	.88	2	.39	.78	.62	.65	.68	.55	.57	52	41	.00	.97	.00	.97	
Mean		.79	.64	1.45	.55	.71	.58	.61	.58	.46	.49	45	34	.00	.53	.01	.53	

Item score ID	*N*	Mean	SD	Max. possible score	*p*+	Poly-serial[a] (*N* = 1,502)	Poly-serial A[b] (*N* = 786)	Poly-serial B[c] (*N* =635)	Pearson correlation[a] (*N* = 1,502)	Pearson correlation A[c] (*N* = 786)	Pearson correlation B[c] (*N* = 635)	Mean item time (sec.)	SD item time (sec.)	Pct. omit	Pct. not reached	Pct. system error	Pct. non-response	Flag
SD		.30	.22	.52	.15	.12	.13	.12	.11	.11	.10	16	9	.00	.28	.02	.27	
Min		.35	.36	1	.35	.53	.39	.43	.42	.32	.34	19	19	.00	.13	.00	.13	
Max		1.18	.96	2	.84	.89	.77	.83	.74	.61	.65	75	51	.00	.97	.06	.97	

Note. R = low item-total polyserial or Pearson correlation; Pct. = percentage.

[a] Polyserial or Pearson item correlation with the total score of Block C2. [b] Polyserial or Pearson item correlation with the total score of PAA-A. [c] Polyserial or Pearson item correlation with the total score of PAA-B.

Table 19

Summary of Correlations Between Item Score and Item Response Time

	All items					SR and C&C items					CR and SCR items				
PAA	*N*	Mean	SD	Min	Max	*N*	Mean	SD	Min	Max	*N*	Mean	SD	Min	Max
Ban Ads	25	.09	.18	-.20	.53	20	.02	.07	-.20	.20	5	.41	.14	.21	.53
Mango Street	14	.07	.24	-.30	.55	10	-.06	.11	-.30	.06	4	.38	.13	.25	.55
PAA A	38	.03	.13	-.23	.40	35	.03	.12	-.21	.40	3	-.01	.28	-.23	.30
PAA B	38	.03	.11	-.12	.34	35	.03	.11	-.12	.34	3	.04	.14	-.09	.20
Block C1+C2	21	.15	.13	-.10	.40	21	.15	.13	-.10	.40					

Note. PAA = periodic accountability assessment; SR = selected response; C&C = click and click; CR = constructed response, SCR = short CR.

Table 20 shows the average item *p*+ values by required skill level, where that designation refers to the categorization in terms of the CBAL reading competency model. Over all items collapsing across PAAs (including linking items), as well as for the items on PAA-A alone, the average item *p*+ values decreased as item skill level increased, a result theoretically in keeping with the CBAL reading competency model categorizations. However, for PAA-B the average item *p*+ values increased from Level 2 to Level 3 for item skills, an inconsistent result that might suggest that the classification of items needs to be refined.

Table 20

Average Reading Item p+ Value by Item Skill Level

	Level 1		Level 2		Level 3	
Test form	Number of items	Mean *p*+ (max. *N*)	Number of items	Mean *p*+ (max. *N*)	Number of items	Mean *p*+ (max. *N*)
PAA-A	7	.65 (1,640)	11	.53 (1,637)	14	.45 (1,639)
PAA-B	10	.62 (1,684)	13	.51 (1,682)	9	.55 (1,677)
All (PAA-A + PAA-B + linking sets C1 and C2)	22	.63 (1,684)	32	.53 (1,682)	29	.48 (1,677)

Note. PAA = periodic accountability assessment.

Differential Item Functioning

Test fairness requires that all test items be fair to all students. Differential item functioning (DIF) analysis is designed to identify items that may have biases against certain student groups. That is, if students having the same ability but from different demographic groups perform differently on an item, then this item shows DIF. DIF in an item may indicate that it measures some construct different from what it is intended to measure. For an item deemed to have DIF, further review by content experts is needed, and depending on the outcome of the review the item may be kept as it is, revised, or discarded.

In this study, the Mantel-Haenszel procedure (Dorans & Holland, 1993; Holland & Thayer, 1988) and the standardized mean difference (Zwick, Donoghue, & Grima, 1993) were used to detect DIF for dichotomous and polytomous items, respectively. ETS DIF procedures (Dorans & Holland, 1993; Zwick et al., 1993) result in classification of items into three categories: A, B, and C. Category A items contain negligible DIF, Category B items exhibit slight or moderate DIF, and Category C items have moderate to large values of DIF. In practice, only Category C items are considered to have substantial DIF and are designated for further review and/or revision.

The DIF analyses were conducted for the following demographic group pairs:

1. gender (male vs.female)
2. race/ethnicity (White vs. Black; White vs. Asian/Pacific Islander except for reading linking items in PAA-A2 and PAA-B1 due to small sample sizes, and White vs. Hispanic)
3. low SES students (no vs. yes)

DIF analyses were not conducted for other demographic groups (i.e., Native American, ELL, test accommodation status) because of small sample sizes.

Table 21 lists the Category C DIF items, and the tables in Appendix C show the DIF category for every item. There were four items in reading PAA-A (note that WP_13 had DIF on two different pairs of groups), one item in reading PAA-B, and four items in the reading linking sets having Category C DIF. And there were no Category C DIF items in the two writing PAAs. Note that some groups had small sample sizes, fewer than 200 (See Tables C1–C5 in Appendix C). Therefore, their DIF results should be interpreted with caution.

Table 21

Category C DIF Items

Item score ID	C DIF description
	Reading PAA-A
WP_12	Favor male over female
WP_13	Favor male over female
WP_13	Favor white over black
A02	Favor female over male
A13	Favor female over male
	Reading PAA-B
B10	Favor male over female
	Reading linking sets
C08	Favor male over female in PAA-A
C16	Favor White over Hispanic in PAA-B
C19	Favor Hispanic over White in PAA-B
C21	Favor female over male in PAA-A

Statistics for Subscores and Total Scores

In this section we present the summary statistics (sample size, mean, and standard deviation), reliabilities (standardized Cronbach alpha[1]), and correlations of subscores and total raw scores.

Tables 22 and 23 show the statistics for the subscores and total raw scores of the two writing PAAs. These tests were moderately difficult as their mean total scores were 52% of the maximum possible scores. The subscores had one through 10 mutually exclusive items and reliabilities ranging from .21 to .88. The subscore reliabilities in Ban Ads varied more than those in Mango Street. For each PAA, the subscore computed from the essay had the highest reliability. Note that each essay subscore contained two scores measuring different aspects of the same essay. The intersubscore correlations were between .25 and .56. The correlations between subscores and total scores ranged from .40 to .91.

Table 22

Ban Ads: Test Subscore and Total Score Summary and Correlations

Score[a]	Number of items	Max poss. score	*N*	Mean	SD	Mean % correct	Standardized alpha[b]	Pearson correlation S1	S2	S3	S4	S5	S6
Subscore 1	4	4	1,717	1.59	.85	40	.21						
Subscore 2	2	4	1,685	1.73	1.46	43	.61	.28					
Subscore 3	10	10	1,717	7.64	1.99	76	.65	.25	.42				
Subscore 4	6	6	1,713	3.77	1.50	63	.49	.27	.49	.40			
Subscore 5	1	8	1,706	2.58	2.36	32		.29	.56	.41	.48		
Subscore 6	2	30	1,485	14.27	6.14	48	.88	.26	.56	.38	.41	.56	
Total	25	62	1,446	32.37	10.81	52	.80	.40	.72	.60	.62	.76	.91

[a] See Table 2 for subscore information. [b] Reliability was not calculated for a subscore with one item.

Table 23

Mango Street: Test Subscore and Total Score Summary and Correlations

Score[a]	Number of items	Max poss. score	*N*	Mean	SD	Mean % correct	Standardized alpha[b]	Pearson correlation S1	S2	S3	S4
Subscore 1	5	5	1,619	3.08	1.43	62	.61				
Subscore 2	1	8	1,687	3.90	1.62	49		.44			
Subscore 3	6	8	1,672	4.94	1.98	62	.67	.56	.46		
Subscore 4	2	20	1,467	8.69	3.39	43	.77	.45	.48	.54	
Total	14	41	1,397	21.27	6.45	52	.81	.69	.70	.78	.88

[a] See Table 3 for subscore information. [b] Reliability was not calculated for a subscore with one item.

