Listening. Learning. Leading.®

Research Memorandum
ETS RM-13-01

Statistical Report of Fall 2009 *CBAL*™ Writing Tests

Jianbin Fu

Seunghee Chung

Maxwell Wise

February 2013

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Statistical Report of Fall 2009 *CBAL*™ Writing Tests

Jianbin Fu, Seunghee Chung, and Maxwell Wise
ETS, Princeton, New Jersey

February 2013

Action Editor: James Carlson

Reviewers: Lora Monfils and Rebecca Zwick

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 4 CBAL Grade 8 Writing tests administered to students in 12 states in fall 2009. Specifically, classical item statistics including rater reliabilities for human-scored items, item *p*+ values, item-total correlations, item missing response rates, differential item functioning (DIF), interscore correlations, and reliabilities of subscores and total scores are reported. Under item response theory, the tests are calibrated and scaled based on the generalized partial credit model. In addition, *t*-tests, multiple comparisons, and mixed models are used to examine the factors influencing test scores, including test form, test order, student, school, gender, and socioeconomic status. The results show that these 4 tests performed reasonably well.

Key words: CBAL, writing test, item analysis, item response theory, statistical report

Acknowledgments

Thanks to Randy Bennett, Jim Carlson, Rick Morgan, Lora Monfils, and Rebecca Zwick for their helpful suggestions and edits on early versions of this report. We are also grateful to Kim Fryer for her editorial assistance.

Table of Contents

Page

List of Tables

List of Figures

Page

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 after learning (of learning), provides timely feedback information 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 still a research project, and CBAL summative tests are not used as accountability assessments. CBAL tests are developed based on the underlying cognitive competency models that incorporate curriculum standards with results of learning sciences' research. The competency models describe skills that students need to learn and their interrelationships, for example, 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.

Four Grade 8 Writing PAAs were administered as described in the next section in the fall of 2009. This report presents the statistical results of the four Writing PAAs in that 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 missing response rates, and differential item functioning (DIF); (c) summary statistics of subscores and total raw scores including means, standard deviations, interscore correlations, and reliabilities; (d) the relationships among lead-in tasks and essays within and across PAAs; (e) results from concurrent calibration and separate calibration based on the generalized partial credit model; and (f) test performance by demographic groups based on gender, socioeconomic status, and race, as well as effects of PAA, test order, student, and school on test scores. Note that in another report, Fu, Wise, and Chung (2011) explored test dimensionality within each PAA.

Test and Sampling Designs

The fall 2009 field test included four PAAs focused on different writing genres: Service Learning, Invasive Plant Species, Ban Ads, and Mango Street. Each PAA had both dichotomous and polytomous items, and item types included constructed-response (CR), short CR, 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, SR/C&C items and subscores, respectively, and possible maximum total raw score for each PAA.)

Each PAA was based on a common scenario, and 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, which are necessary for writing a good essay on a specific genre, and the fourth task was writing an essay. The first three tasks comprised Test Section I and the fourth task was Test Section II. The PAAs were timed at the task level, and each section had to be finished in 50 minutes.

Tables 2 to 5 list the information for each item in the four PAAs, including item score ID, task, and subscore to which an item belongs, item sequence number, item type, scoring type (computer or human scored), score range after score weights were applied, and score weight. For the description of the test design from the content perspective, see Deane, Fowles, Baldwin, and Persky (2011) and Deane et al. (2009).

Table 1

CBAL Writing Test Design

PAA no.	PAA	Writing genre	Number of SR/C&C items	Number of CR/SCR items	Number of subscores	Max total score
1	Service Learning	Persuasive/applying criteria	22	3	4	60
2	Invasive Plant Species	Research-based expository writing	28	4	5	81
3	Ban Ads	Persuasive/argumentative writing	21	5	6	63
4	Mango Street	Writing about literature	10	4	4	41

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

Table 2

Service Learning (PAA 1): Item and Subscore Information

Task number and name	Item sequence	Item score ID	Type	Scoring type	Score range [a]	Score weight	S1	S2	S3	S4
1. Give feedback	1	W_SERVLEARN01	MC, C&C	A	0-1	1	1	-	-	-
	2	W_SERVLEARN02	MC, C&C	A	0-1	1	1	-	-	-
	3	W_SERVLEARN03	MC, C&C	A	0-1	1	1	-	-	-
	4	W_SERVLEARN04	MC, C&C	A	0-1	1	1	-	-	-
	5	W_SERVLEARN05	MC, C&C	A	0-1	1	1	-	-	-
	6	W_SERVLEARN06	MC, C&C	A	0-1	1	1	-	-	-
	7	W_SERVLEARN07	MC, C&C	A	0-1	1	1	-	-	-
2. Compare activities	8	W_SERVLEARN08H	MC, C&C	A	0-1	1	-	1	-	-
	9	W_SERVLEARN08G	MC, C&C	A	0-1	1	-	1	-	-
	10	W_SERVLEARN09H	MC, C&C	A	0-1	1	-	1	-	-
	11	W_SERVLEARN09G	MC, C&C	A	0-1	1	-	1	-	-
	12	W_SERVLEARN10H	MC, C&C	A	0-1	1	-	1	-	-
	13	W_SERVLEARN10G	MC, C&C	A	0-1	1	-	1	-	-
	14	W_SERVLEARN11H	MC, C&C	A	0-1	1	-	1	-	-
	15	W_SERVLEARN11G	MC, C&C	A	0-1	1	-	1	-	-
	16	W_SERVLEARN12H	MC, C&C	A	0-1	1	-	1	-	-
	17	W_SERVLEARN12G	MC, C&C	A	0-1	1	-	1	-	-
	18	W_SERVLEARN13H	MC, C&C	A	0-1	1	-	1	-	-
	19	W_SERVLEARN13G	MC, C&C	A	0-1	1	-	1	-	-
	20	W_SERVLEARN14H	MC, C&C	A	0-1	1	-	1	-	-
	21	W_SERVLEARN14G	MC, C&C	A	0-1	1	-	1	-	-
	22	W_SERVLEARN15	MC	A	0-1	1	-	1	-	-
3. Explain to a student	23	W_SERVLEARN16	CR	H	0-8	2	-	-	1	-

Task number and name	Item sequence	Item score ID	Type	Scoring type	Score range [a]	Score weight	S1	S2	S3	S4
4. Write an essay	24	W_SERVLEARN17_I	CR	H	0-15	3	-	-	-	1
	25	W_SERVLEARN17_III	CR	H	0-15	3	-	-	-	1
		Number of items [b]			24	-	7	14	1	2
		Max. possible score [b]			59	-	7	14	8	30

Note. A = automatically scored by computer, C&C = click & click, CR = constructed response, H = human-scored, PAA = periodic accountability assessment, S1 = subscore for give feedback, S2 = subscore for compare, S3 = subscore for short evaluation, S4 = subscore for essay, SR = selected response.

[a] Score range after score weights are applied. [b] Exclude W_SERVLEARN15 because of its zero item-total correlation; see Table 12.

Table 3

Invasive Plant Species (PAA 2): Item and Subscore Information

Task number and name	Item sequence	Item score ID	Type	Scoring type	Score range [a]	Score weight	S1	S2	S3	S4	S5
1. Gather and evaluate information	1	W_INVASIVE_01_01	CR	H	0-5	1	1	-	-	-	-
	2	W_INVASIVE_01_02	SR	A	0-1	1	-	1	-	-	-
	3	W_INVASIVE_01_03	SR	A	0-1	1	-	1	-	-	-
	4	W_INVASIVE_01_04	SR	A	0-1	1	-	1	-	-	-
	5	W_INVASIVE_01_05	SR	A	0-1	1	-	1	-	-	-
	6	W_INVASIVE_01_06	SR	A	0-1	1	-	1	-	-	-
	7	W_INVASIVE_01_07	SR	A	0-1	1	-	1	-	-	-
	8	W_INVASIVE_01_08	SR	A	0-1	1	-	1	-	-	-
	9	W_INVASIVE_01_09	SR	A	0-1	1	-	1	-	-	-
	10	W_INVASIVE_01_10	SR	A	0-1	1	-	1	-	-	-
	11	W_INVASIVE_01_11	SR	A	0-1	1	-	1	-	-	-