Table 24 shows the statistics for the raw scores of the linking block (C1 or C2), the operational discrete item block (A or B), Section I (*Seasons* or *Wind Power*), and the main reading PAA form (PAA-A or PAA-B) within each of the four reading PAA forms (PAA-A1, PAA-A2, PAA-B1, and PAA-B2). The linking blocks had relatively strong relationships with the operational forms: The correlations of those blocks with the operational discrete item blocks, Section I, and the main PAA forms ranged from .72 to .84. The operational discrete item blocks also had high correlations with the Section I scenario-based task set, ranging from .73 to .78.

Table 24

Test Section Score Summary and Correlations Within Each Reading PAA Form

						Pearson correlation		
Score	Number of items	Max poss. score	*N*	Mean	SD	PAA-A	Wind Power	Block A
		PAA-A1						
PAA-A (Wind Power + Block A)	38	47	787	26.51	10.24			
Wind Power	20	27	890	13.98	6.35	.96		
Block A	18	20	818	11.87	4.75	.93	.78	
Block C1	10	15	822	8.25	3.72	.82	.75	.81
		PAA-A2						
PAA-A (Wind Power + Block A)	38	47	634	26.48	9.73			
Wind Power	20	27	687	14.31	5.80	.95		
Block A	18	20	670	11.91	4.66	.93	.78	
Block C1	11	16	675	9.04	4.20	.81	.73	.81
						PAA-B	Seasons	Block B
		PAA-B1						
PAA-B (Seasons + Block B)	38	50	645	29.98	9.05			
Seasons	20	30	724	16.38	5.47	.94		
Block B	18	20	672	13.30	4.39	.91	.73	
Block C1	10	15	680	8.59	3.50	.80	.72	.78
		PAA-B2						
PAA-B (Seasons + Block B)	38	50	827	28.06	9.56			
Seasons	20	30	901	15.51	5.78	.95		
Block B	18	20	849	12.29	4.52	.91	.74	
Block C2	11	16	856	8.47	4.29	.84	.77	.80

Note. PAA = periodic accountability assessment.

Tables 25 and 26 present the statistics for the subscores and total scores of the reading PAA-A and PAA-B respectively. One can see that, for both PAAs, the intersubscore correlations for subscores with mutually exclusive items were between .41 and .72, and the correlations of subscores with total scores were between .67 and .97. The reliabilities (standardized Cronbach alpha) for subscores were between .51 and .87, and the reliabilities for the PAA-A and PAA-B total scores were .91 and .88, respectively.

Table 25

Reading PAA-A: Test Subscore and Total Score Summary and Correlations

Score[a]	Number of items	Max poss. score	*N*	Mean	SD	Mean % correct	Standard-ized alpha	Pearson correlation[b]					
								S1	S2	S3	S4	S5	S6
Subscore 1	19	25	1,436	14.34	6.03	57	.86						
Subscore 2	7	9	1,448	3.73	2.20	41	.61	.72					
Subscore 3	6	7	1,619	4.07	1.65	58	.51	.62	.50				
Subscore 4	6	6	1,495	4.29	1.54	72	.62	.70	.55	.49			
Subscore 5	22	29	1,436	15.28	6.81	53	.87	*.97*	*.83*	.62	.68		
Subscore 6	2	3	1,499	2.07	1.07	69	.65	*.70*	.46	.41	.54	.58	
Total	38	47	1,421	26.50	10.01	56	.91	.97	.82	.72	.77	.97	.67

Note. PAA = periodic accountability assessment; poss = possible.

[a] See Table 5 for subscore information. [b] Italicized correlations contain items that are mapped to both subscores.

Table 26

Reading PAA-B: Test Subscore and Total Score Summary and Correlations

Score[a]	Number of items	Max poss. score	*N*	Mean	SD	Mean % correct	Standard-ized alpha	Pearson correlation[b]				
								S1	S2	S4	S5	S6
Subscore 1	24	30	1,501	17.10	6.02	57	.84					
Subscore 2	8	14	1,515	7.45	3.13	53	.59	.67				
Subscore 4	6	6	1,531	4.29	1.44	72	.55	.63	.46			
Subscore 5	10	12	1,535	7.10	2.99	59	.73	*.87*	.66	.59		
Subscore 6	22	32	1,508	17.44	6.03	55	.81	*.91*	*.86*	.58	.71	
Total	38	50	1,472	28.90	9.38	58	.88	.96	.83	.71	.87	.96

Note. PAA = periodic accountability assessment; poss = possible.

[a] See Table 6 for subscore information. [b] Italicized correlations contain items that are mapped to both subscores.

Tables 27 and 28 show the statistics for the task scores and total scores of the *Wind Power* and *Seasons* task sets, respectively. The intertask correlations were low to moderate in the range between .31 and .64, and the task-total correlations were moderate to high from .58 to .86. The reliabilities for task scores (within these task sets) were between .32 and .71, and the reliabilities for the *Wind Power* and *Seasons* task sets were .85 and .79, respectively. In Tables

25 through 28, total score means were between 52% and 58% of their maximum possible scores, indicating the tests/task sets were moderately difficult for the samples assessed.

Table 27

Wind Power: Task Score and Total Score Summary and Correlations

Score	Number of items	Max poss. score	*N*	Mean	SD	Mean % correct	Standardized alpha	Pearson correlation				
								Task 1	Task 2	Task 3	Task 4	Task 5
Task 1	4	5	1,633	2.94	1.58	59	.55					
Task 2	4	4	1,630	2.11	1.11	53	.33	.35				
Task 3	4	6	1,632	3.85	1.70	64	.65	.57	.44			
Task 4	4	7	1,625	3.28	1.87	47	.56	.52	.36	.64		
Task 5	4	5	1,594	1.90	1.52	38	.60	.53	.37	.62	.62	
Total	20	27	1,577	14.12	6.11	52	.85	.77	.59	.85	.84	.81

Table 28

Seasons: Task Score and Total Score Summary and Correlations

Score	Number of items	Max. poss. score	*N*	Mean	SD	Mean % correct	Standardized alpha	Pearson correlation			
								Task 1	Task 2	Task 3	Task 4
Task 1	9	11	1,663	5.92	2.83	54	.71				
Task 2	5	9	1,662	5.15	1.81	57	.44	.49			
Task 3	2	2	1,650	1.13	.77	57	.33	.44	.36		
Task 4	4	8	1,641	3.69	2.01	46	.32	.42	.35	.31	
Total	20	30	1,625	15.90	5.66	53	.79	.86	.74	.58	.72

Table 29 shows that the correlations among the total scores of Ban Ads, Mango Street, reading PAA-A, and reading PAA-B were between .66 and .80. Table 29 also displays comparisons of the standardized alphas based on item scores and task scores. (Note that the discrete item block A or B was treated as one task set in the two reading PAAs). For Ban Ads and Mango Street, the alphas based on item scores were close to those based on task scores (commonly known as testlet reliability) with the differences of .02 and .03, respectively, which indicates that testlet effects at the task level were minor for these two writing PAAs. However, for the two reading PAAs, the differences were .08 and .16 for PAA-A and PAA-B, respectively, suggesting that there were some task-level testlet effects.

Table 29

Total Score Summary and Correlations

	Standardized alpha		Pearson correlation		
Total raw score	Task	Item	Ban Ads	Mango Street	PAA-A
Ban Ads	.82	.80			
Mango Street	.78	.81	.71		
PAA-A	.83	.91	.77	.67	
PAA-B	.72	.88	.66	.72	.80

Note. PAA = periodic accountability assessment.

Analyses of Factors Affecting Test Scores

The effects of PAA, test order, teacher instruction, and demographic groups on test raw scores and/or theta estimates were evaluated using *t*-tests, one-way ANOVA, multiple comparisons, correlation, and mixed models.

The item response theory (IRT) model used to calibrate the writing and reading tests was the two-dimensional generalized partial credit model with a simple structure, that is, the reading tests loaded on the reading dimension and the writing tests on the writing dimension. The two dimensions were allowed to be correlated. This was a concurrent calibration, and a student in two test occasions was treated as two different students. Items to which a student did not respond were treated as missing responses. Expected a posteriori (i.e., EAP) theta estimates were obtained after item parameters were estimated with marginal maximum likelihood by means of a stabilized Newton-Raphson algorithm that uses adaptive quadrature (Haberman, 1988, 2006). A computer program developed by Haberman (2011) was used for the IRT estimation. For the reading tests, theta estimates were based only on operational items (i.e., the external linking items were excluded), while for item parameter estimates all items were used. See van Rijn, Fu, and Wise (2012) for the details of the calibrations of these tests using the IRT model.