Task number and name	Item sequence	Item score ID	Type	Scoring type	Score range [a]	Score weight	S1	S2	S3	S4	S5
	12	W_INVASIVE_01_12	SR	A	0-1	1	-	1	-	-	-
	13	W_INVASIVE_01_13	SR	A	0-1	1	-	1	-	-	-
2. Organize information	14	W_INVASIVE_02_01	SR	A	0-1	1	-	-	1	-	-
	15	W_INVASIVE_02_02	SR	A	0-1	1	-	-	1	-	-
	16	W_INVASIVE_02_03	SR	A	0-1	1	-	-	1	-	-
	17	W_INVASIVE_02_04	SR	A	0-1	1	-	-	1	-	-
	18	W_INVASIVE_02_05	SR	A	0-1	1	-	-	1	-	-
	19	W_INVASIVE_02_06	SR	A	0-1	1	-	-	1	-	-
	20	W_INVASIVE_02_07	SR	A	0-1	1	-	-	1	-	-
	21	W_INVASIVE_02_08	SR	A	0-1	1	-	-	1	-	-
	22	W_INVASIVE_02_09	SR	A	0-1	1	-	-	1	-	-
	23	W_INVASIVE_02_10	SR	A	0-1	1	-	-	1	-	-
	24	W_INVASIVE_02_11	SR	A	0-1	1	-	-	1	-	-
	25	W_INVASIVE_02_12	SR	A	0-1	1	-	-	1	-	-
	26	W_INVASIVE_02_13	SR	A	0-1	1	-	-	1	-	-
	27	W_INVASIVE_02_14	SR	A	0-1	1	-	-	1	-	-
	28	W_INVASIVE_02_15	SR	A	0-1	1	-	-	1	-	-
	29	W_INVASIVE_02_16	SR	A	0-1	1	-	-	1	-	-
3. Revise	30	W_INVASIVE_03_01	CR	H	0-8	2	-	-	-	1	-
4. Write pamphlet section	31	W_INVASIVE_04_02_I	CR	H	0-20	4	-	-	-	-	1
	32	W_INVASIVE_04_02_III	CR	H	0-20	4	-	-	-	-	1
		Number of items [b]			31	-	1	11	16	1	2
		Max. possible score [b]			80	-	5	11	16	8	40

Note. A = automatically scored by computer, CR = constructed response, H = human-scored, S1 = subscore for guiding questions, S2 = subscore for evaluate sources, S3 = subscore for organize information, S4 = subscore for revision, S5 = subscore for write pamphlet, SR = selected response.

[a] Score range after score weights are applied. [b] Exclude W_INVASIVE_01_12 because of its negative item-total correlation; see Table 13.

Table 4

Ban Ads (PAA 3): Item and Subscore Information

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	W_BANADS_01A_01	SR	A	0-1	1	1	-	-	-	-	-
	2	W_BANADS_01A_02	SR	A	0-1	1	1	-	-	-	-	-
	3	W_BANADS_01A_03	SR	A	0-1	1	1	-	-	-	-	-
	4	W_BANADS_01A_04	SR	A	0-1	1	1	-	-	-	-	-
	5	W_BANADS_01A_05	SR	A	0-1	1	1	-	-	-	-	-
	6	W_BANADS_01B	CR	H	0-2	1	-	1	-	-	-	-
	7	W_BANADS_01C	CR	H	0-2	1	-	1	-	-	-	-
2. Analyze arguments	8	W_BANADS_02AX_A	SR	A	0-1	1	-	-	1	-	-	-
	9	W_BANADS_02AX_B	SR	A	0-1	1	-	-	1	-	-	-
	10	W_BANADS_02AX_C	SR	A	0-1	1	-	-	1	-	-	-
	11	W_BANADS_02AX_D	SR	A	0-1	1	-	-	1	-	-	-
	12	W_BANADS_02AX_E	SR	A	0-1	1	-	-	1	-	-	-
	13	W_BANADS_02AX_F	SR	A	0-1	1	-	-	1	-	-	-
	14	W_BANADS_02AX_G	SR	A	0-1	1	-	-	1	-	-	-
	15	W_BANADS_02AX_H	SR	A	0-1	1	-	-	1	-	-	-
	16	W_BANADS_02AX_I	SR	A	0-1	1	-	-	1	-	-	-
	17	W_BANADS_02AX_J	SR	A	0-1	1	-	-	1	-	-	-
	18	W_BANADS_02BX_A	SR	A	0-1	1	-	-	-	1	-	-
	19	W_BANADS_02BX_B	SR	A	0-1	1	-	-	-	1	-	-
	20	W_BANADS_02BX_C	SR	A	0-1	1	-	-	-	1	-	-
	21	W_BANADS_02BX_D	SR	A	0-1	1	-	-	-	1	-	-
	22	W_BANADS_02BX_E	SR	A	0-1	1	-	-	-	1	-	-
	23	W_BANADS_02BX_F	SR	A	0-1	1	-	-	-	1	-	-
3. Critique an argument	24	W_BANADS_03	CR	H	0-8	2	-	-	-	-	1	-

Task number and name	Item sequence	Item score ID	Type	Scoring type	Score range [a]	Score weight	S1	S2	S3	S4	S5	S6
4. Write an essay	25	W_BANADS_04_I	CR	H	0-15	3	-	-	-	-	-	1
	26	W_BANADS_04_III	CR	H	0-15	3	-	-	-	-	-	1
Number of items [b]					25	-	4	2	10	6	1	2
Max. possible score [b]					62	-	4	4	10	6	8	30

Note. A = automatically scored by computer, CR = constructed response, H = human-scored, PAA = periodic accountability assessment, S1 = subscore for summary feedback, S2 = subscore for CR summary, S3 = subscore for claims, S4 = subscore for evidence, S5 = subscore for critique, S6 = subscore for essay, SR = selected response.

[a] Score range after score weights are applied. [b] Exclude W_BANADS_01A_01 because of its negative item-total correlation; see Table 14.

Table 5

Mango Street (PAA 4): Item and Subscore Information

Task number and name	Item sequence	Item score ID	Type	Scoring type	Score range [a]	Score weight	S1	S2	S3	S4
1. Support interpretations of the story	1	W_MANGO_01_01	C&C	A	0-1	.5	1	-	-	-
	2	W_MANGO_01_02	C&C	A	0-1	.5	1	-	-	-
	3	W_MANGO_01_03	C&C	A	0-1	.5	1	-	-	-
	4	W_MANGO_01_04	C&C	A	0-1	.5	1	-	-	-
	5	W_MANGO_01_05	C&C	A	0-1	.5	1	-	-	-

Task number and name	Item sequence	Item score ID	Type	Scoring type	Score range [a]	Score weight	S1	S2	S3	S4
2. Explain whether a character’s attitude changes	6	W_MANGO_02_01	CR	H	0-8	2	-	1	-	-
3. Help another student interpret the text	7	W_MANGO_03_01	SR	A	0-1	1	-	-	1	-
	8	W_MANGO_03_02	SR	A	0-1	1	-	-	1	-
	9	W_MANGO_03_03	SR	A	0-1	1	-	-	1	-
	10	W_MANGO_03_04	SR	A	0-1	1	-	-	1	-
	11	W_MANGO_03_05	SR	A	0-1	1	-	-	1	-
	12	W_MANGO_03_06	SCR	H	0-3	1	-	-	1	-
4. Write an essay	13	W_MANGO_04_I	CR	H	0-10	2	-	-	-	1
	14	W_MANGO_04_III	CR	H	0-10	2	-	-	-	1
		Number of items			14	-	5	1	6	2
		Max. possible score			41	-	5	8	8	20

Note. A = automatically scored by computer, CR = constructed response, H = human-scored, PAA = periodic accountability assessment, S1 = subscore for support interpretation, S2 = subscore for interpretive discussion, S3 = subscore for choose interpretation, S4 = subscore for essay, SR = selected response.

[a] Score range after score weights are applied.

The CBAL Writing PAAs were administered online to a convenience sample of 2,580 Grade 8 students from 21 schools in 12 states. (See Table 6 for the sample distribution by state, gender, socioeconomic status [SES] with low SES students defined as having reduced or free lunch, English language learner [ELL] status, and race.) The students took two PAAs out of the four in 1 of the 12 possible orders (see Table 7). For security reasons, a first PAA could not be used as a second PAA in the same school. To accommodate this restriction, the schools were randomly separated into six groups, each associated with four test sequences (see Table 8), and the students in a school group were randomly assigned to one of the four sequences. Ninety-three percent of students completed both PAAs within 1 month.

Classical Item Analyses

Rater Agreement for Human-Scored Items

According to Tables 2 to 5, each PAA had three to six human-scored items with a total of 16 for the four PAAs. Each of the 16 items was scored by two raters. The pairs of raters were not the same across items and students. A third rater would score an item if the difference between the first two raters' scores (before score weights were applied) was larger than one point (except for W_INVASIVE_01_01 which was a sum of the scores of five human-scored dichotomous responses). The raters were familiar with the CBAL writing tests. The adjudication rules for human-scored items were as follows:

1. If there are only two rating scores (the two-rater score difference is 0 or 1), average the two scores.
2. If there are three rating scores (hence the difference between the first two rater's scores is larger than 1), and the third score is closer to one score than the other, average Rater 3's score with the nearest score and discarded the other.
3. If there are three rating scores and the third score is at the middle point of the first two scores, average Rater 3's score with the other two.