Subgroup Comparison

Table 30 provides *t*-test results as well as means and standard deviations of raw scores and theta estimates on each PAA for gender, SES, ELL, and test accommodation status. Statistically significant differences were found for all these demographic groups across the four PAAs. The male, economically disadvantaged, ELL, and test accommodation groups had significantly lower test scores than their respective comparison groups across the four PAAs.

Table 30

Subgroup Comparison on Each PAA

			Theta				Raw score			
Subgroup	Category	*N*	Mean	SD	*t*	*p* value	Mean	SD	*t*	*p* value
Ban Ads										
Gender	M	673	.31	.77	-4.35	.00**	30.52	10.92	-6.86	.00**
	F	718	.48	.70			34.44	10.43		
Low SES	Y	499	.19	.71	9.30	.00**	29.01	10.04	10.66	.00**
	N	784	.57	.71			35.39	10.69		
ELL status	Y	31	-.18	.70	-4.55	.00**	22.42	7.89	-7.46	.00**
	N	1,180	.43	.74			33.24	10.91		
Test accommodation	Y	81	-.31	.66	9.51	.00**	21.33	8.05	13.20	.00**
	N	1,121	.47	.72			33.86	10.68		
Mango Street										
Gender	M	631	.45	.81	-6.98	.00**	19.63	6.62	-9.13	.00**
	F	720	.74	.70			22.78	5.98		
Low SES	Y	491	.42	.79	7.58	.00**	19.40	6.44	9.27	.00**
	N	747	.75	.72			22.79	6.21		
ELL status	Y	22	.05	.77	-3.53	.00**	16.14	6.31	-3.88	.00**
	N	1,154	.63	.77			21.58	6.52		
Test accommodation	Y	76	-.09	.84	8.46	.00**	15.05	6.12	9.12	.00**
	N	1,094	.67	.74			21.91	6.36		
Reading PAA-A										
Gender	M	739	.15	.97	-3.07	.00**	13.70	6.08	-3.05	.00**
	F	782	.30	.98			14.65	6.08		
Low SES	Y	552	-.13	.90	12.65	.00**	11.77	5.73	13.88	.00**
	N	841	.52	.95			16.11	5.69		
ELL status	Y	30	-.65	.96	-5.23	.00**	8.63	6.14	-5.36	.00**
	N	1,293	.30	.98			14.61	6.04		
Test accommodation	Y	93	-.59	.91	9.03	.00**	8.91	5.44	9.34	.00**
	N	1,223	.34	.97			14.87	5.97		
Reading PAA-B										
Gender	M	750	.09	.85	-5.76	.00**	15.07	5.77	-6.36	.00**
	F	786	.34	.82			16.88	5.40		
Low SES	Y	557	-.08	.74	12.24	.00**	14.02	5.06	12.21	.00**
	N	837	.45	.84			17.54	5.60		
ELL status	Y	36	-.26	.79	-3.81	.00**	12.67	5.50	-3.90	.00**
	N	1,284	.27	.83			16.34	5.57		
Test accommodation	Y	99	-.37	.74	7.77	.00**	12.15	5.00	7.59	.00**
	N	1,219	.30	.83			16.52	5.54		

Note. Race had four subgroups to be compared. (Note that Native American was not included in the comparison because of the small sample size.) SES = socioeconomic status; ELL = English language learner; PAA = periodic accountability assessment.

** *p* < .01.

The one-way ANOVAs were first carried out on ethnic groups for theta estimates and raw scores on each PAA. Levene's tests (Levene, 1960) show the group variances in all ANOVA tests were not significantly different at the .01 level. All the one-way ANOVA tests were statistically significant as shown in Table 31. Therefore, multiple comparisons (Tukey HSD test) were conducted on all pairs of racial/ethnic groups.

Table 31

Race Subgroup Comparison on Each PAA

		Theta				Theta: multiple comparison[a]			Raw score				Raw score: multiple comparison[a]		
Race	*N*	Mean	SD	*F*	*p* value	1	2	3	Mean	SD	*F*	*p* value	1	2	3
Ban Ads															
1	116	.67	.72						37.91	10.62					
2	165	.17	.71	17.84	.00**	*			28.83	9.84	29.68	.00**	*		
3	896	.46	.73			*	*		33.49	10.81			*	*	
4	212	.20	.73			*		*	28.50	9.79			*		*
Mango Street															
1	108	.79	.71						22.75	6.51					
2	164	.35	.79	13.72	.00**	*			19.60	6.38	15.70	.00**	*		
3	875	.67	.77				*		21.92	6.46				*	
4	199	.43	.74			*		*	19.20	5.95			*		*
Reading PAA-A															
1	117	.67	.97						17.07	5.62					
2	159	-.07	.91	26.18	.00**	*			11.98	5.79	32.22	.00**	*		
3	987	.31	.96			*	*		14.77	5.94			*	*	
4	252	-.10	.94			*		*	11.92	5.99			*		*
Reading PAA-B															
1	107	.46	.78						17.89	5.54					
2	181	-.01	.83	21.90	.00**	*			14.34	5.40	23.96	.00**	*		
3	987	.31	.85				*		16.62	5.67				*	
4	254	-.06	.74			*		*	14.08	5.00			*		*

Note. PAA = periodic accountability assessment. 1 = Asian/Pacific Islander; 2 = African American; 3 = White; 4 = Hispanic.

[a] Tukey HSD test.

* $p < .05$, ** $p < .01$.

The group pairs having significant differences are shown in Table 31. Table 31 also provides the means and standard deviations of the theta estimates and raw scores for each racial/ethnic group on each PAA. One can see that across the four PAAs, African American and Hispanic students performed similarly, with the lowest mean scores. For Ban Ads and reading PAA-A, Asian/Pacific Islander students performed better than White students, while for Mango Street and reading PAA-B, the two groups were not measurably different.

Correlations Between Instructional Coverage and Test Scores

At each administration, teachers were asked to fill out a questionnaire. For two of the questions, teachers were asked to rate the extent to which they covered specific reading and writing content categories during the last two months, on a one-to-four scale (1 = not at all, 2 = small extent, 3 = moderate extent, 4 = large extent). A composite score related to the coverage of each PAA was then created by summing teachers' ratings on relevant content as follows:

- Ban Ads: informational essays, editorials, and speeches
- Mango Street: stories, poems, song lyrics, journal entries, and book reviews
- Reading PAA-A: exposition, argumentation & persuasion, and procedural texts & documents
- Reading PAA-B: fiction, poetry, and other types of literature

To check the relationships between teacher instructional ratings and student test scores, correlations were calculated between teacher ratings and mean student test scores (thetas and raw scores) by teacher across test occasions. For a test at Time 1, the teacher rating at Time 1 was used, whereas a test at Time 2 used the teacher rating at Time 2.

Summary information for numbers of students by teacher and ratings is shown in Table 32. The correlations are listed in Table 33. One can see that the numbers of students and the ratings varied among teachers. Except for one case, the ratings did not appear to have any linear relationship with mean test scores, as all but one of the correlations were not significantly different from zero. (The exception was the correlation between the teacher ratings for Reading PAA-A and thetas on Mango Street, which had little substantive relationship to one another.)

Table 32

Distributions of Numbers of Students by Teacher and Instructional Ratings

		Number of students by teacher				Teacher rating			
PAA	Number of teachers	Mean	SD	Min	Max	Mean	SD	Min	Max
Ban Ads-related	40	24	13	5	53	6.75	2.00	4	11
Mango-related	47	22	12	3	53	11.83	3.53	6	20
Reading PAA-A-related	46	25	13	3	52	7.74	1.76	4	12
Reading PAA-B-related	41	26	12	3	54	8.76	1.58	6	12

Note. PAA = periodic accountability assessment.

Table 33

Pearson Correlations Between Teacher Instructional Ratings and Mean Test Scores by Teacher

	Theta				Raw score			
PAA (*N*)	Ban Ads	Mango	PAA-A	PAA-B	Ban Ads	Mango	PAA-A	PAA-B
Ban Ads-related (40)	.07	.20	.08	.10	.05	.11	.04	.08
Mango-related (47)	.16	.24	-.04	.08	.08	.14	-.03	.06
PAA-A-related (46)	.11	.36*	.08	.12	.01	.20	.06	.11
PAA-B-related (41)	.07	.07	-.05	-.04	-.09	-.04	-.05	-.08

Note. PAA = periodic accountability assessment.
**p* < .05.