In this section, the first two raters' scores were used to assess rater agreement. Students receiving any omit or not-reached rater score on a human-scored item were excluded from the analysis on this item.

Table 6

Test Sample Distribution by Demographic Characteristic

Demographic	*N*	%
State	-	-
Alabama	122	4.73
Arizona	573	22.21
Arkansas	290	11.24
California	64	2.48
Florida	41	1.59
Georgia	201	7.79
Kentucky	61	2.36
Louisiana	110	4.26
Massachusetts	106	4.11
Mississippi	99	3.84
Ohio	192	7.44
Texas	204	7.91
Unreported	517	20.04
Gender	-	-
Male	1,051	40.74
Female	1,010	39.15
Unreported	519	20.12
Low SES status	-	-
No	701	27.17
Yes	705	27.33
Unreported	1,174	45.50
ELL status	-	-
No	1,077	41.74
Yes	52	2.02
Unreported	1,451	56.24
Race	-	-
African American	374	14.50
Asian/Pacific Islander	58	2.25
Hispanic	196	7.60
Native American	10	.39
White	1,032	40.00
Unreported	910	35.27

Note. Many participant schools failed to fill in the background questionnaire; thus, a lot of demographic information was missing. ELL = English language learner, SES = socioeconomic status.

Table 7

Writing Periodic Accountability Assessment (PAA) Sequences

First PAA	Second PAA			
	PAA1	PAA2	PAA3	PAA4
Service Learning (PAA1)	-	1	2	3
Invasive Plant Species (PAA2)	4	-	5	6
Ban Ads (PAA3)	7	8	-	9
Mango Street (PAA4)	10	11	12	-

Table 8

School Group and Assigned Periodic Accountability Assessment (PAA) Sequences

School group	PAA sequences
Group 1	2, 3, 5, 6
Group 2	1, 3, 8, 9
Group 3	1, 2, 11, 12
Group 4	4, 6, 7, 9
Group 5	4, 5, 10, 12
Group 6	7, 8, 10, 11

Kappa coefficients and percentage agreement. Table 9 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 (ASE) estimate of each weighted kappa coefficient, and the percentage of 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 from 1 (*same scores*) to 0 (*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. 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 to .40

Moderate agreement = .40 to .60

Good agreement = .60 to .80

Very good agreement = .80 to 1.00.

Therefore, all the human-scored items showed good to very good agreement between the first two raters. The percentages of rater agreement ranged from 32% to 78%. Note that Item W_INVASIVE_01_01 was a sum of the scores of five human-scored dichotomous responses. Because rater score differences were cumulated this item had the lowest rater agreement of 32%.

Table 9

Weighted Kappa Coefficient and Percentage of Agreement

Human-scored item	Number of score categories	Sample size	Weighted kappa [a]	ASE of kappa	% of agreement
W_SERVLEARN16	5	1,187	.78	.01	61.92
W_SERVLEARN17_I	6	1,107	.79	.01	53.75
W_SERVLEARN17_III	6	1,104	.79	.01	58.79
W_INVASIVE_01_01	11	1,201	.79	.01	32.47
W_INVASIVE_03_01	5	1,202	.89	.01	67.55
W_INVASIVE_04_02_I	6	1,097	.63	.02	48.04
W_INVASIVE_04_02_III	6	1,100	.62	.02	50.45
W_BANADS_01B	3	1,155	.72	.02	76.36
W_BANADS_01C	3	1,096	.78	.01	78.47
W_BANADS_03	5	1,153	.84	.01	69.12
W_BANADS_04_I	6	1,047	.77	.01	55.01
W_BANADS_04_III	6	1,052	.85	.01	67.87
W_MANGO_02_01	5	1,209	.73	.01	59.47
W_MANGO_03_06	4	1,207	.73	.01	59.32
W_MANGO_04_I	6	1,109	.78	.01	54.46
W_MANGO_04_III	6	1,109	.83	.01	68.80

Note. ASE = asymptotic standard error.

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

Generalizability coefficients. Generalizability theory (Brennan, 2001; Shavelson & Webb, 1991) was used to estimate the rater reliabilities. Treating this as a G-study design, we had a balanced design with one facet (rater) and the object of measurement (students) where rater was nested within student. Based on the variance components from the G-studies, we estimated the generalizability coefficients for the following two D-studies (Crocker & Algina, 1986, pp. 157–171):

1. Each student was rated by one rater, and each student had a different rater;
2. Each student was rated by two raters, each student had different raters, and the final item score was the average of the two rater scores.

Note that the G-study and D-study used the same students. The data in the G-study was used to calculate the rater reliability for the scenario described in each D-study; no separate D-study was conducted. For the D-study where each student has different raters for a given item, the total number of raters is two times the number of students. Table 10 shows the generalizability coefficient estimates. The estimates for the one-rater model ranged from .63 to .89. One can see that compared to the one-rater design, averaging two raters' scores increased the generalizability coefficient estimates from .05 to .15 across all the human-scored items and doubled the accuracy of estimates, as indicated by the signal/noise ratios in Table 10.

Item Summary Statistics

Tables A1 to A4 in Appendix A list the item score frequencies including the frequencies for omit and not reached responses as well as system errors (i.e., the online testing system failed to capture a student's response) for the four PAAs, respectively. Tables 12 to 15 contain item summary statistics for the four PAAs, 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, percentage omit, percentage not reached, percentage system error, and percentage not responding (sum of percentages omit, not reached, and system error), as well as item flags, which as defined in Table 11, single out items with extreme item statistics to be reviewed. At the bottom of Tables 12 to 15, summary statistics across items including mean, standard deviation, minimum, and maximum are also provided. Note that unless explicitly specified, omit was treated as zero across the analyses in this study, while not reached and system error were treated as missing, and a composite score including any missing item score was designated as missing.

Tables 12 to 15 show that the not responding rates were small (less than 5%), except for most items in Task 1 in Invasive Plant Species and W_BANADS_01C. In Task 1 in Invasive Plant Species, starting from the third item, W_INVASIVE_01_03, the not reached rate was 5.36% and increased to 16.42% for the final item, W_INVASIVE_01_13. This was due to speededness: In Task 1 in Invasive Plant Species, there were 15 minutes for 13 items, plus four directions screens and two different stimuli, one of which had four tabs of information.

Table 10

Generalizability Coefficients for Item Rater D-Studies

Item	Number of score categories	*N*	Each student was rated by one rater; each student had different raters		Each student was rated by two raters; each student had different raters	
			Gen. coef.	Signal/noise ratio	Gen. coef.	Signal/noise ratio
W_SERVLEARN16	5	1,187	.78	3.46	.87	6.92
W_SERVLEARN17_I	6	1,107	.79	3.68	.88	7.36
W_SERVLEARN17_III	6	1,104	.79	3.80	.88	7.59
W_INVASIVE_01_01	11	1,201	.78	3.63	.88	7.25
W_INVASIVE_03_01	5	1,202	.89	7.71	.94	15.43
W_INVASIVE_04_02_I	6	1,097	.63	1.72	.77	3.43
W_INVASIVE_04_02_III	6	1,100	.62	1.63	.77	3.26
W_BANADS_01B	3	1,155	.72	2.57	.84	5.14
W_BANADS_01C	3	1,096	.78	3.49	.87	6.99
W_BANADS_03	5	1,153	.84	5.33	.91	10.66
W_BANADS_04_I	6	1,047	.77	3.32	.87	6.65
W_BANADS_04_III	6	1,052	.85	5.48	.92	10.96
W_MANGO_02_01	5	1,209	.73	2.76	.85	5.53
W_MANGO_03_06	4	1,207	.73	2.72	.84	5.45
W_MANGO_04_I	6	1,109	.78	3.47	.87	6.95
W_MANGO_04_III	6	1,109	.83	4.95	.91	9.91

Note. Gen. coef = generalizability coefficient.

The correlation between an item score and the total score is used to indicate the association strength between an item and the construct (represented by total score) that it measures; this 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. The polyserial correlation assumes that the ordinal variable has an underlying standard normal distribution, and the two variables follow a bivariate normal distribution. Tables 12 to 15 provide both polyserial and Pearson item-total correlations because some polyserials did not converge, and for one item (W_INVASIVE_01_01), the polyserial did not exist as it had 21 score categories and was treated as a continuous variable by the LISREL program used to compute the polyserials. One can see that all polyserials were higher in absolute value than their Pearson correlation counterparts. All item-total correlations look reasonable except for three items: one in each of PAAs 1 to 3, W_SERVLEARN15, W_INVASIVE_01_12, and W_BANADS_01A_01, which had polyserial correlations of .00, -.25, and -.27, respectively, and thus were excluded from all the subsequent analyses and reports of summary item statistics. The mean item-total polyserial correlations for the four PAAs were .53, .49, .46, and .65, respectively.

Table 11

Item Flag Definition

Flag value	Reasons for flagging	Criterion	
		Dichotomous	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 not responding	Percentage of not responding > 5%	

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 PAAs 2 to 4 were between .19 and .87 with averages of .57, .58, and .54, respectively. PAA 1 was more difficult than PAAs 2 to 4 as its item *p*+ values were between .17 and .65 with an average of .37.

Table 12

Service Learning (PAA 1): Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,055)	Pearson correlation (*N* = 1,055)	% omit	% not reached	% system error	% not respond	Flag
W_SERVLEARN01	1,193	.34	.47	1	.34	.59	.45	.00	.00	.00	.00	-
W_SERVLEARN02	1,190	.17	.38	1	.17	.33	.22	.00	.25	.00	.25	A
W_SERVLEARN03	1,186	.19	.39	1	.19	.29	.20	.00	.59	.00	.59	A R
W_SERVLEARN04	1,184	.42	.49	1	.42	.43	.34	.00	.75	.00	.75	-
W_SERVLEARN05	1,174	.23	.42	1	.23	.58	.42	.00	1.59	.00	1.59	A
W_SERVLEARN06	1,156	.21	.41	1	.21	.46	.32	.00	3.10	.00	3.10	A
W_SERVLEARN07	1,137	.44	.50	1	.44	.63	.50	.00	4.69	.00	4.69	-
W_SERVLEARN08H	1,192	.29	.45	1	.29	.41	.31	.00	.08	.00	.08	-
W_SERVLEARN08G	1,192	.49	.50	1	.49	.77	.61	.00	.08	.00	.08	-
W_SERVLEARN09H	1,191	.18	.38	1	.18	.13	.09	.00	.17	.00	.17	A R
W_SERVLEARN09G	1,191	.48	.50	1	.48	.57	.45	.00	.17	.00	.17	-
W_SERVLEARN10H	1,190	.20	.40	1	.20	.13	.09	.00	.25	.00	.25	A R
W_SERVLEARN10G	1,190	.24	.43	1	.24	.37	.27	.00	.25	.00	.25	A
W_SERVLEARN11H	1,189	.28	.45	1	.28	.14	.11	.00	.34	.00	.34	R
W_SERVLEARN11G	1,189	.54	.50	1	.54	.74	.59	.00	.34	.00	.34	-
W_SERVLEARN12H	1,186	.35	.48	1	.35	.52	.41	.00	.50	.08	.59	-
W_SERVLEARN12G	1,186	.24	.43	1	.24	.57	.42	.00	.50	.08	.59	A
W_SERVLEARN13H	1,187	.57	.49	1	.57	.65	.53	.00	.50	.00	.50	-
W_SERVLEARN13G	1,187	.64	.48	1	.64	.66	.52	.00	.50	.00	.50	-
W_SERVLEARN14H	1,187	.48	.50	1	.48	[a]	.56	.00	.50	.00	.50	-
W_SERVLEARN14G	1,187	.65	.48	1	.65	.59	.46	.00	.50	.00	.50	-
W_SERVLEARN15	1,185	.14	.34	1	.14	.00	.00	.17	.67	.00	.84	A R
W_SERVLEARN16	1,191	3.40	1.91	8	.43	.78	.77	.17	.17	.00	.34	-
W_SERVLEARN17_I	1,115	6.91	3.48	15	.46	.89	.89	.98	.18	.00	1.16	-
W_SERVLEARN17_III	1,115	6.09	3.19	15	.41	.88	.87	.98	.18	.00	1.16	-

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,055)	Pearson correlation (*N* = 1,055)	% omit	% not reached	% system error	% not respond	Flag
Mean [b]	-	1.00	.75	2.46	.37	.53	.43	.09	.67	.01	.77	-
SD [b]	-	1.77	.83	4.03	.15	.22	.21	.27	1.05	.02	1.04	-
Min [b]	-	.17	.38	1.00	.17	.13	.09	.00	.00	.00	.00	-
Max [b]	-	6.91	3.48	15.00	.65	.89	.89	.98	4.69	.08	4.69	-

Note. See Table 11 for definition of flags. PAA = periodic accountability assessment.

[a] Item-total polyserial correlation did not converge. [b] Excluded W_SERVLEARN15.

Table 13

Invasive Plant Species (PAA 2): Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 911)	Pearson correlation (*N* = 911)	% omit	% not reached	% system error	% not respond	Flag
W_INVASIVE_01_01	1,212	2.42	1.41	5	.48	[a]	.59	.83	.00	.00	.83	R
W_INVASIVE_01_02	1,173	.56	.50	1	.56	.52	.41	1.65	3.22	.00	4.87	-
W_INVASIVE_01_03	1,146	.45	.50	1	.45	.25	.20	.91	5.36	.08	6.35	R N P
W_INVASIVE_01_04	1,136	.84	.36	1	.84	.48	.29	.41	6.27	.00	6.68	N P
W_INVASIVE_01_05	1,130	.42	.49	1	.42	.14	.11	.91	6.77	.00	7.67	R N P
W_INVASIVE_01_06	1,116	.64	.48	1	.64	.16	.13	.33	7.92	.00	8.25	R N P
W_INVASIVE_01_07	1,110	.86	.34	1	.86	.41	.25	.83	8.42	.00	9.24	N P
W_INVASIVE_01_08	1,100	.82	.38	1	.82	.36	.23	1.49	9.24	.00	10.73	N P
W_INVASIVE_01_09	1,081	.55	.50	1	.55	.40	.31	.83	10.81	.00	11.63	N P
W_INVASIVE_01_10	1,067	.69	.46	1	.69	.21	.16	1.57	11.96	.00	13.53	R N P
W_INVASIVE_01_11	1,048	.80	.40	1	.80	.30	.21	.99	13.53	.00	14.52	N P
W_INVASIVE_01_12	1,033	.29	.45	1	.29	-.25	-.18	1.57	14.69	.08	16.34	R N P
W_INVASIVE_01_13	1,013	.51	.50	1	.51	.25	.20	1.65	16.42	.00	18.07	R N P
W_INVASIVE_02_01	1,208	.64	.48	1	.64	.77	.58	.00	.25	.08	.33	-
W_INVASIVE_02_02	1,209	.58	.49	1	.58	.51	.40	.08	.25	.00	.33	-

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 911)	Pearson correlation (*N* = 911)	% omit	% not reached	% system error	% not respond	Flag
W_INVASIVE_02_03	1,207	.39	.49	1	.39	.36	.28	.00	.41	.00	.41	-
W_INVASIVE_02_04	1,207	.56	.50	1	.56	.38	.30	.00	.41	.00	.41	-
W_INVASIVE_02_05	1,207	.19	.39	1	.19	.10	.07	.08	.41	.00	.50	A R
W_INVASIVE_02_06	1,206	.60	.49	1	.60	.61	.47	.00	.50	.00	.50	-
W_INVASIVE_02_07	1,206	.63	.48	1	.63	[b]	.58	.08	.50	.00	.58	-
W_INVASIVE_02_08	1,205	.69	.46	1	.69	.75	.55	.00	.58	.00	.58	-
W_INVASIVE_02_09	1,204	.55	.50	1	.55	.62	.48	.00	.58	.08	.66	-
W_INVASIVE_02_10	1,203	.58	.49	1	.58	.63	.50	.00	.58	.17	.74	-
W_INVASIVE_02_11	1,204	.67	.47	1	.67	.60	.45	.00	.58	.08	.66	-
W_INVASIVE_02_12	1,205	.36	.48	1	.36	.20	.15	.00	.58	.00	.58	R
W_INVASIVE_02_13	1,204	.60	.49	1	.60	.78	.60	.00	.58	.08	.66	-
W_INVASIVE_02_14	1,204	.68	.47	1	.68	.69	.52	.00	.58	.08	.66	-
W_INVASIVE_02_15	1,205	.56	.50	1	.56	.68	.54	.00	.58	.00	.58	-
W_INVASIVE_02_16	1,205	.67	.47	1	.67	.71	.53	.00	.58	.00	.58	-
W_INVASIVE_03_01	1,207	2.95	2.75	8	.37	.70	.68	.33	.41	.00	.74	-
W_INVASIVE_04_02_I	1,102	6.78	4.74	20	.34	.83	.82	.00	.09	.00	.09	-
W_INVASIVE_04_02_I	1,102	5.03	3.67	20	.25	.83	.81	.00	.09	.00	.09	A
Mean [c]	-	1.07	.81	2.58	.57	.49	.40	.42	3.50	.02	3.94	-
SD [c]	-	1.39	1.00	4.78	.16	.22	.20	.56	4.69	.04	5.14	-
Min [c]	-	.19	.34	1.00	.19	.10	.07	.00	.00	.00	.09	-
Max [c]	-	6.78	4.74	20.00	.86	.83	.82	1.65	16.42	.17	18.07	-

Note. See Table 11 for definition of flags.