Mixed Models

Analyses by van Rijn et al. (2012) using the same response data as in this study found that the reading and writing PAAs were on different dimensions. Therefore, for the current analysis, mixed models were built for the reading and writing tests separately, with sequences including both tests within the same subject (i.e., Ban Ads/Mango and Mango/Ban Ads for writing, and PAA-B1/PAA-A2, PAA-B2/PAA-A1, PAA-A1/PAA-B2, and PAA-A2/PAA-B1 for reading). The test sequences with one reading and one writing test were excluded from the mixed models.

Mixed models were built in three stages, with more independent variables added into the models during each stage. In all models, the theta estimate on each PAA (i.e., Ban Ads, Mango Street, Reading PAA-A, or Reading PAA-B) was the dependent variable. In the first stage, the random effects were school, teacher-within-school, and student-within-teacher, and the fixed effects were PAA (Ban Ads vs. Mango Street, or Reading PAA-A vs. Reading PAA-B), test order (Time 1 or Time 2), and their interaction effect. Because the interaction was not significant for both the writing and reading models, it was dropped. In addition, the model comparisons showed that school was not a significant random effect in both the writing and reading models; thus, it was also dropped.

The final model estimates in the first stage for writing and reading are shown in Tables 34 and 35, respectively. These final models indicate that both PAA and test-order effects were significant for writing, while only test-order effect was significant for reading.

Table 36 shows the means and standard deviations of theta estimates by PAA and test order. One can see that for writing, students performed better on Mango Street than Ban Ads in either test order and, overall, students performed better on the first test than the second test; however, the test-order effect mainly came from Ban Ads, that is, students taking Ban Ads at Time 1 had thetas significantly higher than students taking the test at Time 2. For reading, in general students did better on the first test than the second test; however, again, the significant test-order effect mainly came from one test: Students performed better on Reading PAA-A at Time 1 than those taking Reading PAA-A at Time 2.

In the second stage, teacher ratings were added into the models as independent variables. In particular, the main effects of Ban Ads-related and Mango-related ratings at Times 1 and 2, as well as their three-way interaction effects with PAA and test order were added into the final writing model from Stage 1. Similarly, for the reading model, the main effects of PAA-A-related and PAA-B-related ratings at Times 1 and 2, as well as their three-way interaction effects with PAA and test order were inserted. None of these effects was significant at the .01 level. Only one effect, the three-way interaction of Ban Ads-related ratings at Time 1, PAA, and test order, was significant at the .05 level (p = .03). For this effect, the regression coefficients show that a teacher with high Ban Ads-related ratings at Time 1 had significantly lower mean theta score on Ban Ads than that on Mango Street at Time 2, which is not of interest and fully understandable. These results are consistent with the low correlations between ratings and mean test scores by teacher as shown in Table 33. Therefore, all the effects related to teacher instructional ratings were dropped from the models in the second stage.

In the third stage, the following background variables were added as fixed main effects into the final models from the second stage: (a) five student demographic variables compared in the section above (i.e., gender, SES, ELL, test accommodation, and race); (b) three school variables (i.e., percentage of free/reduced price lunch, percentage of minority, and percentage of student over teacher); and (c) one teacher variable (years teaching English). The final models were selected by sequentially dropping statistically nonsignificant background variable(s) and are shown in Tables 37 and 38 for the writing and reading PAAs, respectively. For the writing model, the main effects of PAA, gender, race, SES, test accommodation, and percentage of free/reduced price lunch were statistically significant, while test order became nonsignificant. For the reading model, the main effects of test order, gender, SES, and test accommodation were statistically significant, while PAA remained as nonsignificant.

Table 34

Mixed Model for Writing PAA and Test Order Effects (N = 1,261)

Fixed effect	Numerator *df*	Denominator *df*	*F*	*p* value	Random effect	Variance
PAA	1	526	54.18	.00	Teacher	.11
Order	1	526	7.27	.01	Student nested in Teacher	.25
					Residual	.23

Note. PAA = periodic accountability assessment.

Table 35

Mixed Model for Reading PAA and Test Order Effects (N = 1,370)

Fixed effect	Numerator *df*	Denominator *df*	*F*	*p* value	Random effect	Variance
PAA	1	624	.08	.78	Teacher	.20
Order	1	624	28.60	.00	Student nested in Teacher	.41
					Residual	.28

Note. PAA = periodic accountability assessment.

Table 36

Means and Standard Deviations of Theta Estimates by Test Order and PAA

	Writing						Reading					
	Ban Ads		Mango		Total		PAA-A		PAA-B		Total	
Test order	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	.47	.79	.60	.78	.53	.79	.41	1.00	.21	.79	.31	.91
2	.32	.71	.64	.75	.49	.75	.09	1.00	.24	.93	.17	.96
Total	.40	.76	.62	.76	.51	.77	.26	1.01	.22	.86	.24	.94

Note. PAA = periodic accountability assessment.

Table 37

Mixed Model With Subgroup Comparisons for Writing PAAs (N = 1,097)

Fixed effect	Numerator *df*	Denominator *df*	*F*	*p* value	Random effect	Variance
PAA	1	462	35.99	.00	Teacher	.05
Order	1	462	3.31	.07	Student nested in Teacher	.20
Gender	1	462	6.77	.01	Residual	.22
Race	3	462	3.33	.02		
SES	1	462	10.03	.00		
Test accommodation	1	462	38.01	.00		
Percentage of free/ reduced price lunch	1	462	7.04	.01		

Note. PAA = periodic accountability assessment; SES = socioeconomic status.

Table 38

Mixed Model With Subgroup Comparisons for Reading PAAs (N = 1,187)

Fixed effect	Numerator *df*	Denominator *df*	*F*	*p* value	Random effect	Variance
PAA	1	541	.46	.50	Teacher	.12
Order	1	541	20.44	.00	Student nested in teacher	.36
Gender	1	541	22.10	.00	Residual	.29
SES	1	541	55.02	.00		
Test accommodation	1	541	9.24	.00		

Note. PAA = periodic accountability assessment; SES = socioeconomic status.

Results of Student Survey

After taking the tests at each occasion, students completed a survey regarding their experience with CBAL tests, which contained the following four questions (with response options, 1 = yes, 2 = somewhat, and 3 = no):

Q1 – Was the test hard for you?

Q2 – Did you have enough time?

Q3 – Did you try your best?

Q4 – Did you find the test more interesting than a typical test you take?Table 39 shows the sample size, mean, and standard deviation for each question on each test form for each test administration as well as the two administrations combined. In general, students reported that the CBAL tests were not too hard, they had enough test time, and they tried their best. In addition, they felt these CBAL tests were moderately more interesting than a traditional test. Finally, by comparing the two administrations, students appeared to try harder and feel the tests were more interesting at Time 1 than at Time 2.