[a] This item score had 21 score categories and was treated as a continuous variable. [b] Item-total polyserial correlation did not converge.

[c] Excluded W_INVASIVE_01_12.

Table 14

Ban Ads (PAA 3): Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,025)	Pearson correlation (*N* = 1,025)	% omit	% not reached	% system error	% not respond	Flag
W_BANADS_01A_01	1,160	.62	.49	1	.62	-.27	-.21	.00	.00	.00	.00	R
W_BANADS_01A_02	1,159	.77	.42	1	.77	.39	.27	.00	.00	.09	.09	-
W_BANADS_01A_03	1,160	.58	.49	1	.58	.38	.30	.00	.00	.00	.00	-
W_BANADS_01A_04	1,160	.40	.49	1	.40	.14	.11	.00	.00	.00	.00	R
W_BANADS_01A_05	1,160	.78	.41	1	.78	.40	.28	.09	.00	.00	.09	-
W_BANADS_01B	1,159	.53	.62	2	.27	.70	.65	.34	.09	.00	.43	A
W_BANADS_01C	1,124	.76	.68	2	.38	.63	.60	2.33	3.10	.00	5.43	P
W_BANADS_02AX_A	1,158	.75	.43	1	.75	.46	.33	.00	.17	.00	.17	-
W_BANADS_02AX_B	1,158	.79	.41	1	.79	.42	.28	.00	.17	.00	.17	-
W_BANADS_02AX_C	1,158	.63	.48	1	.63	.22	.17	.00	.17	.00	.17	R
W_BANADS_02AX_D	1,158	.81	.40	1	.81	.56	.38	.00	.17	.00	.17	-
W_BANADS_02AX_E	1,158	.79	.41	1	.79	.54	.37	.00	.17	.00	.17	-
W_BANADS_02AX_F	1,158	.83	.38	1	.83	.43	.28	.00	.17	.00	.17	-
W_BANADS_02AX_G	1,158	.72	.45	1	.72	.35	.26	.00	.17	.00	.17	-
W_BANADS_02AX_H	1,158	.87	.33	1	.87	.53	.31	.00	.17	.00	.17	-
W_BANADS_02AX_I	1,158	.75	.43	1	.75	.52	.37	.00	.17	.00	.17	-
W_BANADS_02AX_J	1,158	.62	.49	1	.62	.44	.34	.00	.17	.00	.17	-
W_BANADS_02BX_A	1,157	.60	.49	1	.60	.15	.11	.17	.26	.00	.43	R
W_BANADS_02BX_B	1,157	.41	.49	1	.41	.26	.20	.26	.26	.00	.52	R
W_BANADS_02BX_C	1,157	.43	.50	1	.43	.37	.30	.26	.26	.00	.52	-
W_BANADS_02BX_D	1,157	.63	.48	1	.63	.39	.31	.26	.26	.00	.52	-
W_BANADS_02BX_E	1,157	.28	.45	1	.28	.22	.17	.34	.26	.00	.60	R
W_BANADS_02BX_F	1,157	.49	.50	1	.49	.43	.34	.34	.26	.00	.60	-
W_BANADS_03	1,155	1.93	2.09	8	.24	.79	.75	.00	.43	.00	.43	A
W_BANADS_04_I	1,056	6.08	3.44	15	.41	.92	.91	.19	.28	.00	.47	-

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,025)	Pearson correlation (*N* = 1,025)	% omit	% not reached	% system error	% not respond	Flag
W_BANADS_04_III	1,056	5.76	3.30	15	.38	.90	.89	.19	.28	.00	.47	-
Mean [a]	-	1.12	.76	2.48	.58	.46	.37	.19	.30	.00	.49	-
SD [a]	-	1.45	.84	3.94	.20	.20	.21	.46	.58	.02	1.03	-
Min [a]	-	.28	.33	1.00	.24	.14	.11	.00	.00	.00	.00	-
Max [a]	-	6.08	3.44	15.00	.87	.92	.91	2.33	3.10	.09	5.43	-

Note. See Table 11 for definition of flags.

[a] Excluded W_BANADS_01A_01.

Table 15

Mango Street (PAA 4): Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,067)	Pearson correlation (*N* = 1,067)	% omit	% not reached	% system error	% not respond	Flag
W_MANGO_01_01	1,213	.34	.46	1	.34	.56	.44	.00	.00	.00	.00	R
W_MANGO_01_02	1,205	.59	.47	1	.59	.67	.56	.00	.66	.00	.66	-
W_MANGO_01_03	1,195	.65	.47	1	.65	.39	.31	.00	1.48	.00	1.48	R
W_MANGO_01_04	1,178	.70	.46	1	.70	.67	.51	.00	2.89	.00	2.89	-
W_MANGO_01_05	1,165	.42	.48	1	.42	.70	.57	.00	3.96	.00	3.96	-
W_MANGO_02_01	1,211	3.63	1.79	8	.45	.79	.77	.00	.16	.00	.16	-
W_MANGO_03_01	1,210	.67	.47	1	.67	.50	.38	.00	.25	.00	.25	-
W_MANGO_03_02	1,209	.70	.46	1	.70	.73	.55	.00	.33	.00	.33	-
W_MANGO_03_03	1,209	.55	.50	1	.55	.56	.44	.00	.33	.00	.33	-
W_MANGO_03_04	1,209	.70	.46	1	.70	.64	.48	.00	.33	.00	.33	-

Item score ID	*N*	Mean	SD	Max possible score	*p*+	Polyserial (*N* = 1,067)	Pearson correlation (*N* = 1,067)	% omit	% not reached	% system error	% not respond	Flag
W_MANGO_03_05	1,209	.56	.50	1	.56	.48	.37	.08	.33	.00	.41	-
W_MANGO_03_06	1,207	1.38	.93	3	.46	.70	.67	.00	.49	.00	.49	-
W_MANGO_04_I	1,113	4.30	2.26	10	.43	.89	.88	.36	.27	.00	.63	-
W_MANGO_04_III	1,113	3.79	1.98	10	.38	.88	.86	.36	.27	.00	.63	-
Mean	-	1.36	.84	2.93	.54	.65	.56	.06	.84	.00	.90	-
SD	-	1.41	.65	3.54	.13	.15	.18	.13	1.17	.00	1.14	-
Min	-	.34	.46	1.00	.34	.39	.31	.00	.00	.00	.00	-
Max	-	4.30	2.26	10.00	.70	.89	.88	.36	3.96	.00	3.96	-

Note. See Table 11 for definition of flags.

Differential Item Functioning (DIF)

Test fairness requires that all test items be fair to all students. DIF analysis is designed to identify items that may have biases against certain groups of students. That is, if students having the same ability but from different demographic groups perform differently on an item, then this item shows DIF. A DIF 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; Zwick, Donoghue, & Grima, 1993) was used to detect DIF. ETS DIF procedures (Dorans & Holland, 1993) result in classification of items into three DIF 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 in the following demographic group pairs:

1. gender (male vs. female),
2. race/ethnicity (White vs. Black; White vs. combination of Native American, Asian/Pacific Islander, and Hispanic), and
3. low SES students (no vs. yes).

The reason for combining Native American, Asian/Pacific Islander, and Hispanic is that these ethnic groups had sample sizes too small to conduct DIF analyses separately. Table 16 lists the Category C DIF items, and the tables in Appendix B show the DIF category for every item. Only one item in each of PAA 1 and PAA 2 has Category C DIF. Note that the Black and race/ethnicity combination had small sample sizes, fewer than 200. Therefore, their DIF results should be interpreted with caution.

Table 16

Category C Differential Item Functioning (DIF) Items

Item score ID	C DIF description
W_SERVLEARN16	Favor female over male
W_BANADS_01C	Favor White over Black

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, and explore the relationships between lead-in tasks (i.e., Tasks 1 to 3) and essays.

Subscores and Total Scores

Tables 17 to 20 show the statistics for the subscores and total raw scores of the four PAAs. These tests were relatively difficult as their mean total scores were 42% to 47% of the maximum possible scores. The subscores had 1 to 16 items (see Tables 2 to 5) and reliabilities ranging from .24 to .92. 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 .18 and .64. The correlations between subscores and total scores ranged from .43 to .93.

Table 21 shows that the correlations among the four PAA total scores were between .66 and .76. Table 21 also displays comparisons of the standardized alphas based on item scores and task scores. The alphas based on item scores ranged from .79 to .86, and the alphas based on task scores (commonly known as testlet reliability) were close to those based on item scores with differences of between .01 and .06, which indicates that testlet effects at the task level were minor for these four PAAs. For comparison purposes, the alphas based on item and task raw scores for the four PAAs are shown in Appendix C.

Table 17

Service Learning (PAA 1): Test Subscore and Total Score Summary and Correlations

Score	*N*	Mean	SD	Standardized alpha[a]	Pearson correlation			
					S1	S2	S3	S4
S1	1,137	2.02	1.72	.62	-	-	-	-
S2	1,186	5.63	3.19	.74	.58	-	-	-
S3	1,191	3.40	1.91		.40	.55	-	-
S4	1,115	13.00	6.38	.91	.42	.57	.64	-
Total	1,057	24.51	11.03	.85	.64	.81	.77	.92

Note. PAA = periodic accountability assessment, S1 = subscore for give feedback, S2 = subscore for compare, S3 = subscore for short evaluation, S4 = subscore for essay.

[a] Reliability was not calculated for a subscore with one item.

Table 18

Invasive Plant Species (PAA 2): Test Subscore and Total Score Summary and Correlations

Score	*N*	Mean	SD	Standardized alpha[a]	Pearson correlation				
					S1	S2	S3	S4	S5
S1	1,212	2.42	1.41	-	-	-	-	-	-
S2	1,012	7.20	1.85	.41	.36	-	-	-	-
S3	1,200	8.97	4.22	.84	.52	.50	-	-	-
S4	1,207	2.95	2.75	-	.45	.39	.55	-	-
S5	1,102	11.81	7.89	.86	.37	.35	.49	.42	-
Total	912	33.87	14.17	.86	.59	.59	.80	.68	.87

Note. PAA = periodic accountability assessment, S1 = subscore for guiding questions, S2 = subscore for evaluate sources, S3 = subscore for organize information, S4 = subscore for revision, S5 = subscore for write pamphlet,

[a] Reliability was not calculated for a subscore with one item.

Table 19

Ban Ads (PAA 3): Test Subscore and Total Score Summary and Correlations

Score	*N*	Mean	SD	Standardized alpha[a]	Pearson correlation					
					S1	S2	S3	S4	S5	S6
S1	1,159	2.54	1.00	.24						
S2	1,124	1.29	1.13	.68	.30					
S3	1,158	7.56	2.07	.66	.25	.40				
S4	1,157	2.84	1.41	.35	.18	.34	.26			
S5	1,155	1.93	2.09		.27	.56	.38	.34		
S6	1,056	11.84	6.49	.92	.30	.59	.44	.34	.60	
Total	1,025	28.08	10.93	.79	.43	.72	.62	.50	.76	.93

Note. PAA = periodic accountability assessment, S1 = subscore for summary feedback, S2 = subscore for CR summary, S3 = subscore for claims, S4 = subscore for evidence, S5 = subscore for critique, S6 = subscore for essay.

[a] Reliability was not calculated for a subscore with one item.

Table 20

Mango Street (PAA 4): Test Subscore and Total Score Summary and Correlations

Score	*N*	Mean	SD	Standardized alpha[a]	Pearson correlation			
					S1	S2	S3	S4
S1	1,165	2.71	1.51	.64				
S2	1,211	3.63	1.79		.48			
S3	1,207	4.57	2.06	.66	.60	.53		
S4	1,113	8.09	4.01	.89	.54	.61	.60	
Total	1,067	19.20	7.89	.85	.74	.77	.81	.91

Note. PAA = periodic accountability assessment, S1 = subscore for support interpretation, S2 = subscore for interpretive discussion, S3 = subscore for choose interpretation, S4 = subscore for essay.

[a] Reliability was not calculated for a subscore with one item.

Table 21

Total Score Summary and Correlations

Total raw score	Standardized alpha		Pearson correlation (*N*)		
	Task	Item	PAA 1	PAA 2	PAA 3
Service Learning (PAA 1)	.82	.85			
Invasive Plant Species (PAA 2)	.80	.86	.66 (271)		
Ban Ads (PAA 3)	.81	.79	.75 (375)	.66 (200)	
Mango Street (PAA 4)	.84	.85	.71 (246)	.76 (326)	.74 (286)

Note. PAA = periodic accountability assessment.

Table 22 shows the correlations of the four PAAs with some Grade 7 state tests by state on English language arts (ELA), math, reading, and writing. The numbers of available state tests on ELA, math, reading, and writing were 8, 10, 6, and 4. (Please note the limited sample sizes used in calculating these correlations: Most correlations were based on sample sizes smaller than 100.) The mean correlations between the four PAAs and the state tests were between .46 and .66, which provided some supportive evidence for the validity of CBAL Writing tests. The mean correlations with the state ELA and reading tests were slightly higher than those with the math state tests; however, the mean correlations with the state writing tests appeared to be slightly lower than those with the state math tests, which indicated some differences between the CBAL writing tests and the state writing tests. One such difference is that each CBAL writing test included three lead-in tasks to measure reading abilities related to essay writing. A second difference is that each CBAL Writing test includes extensive source materials that students must read in preparation for writing their culminating essay. The high reading demand made by CBAL writing tests is evident in the pattern of correlations shown in Table 22.

Lead-In Tasks and Essay

The lead-in tasks measure reading and/or critical thinking ability, and the essays directly evaluate writing ability. It is interesting to explore the relationships between the lead-in tasks and the essays. Figures 1 to 4 show the scatter plots of the subscores in the lead-in tasks versus the essay scores (Task 4 scores) with the LOWESS (locally weighted scatter plot smoothing; Cook & Weisberg, 1999, pp. 42–45) regression lines with a smoothing parameter of 0.6 for the four PAAs, respectively. The LOWESS regression is a locally empirical regression method that does not assume a parametric form. One can see that for each PAA essay score the increase was monotonic with each lead-in subscore, although the increase might not be strictly linear.

The influence of test form on the relationships among the lead-in tasks and the essays was examined. Because each form has a different topic and tests a rather different type of critical thinking skill, it is reasonable to assume that the lead-in tasks are more strongly associated with the final essay in each form than across forms. On the other hand, the lead-in tasks often involve much simpler skills, reflecting a general level of development of reading and/or critical thinking ability, while the essays show common variance due to general verbal fluency and document construction skills. Therefore, it is also possible that the associations among the lead-in tasks across test forms and among the essays across test forms are stronger than those between the lead-in tasks and the essay within test forms. Table 23 shows the means and standard deviations of lead-in scores (sums of the first three task scores) and essay scores, and their correlations across the four PAAs. The comparison of the correlations among lead-in and essay raw scores across four PAAs did not reveal a systemic pattern: The relationships depended on the specific PAAs. See Table 24 for the comparison results separated by each PAA and section.

Table 22

Correlations of the Four Writing Periodic Accountability Assessments (PAAs) With State Tests

State/school	CBAL PAA	Pearson correlation (*N*)			
		ELA	Math	Reading	Writing
Alabama (School A) [a]	Service Learning	.80 (29)	.52 (29)	.74 (29)	
	Invasive Plant Species	.54 (29)	.40 (29)	.58 (29)	
	Ban Ads	.68 (28)	.59 (28)	.61 (28)	
	Mango Street	.68 (33)	.52 (33)	.65 (33)	
Alabama (School B) [a]	Service Learning	.36 (39)	.55 (39)	.43 (39)	
	Invasive Plant Species	.70 (18)	.66 (18)	.62 (18)	
	Ban Ads	.36 (15)	.28 (15)	.32 (15)	
	Mango Street	.39 (21)	.18 (21)	.27 (21)	
Arkansas	Service Learning	.60 (145)	.58 (145)		
	Invasive Plant Species	.56 (117)	.59 (117)		
	Ban Ads	.48 (123)	.53 (123)		
	Mango Street	.62 (109)	.60 (109)		
Arizona	Service Learning	.66 (36)	.54 (152)	.61 (171)	.57 (206)
	Invasive Plant Species	.76 (33)	.55 (113)	.69 (123)	.52 (156)
	Ban Ads	.56 (14)	.66 (117)	.71 (153)	.58 (166)
	Mango Street	.68 (34)	.59 (129)	.67 (137)	.58 (170)
California	Service Learning	.52 (31)	.30 (31)	.49 (31)	.25 (31)
	Invasive Plant Species	.75 (19)	.34 (19)	.68 (19)	.67 (19)
	Ban Ads	.55 (30)	.56 (30)	.46 (30)	.25 (30)
	Mango Street	.68 (28)	.12 (28)	.64 (28)	.45 (28)
Florida	Service Learning		.67 (21)	.68 (21)	
	Invasive Plant Species		.86 (14)	.89 (14)	
	Ban Ads		.76 (18)	.76 (18)	
	Mango Street		.78 (25)	.75 (25)	
Georgia	Service Learning	.60 (42)	.60 (42)	.65 (16)	
	Invasive Plant Species	.48 (65)	.51 (65)	.45 (41)	
	Ban Ads	.58 (39)	.43 (39)	.70 (13)	
	Mango Street	.57 (70)	.53 (70)	.66 (47)	
Kentucky	Service Learning	.70 (19)	.57 (19)		
	Invasive Plant Species	.80 (18)	.67 (18)		
	Ban Ads	.66 (38)	.64 (38)		
	Mango Street	.60 (31)	.52 (31)		
Mississippi	Service Learning	.39 (23)	.35 (23)		
	Invasive Plant Species	.77 (15)	.55 (15)		
	Ban Ads	.57 (42)	.47 (42)		
	Mango Street	.41 (28)	.42 (28)		