Table 39

Means and Standard Deviations of Student Survey Questions

	Q1			Q2			Q3			Q4		
Test form	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD
Time 1												
Writing												
Ban Ads	885	2.37	.61	885	1.23	.52	885	1.35	.57	885	1.86	.82
Mango Street	749	2.44	.60	749	1.14	.43	749	1.24	.49	749	1.71	.78
Reading												
PAA-A1	452	2.38	.61	452	1.13	.41	452	1.46	.60	452	1.88	.81
PAA-A2	315	2.42	.57	314	1.06	.31	313	1.29	.52	312	1.71	.71
PAA-A1 + A2	767	2.40	.59	766	1.10	.38	765	1.39	.58	764	1.81	.77
PAA-B1	404	2.37	.58	404	1.14	.46	404	1.38	.58	404	1.81	.80
PAA-B2	453	2.29	.63	452	1.21	.53	452	1.44	.60	452	2.04	.81
PAA-B1 + B2	857	2.33	.61	856	1.18	.50	856	1.41	.59	856	1.93	.81
Time 2												
Writing												
Ban Ads	686	2.36	.65	686	1.20	.51	686	1.41	.60	686	1.94	.80
Mango Street	853	2.44	.65	853	1.17	.47	853	1.50	.66	853	2.01	.81
Reading												
PAA-A1	405	2.36	.62	405	1.21	.54	405	1.58	.66	405	2.28	.76
PAA-A2	380	2.34	.66	380	1.19	.52	380	1.62	.72	380	2.10	.82
PAA-A1 + A2	785	2.35	.64	785	1.20	.53	785	1.60	.69	785	2.19	.79
PAA-B1	302	2.46	.61	300	1.13	.43	300	1.45	.63	298	2.01	.79
PAA-B2	434	2.31	.66	434	1.17	.48	434	1.63	.70	433	2.06	.83
PAA-B1 + B2	736	2.37	.64	734	1.15	.46	734	1.56	.68	731	2.04	.81
Combined												
Writing												
Ban Ads	1,571	2.37	.63	1,571	1.22	.51	1,571	1.38	.58	1,571	1.89	.82
Mango Street	1,602	2.44	.63	1,602	1.16	.45	1,602	1.38	.60	1,602	1.87	.81
Reading												
PAA-A1	857	2.37	.61	857	1.17	.48	857	1.52	.63	857	2.07	.81
PAA-A2	695	2.38	.62	694	1.13	.44	693	1.47	.66	692	1.92	.79
PAA-A1 + A2	1,552	2.37	.62	1,551	1.15	.46	1,550	1.50	.64	1,549	2.00	.81
PAA-B1	706	2.41	.59	704	1.13	.45	704	1.41	.61	702	1.89	.80
PAA-B2	887	2.30	.64	886	1.19	.51	886	1.53	.66	885	2.05	.82
PAA-B1 + B2	1,593	2.35	.62	1,590	1.17	.48	1,590	1.48	.64	1,587	1.98	.81

Note. PAA = periodic accountability assessment.

Table 40 shows the Pearson correlations[2] between survey items and test raw scores for each PAA by administration, as well as both administrations combined. All the correlations were low (none larger than .32 in absolute value), but most were statistically significant due to the large sample sizes. For Q1, all the statistically significant correlations were positive, while for other questions all the statistically significant correlations were negative. Surprisingly, students who achieved higher scores tended to be slightly more likely to report finding the CBAL tests to be harder, less interesting, more speeded, and less motivating than students who obtained lower scores.

Table 40

Pearson Correlations Between Student Survey Questions and PAA Raw Scores

Test form	Q1	*N*	Q2	*N*	Q3 (	*N*)	Q4	*N*
				Time 1				
Ban Ads	.18**	815	00	815	-.18**	815	-.12**	815
Mango Street	.19**	671	-.13**	671	-.20**	671	-.05	671
Reading PAA-A	.32**	721	-.12**	720	-.19**	720	-.08*	720
Reading PAA-B	.27**	832	-.21**	831	-.18**	831	-.05	831
				Time 2				
Ban Ads	.20**	624	.00	624	-.19**	624	-.10*	624
Mango Street	.23**	706	-.11**	706	-.17**	706	-.11**	706
Reading PAA-A	.25**	743	-.13**	743	-.17**	743	-.07	743
Reading PAA-B	.28**	688	-.14**	687	-.27**	687	-.13**	686
				Combined				
Ban Ads	.19**	1,439	.01	1,439	-.18**	1,439	.11**	1,439
Mango Street	.21**	1,377	-.12**	1,377	-.19**	1,377	-.08**	1,377
Reading PAA-A	.28**	1,464	-.14**	1,463	-.20**	1,463	-.11**	1,463
Reading PAA-B	.27**	1,520	-.18**	1,518	-.22**	1,518	-.08**	1,517

Note. PAA = periodic accountability assessment.

*p < .05. **p < .01.

Summary

Some basic psychometric properties of the CBAL writing and reading PAAs in the 2011 multistate administration were presented. The main findings are as follows:

1. The classical item statistics show all items performed reasonably well except for one item, BA_01A_02, which had a polyserial correlation of -.17 with the total test

scores and was removed from the test analyses. For the human-scored items, about 10% of the total responses were scored by two or three raters, and the weighted kappa coefficients showed good to very good rater agreement. The missing response rates were no more than 3.72%, indicating students had enough time to complete the tests. The item skill-level classification for the reading PAA-A was reasonable; however, for PAA-B the items or their classifications could be improved as the Level 2 items were unexpectedly more difficult than the Level 3 items. There were four items in the reading PAA-A, one item in the reading PAA-B, and four items in the reading linking sets having Category C DIF.

2. The correlations between item response times and item scores varied across items. For CR and SCR items in Ban Ads and Mango Street, the correlations were moderate with means .41 and .38, respectively.

3. Ban Ads and Mango Street were moderately difficult and had reliabilities (standardized Cronbach alpha of test raw scores) of .80 and .81, respectively. These reliability estimates were close to the testlet reliabilities based on task scores, indicating that dependency among items within a task did not appear to have significant effects on the two PAAs. The subscores' reliabilities ranged from .21 to .88, the intersubscore correlations were between .25 and .56, and the correlations between subscores and total scores ranged from .40 to .91.

4. The reading PAA-A and PAA-B had similar difficulty (slightly above average) and high reliabilities of .91 and .88, respectively. The reliabilities were .08 and .16 smaller than the testlet reliability for PAA-A and PAA-B, respectively, suggesting that there were some testlet effects at the task level. The subscores' reliabilities were between .51 to .87, the intersubscore correlations for subscores with mutually exclusive items ranged between .41 and .72, and the correlations of subscores with total scores were between .67 and .97. The correlations among the total scores of Ban Ads, Mango Street, PAA-A, and PAA-B were between .66 and .80. Strong correlations and similar distributions of item $p+$ values between the reading linking sets and the operational forms (PAA-A and PAA-B) were indicators of well-performing linking sets.

5. The *t*-test and one-way ANOVA analyses showed that for the five demographic variables, gender, SES, ELL, test accommodation, and race/ethnicity, the subgroup means on the test raw and theta scores of all PAAs in each demographic variable were significantly different. The results from the multilevel models for the writing theta scores were that PAA, gender, race, SES, test accommodation, and percent free/reduced price lunch were statistically significant factors. Similarly, for the reading theta scores the multilevel results show that test order, gender, SES, and test accommodation were statistically significant. For both subjects, student and teacher were statistically significant random effects. Both the correlations and multilevel models indicate that teachers' instructional ratings had no statistically significant effects on thetas and/or raw scores in either subject.
6. The student survey was conducted after each test administration. The results show that, on average, students reported having enough time, trying their best to answer the questions, and perceiving the tests as not too hard and moderately more interesting than an ordinary test.

However, the limitation of this study should be noted here. The study was based on a convenient sample recruited across the nation. A more representative and larger sample is needed to verify the results reported here. Especially for the mixed models, another validation sample different from the one used for model selections is needed to cross-validate the final model. However, this is infeasible in the current study because of the small sample sizes used.

References

Altman, D. G. (1991). *Practical statistics for medical research*. London, England: Chapman & Hall.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8,* 70–91.

CBAL ELA Team. (2011). *Reading summative final report 2010*. Unpublished manuscript, ETS, Princeton, NJ.

Deane, P. (2011). *Writing assessment and cognition* (ETS Research Report No. RR-11-14). Princeton, NJ: ETS.

Deane, P., Fowles, M., Baldwin, D., & Persky, H. (2011). *The CBAL summative writing assessment: A draft eight-grade design* (ETS Research Memorandum No. RM-11-01). Princeton, NJ: ETS.

Deane, P., Fowles, M., Persky, H., Baldwin, D., Cooper, P., Ecker, M., …, Wagner, M. (2009). *Progress on designing the CBAL summative writing assessment: Design principles and results.* Unpublished manuscript, ETS, Princeton, NJ.

Dorans, N. J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613–619.

Garson, G. D. (2012). *Correlation*. Asheboro, NC: Statistical Associates Publishers. Retrieved from http://www.statisticalassociates.com/correlation.htm

Graf, E. A. (2009). *Defining mathematics competency in the service of cognitively based assessment for grades 6 through 8* (ETS Research Report No. RR-09-42). Princeton, NJ: ETS.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology, 18,* 193–211.

Haberman, S. J. (2006). *Adaptive quadrature for item response models* (ETS Research Report No. RR-06-29). Princeton, NJ: ETS.

Haberman, S. J. (2011). *A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm*. Unpublished manuscript, ETS, Princeton, NJ.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Levene, H. (1960). Robust tests for the equality of variance. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278–292). Palo Alto, CA: Stanford University Press.

O'Reilly, T., & Sheehan, K. M. (2009a). *Cognitively based assessment of, for, and as learning: A framework for assessing reading competency* (ETS Research Report No. RR-09-26). Princeton, NJ: ETS.

O'Reilly, T., & Sheehan, K. M. (2009b). *Cognitively based assessment of, for, and as learning: A 21st century approach for assessing reading competency* (ETS Research Memorandum No. RM-09-04). Princeton, NJ: ETS.

van Rijn, P., Fu, J., & Wise, M. (2012, January). *2011 ELA multi-state results update.* Paper presented at the meeting of CBAL technical advisory committee, Princeton, NJ.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251

Notes

[1] The reason for using standardized alpha is to remove the impact of item variances. Note that in the two writing PAAs, item scores had various score ranges and thus their score variances varied considerably.

[2] Polyserial correlations were not used here because the assumption of normality of survey item scores seemed too strong in this case.

List of Appendices

Page

Appendix A

Data Cleaning

The 2011 CBAL multistate writing and reading data were cleaned by the following steps:

1. For each test record, counted the number of valid responses (i.e., not an omit, not-reached, system error, or blank) for the whole test and each section for reading tests or each task for writing tests. If the number of valid responses in a test section for a reading test or a task for a writing test was zero, the item scores in this section/task were set to blanks. Removed a student record if the item responses in both tests were blanks, and there was no such a record.
2. Because there were negative, zero, and extreme large item response times due to computer glitches, item response times were cleaned up.
   a. If item response times were equal to or smaller than 0, they were set to missing.
   b. If item response times in the respective test section were equal to or larger than the times (in seconds) shown below, they were set to missing; an exception was for the first item in a reading test section or writing task.

Ban Ads lead-in section (Tasks 1-3):	700
Mango Street lead-in section (Tasks 1-3):	642
PAA-A Section I (Wind Power):	539
PAA-A1 Section II:	320
PAA-A2 Section II:	374
PAA-B Section I (Seasons):	565
PAA-B1 Section II:	346
PAA-B2 Section II:	300

   c. There was one case in Wind Power where the section response time was larger than the time limit (3,000 seconds). In this case, the item response time for WP_12 was 400 seconds. This item response time was adjusted by subtracting time for WP_11, which brought down the section response time under the time limit.
3. A student's item scores in a test section were set to missing if that student met the criterion in Table A1. The criteria were used to judge students' absolute motivation levels; that is, a student's item scores in a test section were removed if that student

completed this section in a very short time and received a very low score. These criteria were relatively conservative.

Table A1

Response Time and Score Criteria to Remove a Test Section

Test section	Time limit (sec.)	Max total score	Time criterion (smaller than or equal to; sec.)	Score criterion [a] (smaller than or equal to)	Total number of students	Number of students meeting the criterion
PAA-A Section I (Wind Power)	3,000	27	600	5	1,688	35
PAA-B Section I (Season)	3,000	30	600	5	1,712	19
PAA-A1 Section II	3,000	35	300	6	865	30
PAA-B1 Section II	3,000	35	300	6	708	7
PAA-A2 Section II	3,000	36	300	6	701	11
PAA-B2 Section II	3,000	36	300	6	895	15
Ban Ads: Lead-in section (Tasks 1-3)	2,700	34	600	6	1,737	8
Ban Ads: Essay section (Task 4)	2,700	30	200	5	1,595	76
Mango: Lead-in section (Tasks 1-3)	2,700	21	300	3.5	1,714	17
Mango: Essay section (Task 4)	2,700	20	150	4	1,630	142

Note. PAA = periodic accountability assessment.

[a] Weighted scores for writing tests.

4. Ran a simple regression with Time 1 test raw score as independent variable and Time 2 test raw score as dependent variable; excluded linking items from total scores; and excluded any students having any not-reached or blank score in Time 1 or Time 2 tests (excluding linking sets). Removed a student's item score in a test if that student's regression residual was equal to or larger than three standard deviations of errors. Removed Test 1 if the residual was positive or Test 2 if the residual was negative. There were only 16 such cases. This step was to remove students with relatively low motivation levels (*relatively* referred to the comparison between two test occasions).

5. Removed 17 student records because their item responses in both tests were blanks.

Appendix B

Item Score Frequency Tables

Table B1

Ban Ads: Item Score Frequency

			Score																							
	Total		0		1		2		3		4		6		8		9		12		15		OM		NR		SE	
Item score ID	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
BA 01A 01	1,719	100	1,451	84	266	15																	1	0			1	0
BA_01A_02	1,719	100	824	48	894	52																					1	0
BA_01A_03	1,719	100	975	57	744	43																						
BA_01A_04	1,719	100	1,487	87	231	13																			1	0		
BA_01A_05	1,719	100	225	13	1,493	87																			1	0		
BA_01B	1,719	100	708	41	506	29	380	22	114	7													8	0	3	0		
BA_01C	1,719	100	673	39	687	40	279	16	24	1													22	1	34	2		
BA_02AX_A	1,718	100	401	23	1,316	77																					1	0
BA_02AX_B	1,718	100	274	16	1,443	84																					1	0
BA_02AX_C	1,718	100	711	41	1,006	59																					1	0
BA_02AX_D	1,718	100	262	15	1,455	85																					1	0
BA_02AX_E	1,718	100	318	19	1,399	81																					1	0
BA_02AX_F	1,718	100	280	16	1,437	84																					1	0
BA_02AX_G	1,718	100	483	28	1,234	72																					1	0
BA_02AX_H	1,718	100	217	13	1,500	87																					1	0
BA_02AX_I	1,718	100	414	24	1,303	76																					1	0
BA_02AX_J	1,718	100	688	40	1,029	60																					1	0
BA_02BX_A	1,718	100	671	39	1,046	61																			1	0		
BA_02BX_B	1,718	100	733	43	981	57																	1	0	3	0		
BA_02BX_C	1,718	100	445	26	1,269	74																			4	0		
BA_02BX_D	1,718	100	634	37	1,079	63																			5	0		
BA_02BX_E	1,718	100	922	54	791	46																			5	0		
BA_02BX_F	1,718	100	410	24	1,303	76																			5	0		
BA_03	1,706	100	608	36			338	20			483	28	215	13	62	4												
BA_04_I	1,485	100	38	3					349				484	33			414	28	179	12	20	1	1	0				
BA_04_III	1,485	100	31	2					253				477	32			460	31	218	15	46	3						

Note. OM = omit; NR = not reached; SE = system error.

Table B2

Mango Street: Item Score Frequency

			Score																							
	Total		0		.5		1		2		3		4		6		8		10		OM		NR		SE	
Item score ID	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
MG_01_01	1,693	100	913	54	77	5	698	41																	5	0
MG_01_02	1,693	100	481	28	138	8	1,041	61															29	2	4	0
MG_01_03	1,693	100	513	30	38	2	1,100	65															39	2	3	0
MG_01_04	1,693	100	364	22	6	0	1,275	75															48	3		
MG_01_05	1,693	100	784	46	97	6	749	44															63	4		
MG_02_01	1,687	100	63	4					363	22			902	53	315	19	44	3								
MG_03_01	1,684	100	465	28			1,216	72																	3	0
MG_03_02	1,684	100	359	21			1,322	79													1	0			2	0
MG_03_03	1,684	100	639	38			1,041	62													1	0	1	0	2	0
MG_03_04	1,684	100	351	21			1,331	79															2	0		
MG_03_05	1,684	100	641	38			1,037	62													1	0	2	0	3	0
MG_03_06	1,684	100	350	21			467	28	698	41	162	10									4	0	3	0		
MG_04_I	1,467	100	23	2					246	17			710	48	326	22	139	9	23	2						
MG_04_III	1,467	100	11	1					369	25			728	50	220	15	121	8	18	1						

Note. OM = omit; NR = not reached; SE = system error.