State/school	CBAL PAA	Pearson correlation (*N*)			
		ELA	Math	Reading	Writing
Ohio	Service Learning	.45 (58)	.42 (58)		.57 (58)
	Invasive Plant Species	.55 (85)	.54 (85)		.43 (85)
	Ban Ads	.59 (88)	.55 (88)		.55 (88)
	Mango Street	.52 (65)	.27 (65)		.55 (65)
Texas	Service Learning		.56 (75)	.47 (73)	.54 (74)
	Invasive Plant Species		.57 (68)	.44 (66)	.52 (67)
	Ban Ads		.54 (81)	.56 (81)	.47 (82)
	Mango Street		.55 (69)	.42 (68)	.46 (69)
Mean [b]	Service Learning	.56	.51	.58	.49
	Invasive Plant Species	.66	.57	.62	.54
	Ban Ads	.56	.55	.59	.46
	Mango Street	.57	.46	.58	.51

Note. ELA = English language arts, PPA = periodic accountability assessment.

[a] Schools A and B reported test scores at different scales. Therefore, their correlations were calculated separately. [b] The simple average of correlations.

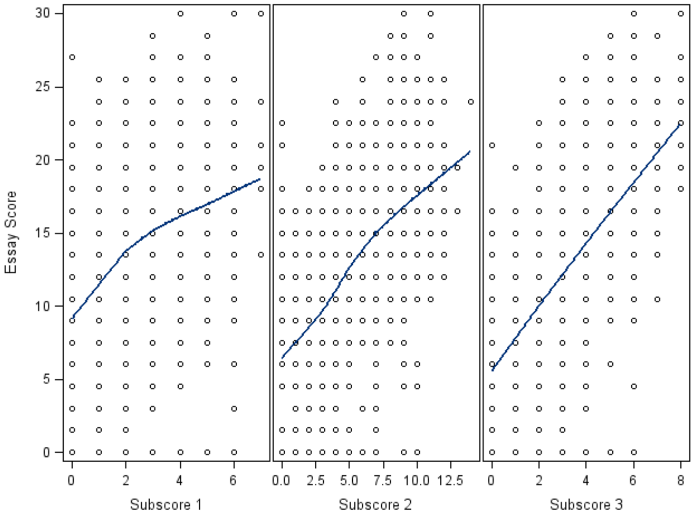

***Figure 1*. Service Learning (PAA 1): scatter plots of lead-in subscores versus essay scores with LOWESS regression lines.**

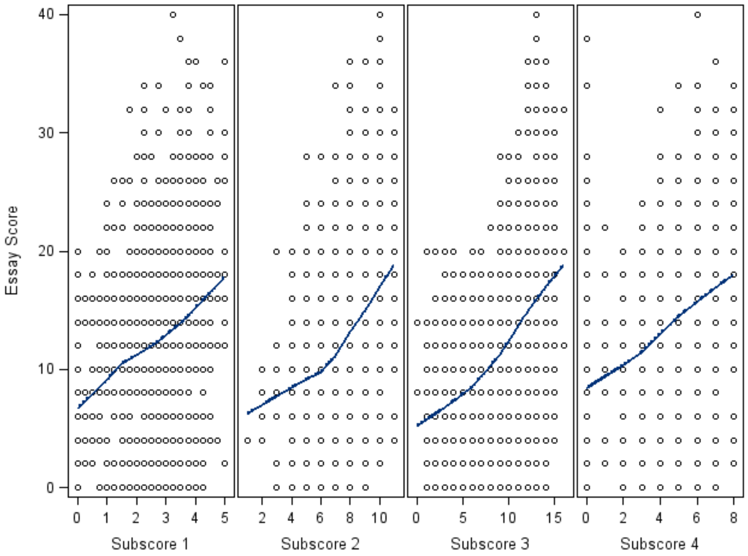

***Figure 2*. Invasive Plant Species (PAA 2): scatter plots of lead-in subscores versus essay scores with LOWESS regression lines.**

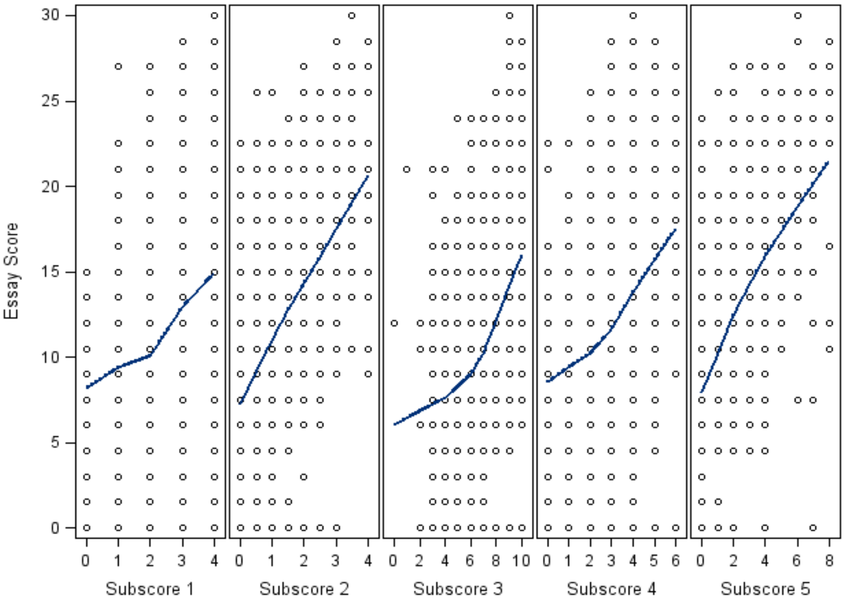

***Figure 3.* Ban Ads (PAA 3): scatter plots of lead-in subscores versus essay scores with LOWESS regression lines.**

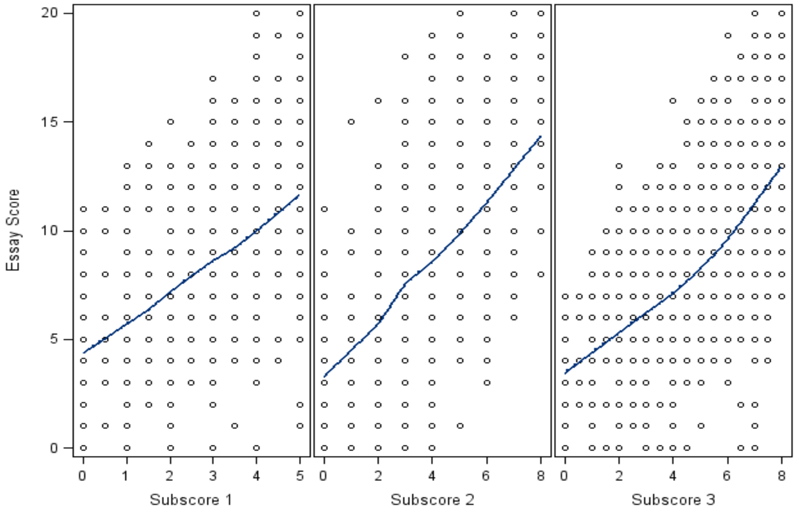

Figure 4. **Mango Street (PAA 4): scatter plots of lead-in subscores versus essay scores with LOWESS regression lines.**

Table 23

Correlations Among Lead-In and Essay Raw Scores Across Four Periodic Accountability Assessments (PAAs)