Table B3

PAA-A: Item Score Frequency

			Score													
	Total		0		1		2		OM		NR		SE			
Item score ID	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%		
WP_11	1,640	100	730	45	910	55										
WP_12	1,640	100	554	34	1,083	66							3	0		
WP_13	1,640	100	494	30	198	12	945	58					3	0		
WP_14	1,640	100	906	55	730	45							4	0		
WP_21	1,640	100	818	50	819	50					1	0	2	0		
WP_22	1,640	100	516	31	1,122	68					1	0	1	0		
WP_23	1,640	100	1,076	66	557	34					1	0	6	0		
WP_24	1,640	100	685	42	954	58					1	0				
WP_31	1,640	100	768	47	864	53			3	0	5	0				
WP_32	1,640	100	447	27	1,186	72					5	0	2	0		
WP_33	1,640	100	325	20	616	38	692	42	2	0	5	0				
WP_34	1,640	100	189	12	647	39	798	49			6	0				
WP_41	1,640	100	273	17	778	47	579	35			9	1	1	0		
WP_42	1,640	100	1,207	74	422	26					11	1				
WP_43	1,640	100	485	30	372	23	770	47			13	1				
WP_44	1,640	100	863	53	446	27	313	19	4	0	14	1				
WP_51	1,640	100	1,100	67	520	32					20	1				
WP_52	1,640	100	1,403	86	208	13					25	2	4	0		
WP_53	1,640	100	657	40	948	58					33	2	2	0		
WP_54	1,640	100	722	44	383	23	491	30	4	0	40	2				
A01	1,502	100	358	24	1,137	76					7	0				
A02	1,502	100	443	29	1,057	70							2	0		
A03	1,502	100	296	20	392	26	813	54					1	0		
A04	1,502	100	408	27	1,094	73										
A05	1,502	100	903	60	598	40							1	0		
A06	1,502	100	746	50	756	50										
A07	1,502	100	302	20	1,200	80										
A08	1,502	100	364	24	1,138	76										
A09	1,502	100	862	57	640	43										
A10	1,502	100	560	37	942	63										
A11	1,502	100	781	52	716	48					5	0				
A12	1,502	100	1,288	86	208	14					5	0	1	0		
A13	1,502	100	493	33	1,004	67					5	0				
A14	1,502	100	904	60	593	39					5	0				
A15	1,502	100	475	32	1,021	68			1	0	5	0				
A16	1,502	100	497	33	400	27	598	40			6	0	1	0		
A17	1,502	100	681	45	814	54					6	0	1	0		
A18	1,502	100	226	15	1,270	85					6	0				

Note. PAA = periodic accountability assessment; OM = omit; NR = not reached; SE = system error.

Table B4

PAA-B: Item Score Frequency

			Score											
	Total		0		1		2		OM		NR		SE	
Item score ID	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
SS_11	1,684	100	274	16	1,408	84			2	0				
SS_12	1,684	100	834	50	849	50							1	0
SS_13	1,684	100	475	28	490	29	717	43			1	0	1	0
SS_14	1,684	100	778	46	901	54			3	0	2	0		
SS_15	1,684	100	1,084	64	595	35					3	0	2	0
SS_16	1,684	100	954	57	719	43					5	0	6	0
SS_17	1,684	100	815	48	861	51					6	0	2	0
SS_18	1,684	100	1,031	61	644	38					7	0	2	0
SS_19	1,684	100	499	30	303	18	875	52			7	0		
SS_21	1,684	100	302	18	215	13	1,158	69			9	1		
SS_22	1,684	100	639	38	773	46	262	16			9	1	1	0
SS_23	1,684	100	363	22	979	58	330	20			12	1		
SS_24	1,684	100	105	6	681	40	878	52			19	1	1	0
SS_25	1,684	100	981	58	683	41					20	1		
SS_31	1,684	100	788	47	872	52					22	1	2	0
SS_32	1,684	100	657	39	994	59					25	1	8	0
SS_41	1,684	100	570	34	226	13	859	51			28	2	1	0
SS_42	1,684	100	805	48	528	31	317	19			34	2		
SS_43	1,684	100	638	38	234	14	772	46			36	2	4	0
SS_44	1,684	100	916	54	264	16	466	28			38	2		
B01	1,567	100	679	43	858	55					30	2		
B02	1,567	100	420	27	1,143	73							4	0
B03	1,567	100	570	36	997	64								
B04	1,567	100	204	13	1,362	87							1	0
B05	1,567	100	535	34	1,031	66							1	0
B06	1,567	100	611	39	355	23	600	38			1	0		
B07	1,567	100	309	20	415	26	840	54			2	0	1	0
B08	1,567	100	592	38	972	62					2	0	1	0
B09	1,567	100	454	29	1,110	71					3	0		
B10	1,567	100	547	35	1,014	65					3	0	3	0
B11	1,567	100	375	24	1,169	75					21	1	2	0
B12	1,567	100	662	42	883	56					22	1		
B13	1,567	100	359	23	1,183	75					24	2	1	0
B14	1,567	100	961	61	580	37			1	0	24	2	1	0
B15	1,567	100	472	30	1,069	68					25	2	1	0
B16	1,567	100	852	54	690	44					25	2		
B17	1,567	100	684	44	857	55					26	2		
B18	1,567	100	344	22	1,194	76					28	2	1	0

Note. PAA = periodic accountability assessment; OM = omit; NR = not reached; SE = system error.

Table B5

Linking Blocks C1 and C2: Item Score Frequency

			Score									
	Total		0		1		2		NR		SE	
Item score ID	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
Block C1												
C01	1,522	100	573	38	942	62			5	0	2	0
C05	1,522	100	417	27	1,103	72			1	0	1	0
C06	1,522	100	785	52	733	48			2	0	2	0
C08	1,522	100	483	32	404	27	629	41	5	0	1	0
C10	1,522	100	108	7	284	19	1,126	74	3	0	1	0
C13	1,522	100	305	20	944	62	267	18	5	0	1	0
C15	1,522	100	977	64	78	5	460	30	5	0	2	0
C16	1,522	100	397	26	210	14	909	60	6	0		
C22	1,522	100	868	57	646	42			8	1		
C23	1,522	100	907	60	605	40			9	1	1	0
Block C2												
C02	1,547	100	244	16	1,301	84			2	0		
C03	1,547	100	646	42	899	58			2	0		
C04	1,547	100	759	49	106	7	675	44	6	0	1	0
C07	1,547	100	387	25	1,154	75			6	0		
C11	1,547	100	936	61	604	39			7	0		
C12	1,547	100	1000	65	539	35			8	1		
C14	1,547	100	306	20	713	46	519	34	9	1		
C18	1,547	100	643	42	895	58			9	1		
C19	1,547	100	533	34	202	13	800	52	12	1		
C20	1,547	100	365	24	525	34	643	42	14	1		
C21	1,547	100	808	52	267	17	457	30	15	1		

Note. NR = not reached; SE = system error.

Appendix C

Item DIF Results

Table C1

Ban Ads: Item DIF Categories

Item score ID	Male (*N* = 673) vs. female (*N* = 718)	White (*N* = 896) vs. Black (*N* = 165)	White (*N* = 896) vs. Asian/Pacific Islander (*N* = 116)	White (*N* = 896) vs. Hispanic (*N* = 212)	Low SES: No (*N* = 784) vs. yes (*N* = 499)	Number of C DIF (if not 0)
BA_01A_01	A	A	A	A	A	
BA_01A_03	A	A	A	A	A	
BA_01A_04	A	A	A	A	A	
BA_01A_05	A	A	A	B-	A	
BA_01B	A	A	A	A	A	
BA_01C	A	B-	A	A	A	
BA_02AX_A	A	B+	A	A	A	
BA_02AX_B	B-	A	A	A	A	
BA_02AX_C	A	A	A	A	A	
BA_02AX_D	A	A	A	B-	A	
BA_02AX_E	A	A	A	A	A	
BA_02AX_F	A	A	A	A	A	
BA_02AX_G	A	A	A	A	A	
BA_02AX_H	A	A	A	A	A	
BA_02AX_I	A	A	A	A	A	
BA_02AX_J	A	A	A	A	A	
BA_02BX_A	A	A	A	A	A	
BA_02BX_B	A	A	A	A	A	
BA_02BX_C	B-	A	A	A	A	
BA_02BX_D	A	A	B+	A	A	
BA_02BX_E	A	A	A	A	A	
BA_02BX_F	A	A	A	A	A	
BA_03	A	A	A	A	A	
BA_04_I	A	A	A	A	A	
BA_04_III	A	A	A	A	A	

Note. The first group is the reference group, and the second group is the focus group. A positive sign favors the focus group, while a negative sign favors the reference group. DIF = differential item functioning; SES = socioeconomic status.

Table C2

Mango Street: Item DIF Categories

Item score ID	Male (*N* = 631) vs. female (*N* = 720)	White (*N* = 875) vs. Black (*N* = 164)	White (*N* = 875) vs. Asian/Pacific Islander (*N* = 108)	White (*N* = 875) vs. Hispanic (*N* = 199)	Low SES: No (*N* = 747) vs. yes (*N* = 491)	Number of C DIF (if not 0)
MG_01_01	B-	B-	B-	A	A	
MG_01_02	A	A	A	A	A	
MG_01_03	A	A	B+	B+	A	
MG_01_04	A	B+	A	A	A	
MG_01_05	A	B-	B+	B+	A	
MG_02_01	A	A	A	A	A	
MG_03_01	A	A	A	A	A	
MG_03_02	A	A	A	A	A	
MG_03_03	A	A	A	A	A	
MG_03_04	A	A	A	A	A	
MG_03_05	A	A	A	A	A	
MG_03_06	A	A	A	A	A	
MG_04_I	A	A	A	A	A	
MG_04_III	A	A	A	A	A	

Note. The first group is the reference group, and the second group is the focus group. A positive sign favors the focus group, while a negative sign favors the reference group. DIF = differential item functioning; SES = socioeconomic status.

Table C3

PAA-A: Item DIF Categories

Item score ID	Male (*N* = 656) vs. female (*N* = 724)	White (*N* = 909) vs. Black (*N* = 141)	White (*N* = 909) vs. Asian/ Pacific Islander (*N* = 107)	White (*N* = 909) vs. Hispanic (*N* = 218)	Low SES: No (*N* = 777) vs. yes (*N* = 502)	Number of C DIF (if not 0)
WP 11	A	B+	B+	A	A	
WP 12	C-	B-	A	A	A	1
WP 13	C-	C-	A	A	B-	2
WP 14	A	A	A	A	A	
WP 21	A	B+	A	A	A	
WP 22	A	A	A	A	A	
WP 23	A	A	B+	A	A	
WP 24	A	A	A	A	A	
WP 31	B+	A	A	A	A	
WP 32	A	A	A	A	A	
WP 33	A	A	A	B-	A	
WP 34	A	A	A	B-	A	
WP 41	A	A	A	A	A	
WP 42	A	A	A	B+	A	
WP 43	B-	A	A	A	A	
WP 44	B+	A	A	A	A	
WP 51	A	A	A	A	A	
WP 52	B+	A	A	A	A	
WP 53	A	A	A	A	A	
WP 54	A	A	A	B+	A	
A01	A	A	A	A	A	
A02	C+	A	A	A	A	1
A03	A	A	A	A	A	
A04	A	A	A	A	A	
A05	A	A	A	A	A	
A06	A	A	A	A	A	
A07	B+	B+	A	B+	A	
A08	B-	A	A	A	A	
A09	B-	A	A	A	A	
A10	A	A	A	A	A	
A11	A	A	A	A	A	
A12	A	A	A	A	A	
A13	C+	B+	B-	A	A	1
A14	A	A	A	A	A	
A15	A	A	A	A	A	
A16	B+	A	A	A	A	
A17	A	A	A	A	A	
A18	B+	A	B-	A	B+	

Note. The first group is the reference group, and the second group is the focus group. A positive sign favors the focus group, while a negative sign favors the reference group. DIF = differential item functioning; PAA = periodic accountability assessment; SES = socioeconomic status.

Table C4

PAA-B: Item DIF Categories

Item score ID	Male (*N* = 687) vs. female (*N* = 734)	White (*N* = 915) vs. Black (*N* = 165)	White (*N* = 915) vs. Asian/Pacific Islander (*N* = 101)	White (*N* = 915) vs. Hispanic (*N* = 235)	Low SES: No (*N* = 784) vs. yes (*N* = 507)	Number of C DIF (if not 0)
SS 11	A	A	A	A	A	
SS 12	A	B-	A	A	A	
SS 13	A	A	A	B+	B+	
SS 14	A	B-	B-	A	A	
SS 15	A	A	A	A	A	
SS 16	A	A	A	A	A	
SS 17	A	B-	A	B-	A	
SS 18	A	A	A	A	A	
SS 19	A	B-	A	A	B-	
SS 21	A	B+	A	A	B+	
SS 22	A	A	A	A	A	
SS 23	A	B+	A	A	A	
SS 24	A	B-	A	A	A	
SS 25	A	A	A	B-	A	
SS 31	A	A	A	A	A	
SS 32	A	A	A	A	A	
SS 41	B+	A	B-	A	A	
SS 42	B+	A	B-	A	A	
SS 43	A	B+	B+	A	A	
SS 44	A	B+	B+	A	A	
B01	A	A	B-	A	A	
B02	A	A	A	A	A	
B03	A	A	A	A	A	
B04	A	A	A	A	A	
B05	A	A	A	A	A	
B06	A	B+	B-	A	A	
B07	B-	A	A	B+	A	
B08	A	A	A	A	A	
B09	A	A	A	A	A	
B10	C-	A	A	A	A	1
B11	A	B-	A	A	B-	
B12	A	B+	A	A	A	
B13	A	A	A	A	A	
B14	A	A	A	A	A	
B15	A	A	A	A	A	
B16	A	A	A	A	A	
B17	A	A	A	A	A	
B18	A	A	A	A	A	

Note. The first group is the reference group, and the second group is the focus group. A positive sign favors the focus group, while a negative sign favors the reference group. DIF = differential item functioning; PAA = periodic accountability assessment; SES = socioeconomic status.

Table C5

Reading Linking Set C1: Item DIF Categories

Item score ID	PAA A: Male (*N* = 362) vs. female (*N* = 403)	PAA B: Male (*N* = 303) vs. female (*N* = 326)	PAA A: White (*N* = 445) vs. Black (*N* = 83)	PAA B: White (*N* = 478) vs. Black (*N* = 62)	PAA A: White (*N* = 445) vs. Asian/ Pacific Islander (*N* = 80)	PAA A: White (*N* = 445) vs. Hispanic (*N* = 153)	PAA B: White (*N* = 478) vs. Hispanic (*N* = 61)	PAA A: Low SES: No (*N* = 447) vs. yes (*N* = 286)	PAA B: Low SES: No (*N* = 340) vs. yes (*N* = 204)	Number of C DIF (if not 0)
C01	A	A	B+	A	A	B+	A	A	A	
C05	B+	A	A	A	A	A	A	A	A	
C06	A	A	A	A	B+	A	A	A	A	
C08	C-	B-	A	A	A	B-	B-	B-	A	1
C10	A	A	A	A	A	A	A	A	A	
C13	A	A	B+	A	A	A	A	A	A	
C15	A	A	B-	A	B-	A	A	A	B+	
C16	B+	B+	A	A	B-	A	C-	A	A	1
C22	A	A	A	A	A	A	A	A	A	
C23	B+	A	A	A	A	A	A	A	A	

Note. The first group is the reference group, and the second group is the focus group. A positive sign favors the focus group, while a negative sign favors the reference group. DIF = differential item functioning; PAA = periodic accountability assessment; SES = socioeconomic status.

Table C6

Reading Linking Set C2: Item DIF Categories

Item score ID	PAA A: Male (*N* = 294) vs. female (*N* = 320)	PAA B: Male (*N* = 377) vs. female (*N* = 405)	PAA A: White (*N* = 463) vs. Black (*N* = 58)	PAA B: White (*N* = 430) vs. Black (*N* = 100)	PAA B: White (*N* = 430) vs. Asian/Pacific Islander (*N* = 74)	PAA A: White (*N* = 463) vs. Hispanic (*N* = 65)	PAA B: White (*N* = 430) vs. Hispanic (*N* = 174)	PAA A: Low SES: No (*N* = 330) vs. yes (*N* = 215)	PAA B: Low SES: No (*N* = 439) vs. yes (*N* = 298)	Number of C DIF (if not 0)
C02	A	B-	A	A	A	A	A	A	A	
C03	A	A	A	A	A	A	A	A	A	
C04	A	B-	A	A	A	A	A	A	A	
C07	A	A	A	A	A	B+	A	A	A	
C11	A	A	A	A	A	A	A	A	A	
C12	A	A	A	A	B+	A	A	A	B-	
C14	B+	A	A	A	B+	A	A	A	A	
C18	B+	A	A	A	A	A	A	A	A	
C19	A	B+	A	A	A	A	C+	A	B+	1
C20	B+	B+	A	A	A	A	B-	A	A	
C21	C+	B+	B+	B+	A	A	A	A	A	1

Note. The first group is the reference group, and the second group is the focus group. A positive sign favors the focus group, while a negative sign favors the reference group. DIF = differential item functioning; PAA = periodic accountability assessments; SES = socioeconomic status.