PAA	Section	Mean	SD	Service Learning (*N*)		Invasive Plant Species (*N*)		Ban Ads (*N*)		Mango Street (*N*)
				Lead-in	Essay	Lead-in	Essay	Lead-in	Essay	Lead-in
Service Learning	Lead-in	11.34	5.71	-	-	-	-	-	-	-
	Essay	13.17	6.39	.66(1,057)	-	-	-	-	-	-
Invasive Plant Species	Lead-in	21.78	8.19	.64(271)	.60(271)	-	-	-	-	-
	Essay	12.09	7.98	.40(271)	.44(271)	.53(912)	-	-	-	-
Ban Ads	Lead-in	16.22	5.38	.65(375)	.59(375)	.57(200)	.54(200)	-	-	-
	Essay	11.86	6.54	.63(375)	.65(375)	.49(200)	.55(200)	.68(1,025)	-	-
Mango Street	Lead-in	11.03	4.50	.64(246)	.61(246)	.74(326)	.57(326)	.63(286)	.63(286)	-
	Essay	8.17	4.05	.55(246)	.58(246)	.62(326)	.56(326)	.58(286)	.67(286)	.70(1,067)

Table 24

Correlation Comparison Results Among Lead-In and Essay Across Four Periodic Accountability Assessments (PAAs)

PAA	Section	Comparison
Service Learning	Lead-in	The correlation with the within PAA essay was similar to those with lead-ins, and correlations with lead-ins were higher than with essays.
	Essay	The correlation with the within PAA lead-in was the highest, and correlations with lead-ins were higher than with essays except for Ban Ads where the order was the reverse.
Invasive Plant Species	Lead-in	The correlations with lead-ins were higher than with essays, and all correlations were higher than the correlation with the within essay except for with the Ban Ads essay.
	Essay	All correlations were close, except for low correlations with the Service Learning lead-in and essay.
Ban Ads	Lead-in	The correlation with the within PAA essay was the highest, and correlations with lead-ins were higher than with essays.
	Essay	The correlation with the within PAA lead-in was the highest, and correlations with essays were higher than with lead-ins.
Mango Street	Lead-in	The correlations with lead-ins were higher than with essays except for Ban Ads, where they were equal; the correlation with the Invasive lead-in was higher than with the within PAA essay.
	Essay	The correlation with the within PAA essay was the highest, and correlations with essays were higher than with lead-ins except for Invasive where the order was reverse.

Item Response Theory (IRT) Item Calibration and Scaling

The four PAAs were calibrated using the unidimensional generalized partial credit model (GPCM; Muraki, 1992). Two calibration approaches, concurrent calibration and separate calibration, were carried out, and the item parameter estimates and ability (theta) estimates were compared. In this study, the GPCM was formulated as the following:

$$P_{ijs_{im}} = P(x_{ij} = s_{im} \mid \theta_j, a_i, b_i, \mathbf{d}_i) = \frac{\exp(a_i\theta_j s_{im} - b_i s_{im} + \sum_{h=0}^{m} d_{ih})}{\sum_{v=0}^{M_i-1} \exp(a_i\theta_j s_{iv} - b_i s_{iv} + \sum_{h=0}^{v} d_{ih})}$$

where

$a_i\theta_j s_{i0} - b_i s_{i0} + d_{i0} \equiv 0$;

x_{ij} is examinee *j*'s score on item *i*;

s_{im} is the score of item *i*'s score category *m* ($m = 0$ to $M_i - 1$);

b_i is the location parameter (or difficulty parameter for dichotomous items) for item i;

d_{ih} is the step parameter for score category h (h = 0 to M_i –1), and $\sum_{h=1}^{M_i-1} d_{ih}$ is constrained to 0 for model identification purpose;

$\mathbf{d}_i$ is the vector with elements d_{ih};

a_i is the discrimination (slope) parameter for item i;

θ_j is examinee j's latent (theta) score; and

$P_{ijs_{im}}$ is the probability of getting score s_{im} on item i conditioned on examinee j's theta and item i's parameters.

In the four PAAs, dichotomous items had scores 0 or 1. And the polytomous items had up to 11 score categories, and the scores assigned to each category are shown in Table 25. For example, W_SERVLEARN17_I had 11 score categories ranging from 0 to 15 with the interval of 1.5.

Concurrent Versus Separate Calibrations

Recall that in the current test design each student took two Writing PAAs out of the four PAAs within a short period, and there was no common item between PAAs. In the concurrent calibration, all the items were calibrated together, and for the test forms that a student did not take, their item responses were treated as missing in estimating item parameters. Examinees were assumed to be from a common population, and PAAs were linked together by the common PAAs that examinees took. Then, the item parameter estimates from the concurrent calibration were used to estimate thetas for each examinee on each PAA. In the separate calibration, each PAA was calibrated separately, and the item parameter and theta estimates of the four PAAs from the separate calibrations were assumed to be on the same scale by the assumption of equivalent examinee groups. The expected a posterior (EAP) method was used to estimate theta.

Table 26 shows the sample sizes used in the item calibration and EAP theta estimation for each calibration. Note that a student with any missing value in a PAA was excluded from the theta estimation in the PAA.

Table 25

Score Categories (SC) for Polytomous Items

Item	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10	SC11
W_BANADS_01B	0	.5	1	1.5	2	-	-	-	-	-	-
W_BANADS_01C	0	.5	1	1.5	2	-	-	-	-	-	-
W_BANADS_03	0	1	2	3	4	5	6	7	8	-	-
W_BANADS_04_I	0	1.5	3	4.5	6	7.5	9	10.5	12	13.5	15
W_BANADS_04_III	0	1.5	3	4.5	6	7.5	9	10.5	12	13.5	15
W_INVASIVE_01_01[a]	0	.5	1	1.5	2	2.5	3	3.5	4	4.5	5
W_INVASIVE_03_01	0	1	2	3	4	5	6	7	8		-
W_INVASIVE_04_02_I [b]	0	2	4	6	8	10	12	14	16	18	-
W_INVASIVE_04_02_III [b]	0	2	4	6	8	10	12	14	16	18	-
W_MANGO_01_01	0	.5	1	-	-	-	-	-	-	-	-
W_MANGO_01_02	0	.5	1	-	-	-	-	-	-	-	-
W_MANGO_01_03	0	.5	1	-	-	-	-	-	-	-	-
W_MANGO_01_04	0	.5	1	-	-	-	-	-	-	-	-
W_MANGO_01_05	0	.5	1	-	-	-	-	-	-	-	-
W_MANGO_02_01	0	1	2	3	4	5	6	7	8	-	-
W_MANGO_03_06	0	.5	1	1.5	2	2.5	3	-	-	-	-
W_MANGO_04_I	0	1	2	3	4	5	6	7	8	9	10
W_MANGO_04_III	0	1	2	3	4	5	6	7	8	9	10
W_SERVLEARN16	0	1	2	3	4	5	6	7	8		
W_SERVLEARN17_I	0	1.5	3	4.5	6	7.5	9	10.5	12	13.5	15
W_SERVLEARN17_III	0	1.5	3	4.5	6	7.5	9	10.5	12	13.5	15

[a] Decimal scores .25 and .75 were rounded to .50 and 1, respectively. [b] Score 20 was combined with Score 18 in order for smooth calibration.

Table 26

Sample Sizes Used in Item Response Theory (IRT) Calibrations

Estimation	Separate calibration				Concurrent calibration
	Service Learning	Invasive Plant Species	Ban Ads	Mango Street	
Item parameter	1,195	1,219	1,161	1,213	2,580
EAP theta	1,057	912	1,025	1,067	NA

Note. EAP = expected a posterior.

In Figure 5 we compare estimates of item discrimination parameters (a_i), item location parameters (b_i), and item step parameters (d_{ih}) between the concurrent and separate calibrations for the four PAAs. The item parameter estimates from both calibrations were highly correlated (≥ 0.96) except for the slope parameters in Ban Ads and Mango Street, and the location parameters in Service Learning. The low correlation for the location parameters in Service Learning was caused by the large differences, relative to other items, in the location parameter estimates of the two items, W_SERVLEARN09H and W_SERVLEARN10H, between the two calibrations.

Figure 6 shows the comparisons of EAP theta estimates between the two calibration approaches for the four PAAs. The EAP theta estimates from the separate calibrations were transformed to have the same mean and standard deviation of the combined four PAAs as the ones from the concurrent calibration so that they were in the same metric. EAP theta estimates were almost perfectly correlated for all PAAs (≥ 0.99), and the root mean squared differences (RMSDs) for Service Learning and Invasive Plant Species were quite small. It appears that the two items having large differences in location parameter estimates did not have much influence on EAP theta estimates. However, the RMSDs for Ban Ads and Mango Street were much higher, at .17 and .13, respectively. From the plots, one can see that some points deviated considerably from the diagonal line.

In conclusion, there were some differences in item parameter estimates from the separate and concurrent calibrations, especially for Ban Ads, Mango Street, and Service Learning. The calibrations produced very similar EAP theta estimates for Service Learning and Invasive Plant Species, and highly correlated but somewhat different EAP theta estimates for Ban Ads and Mango Street. Because in practice theta estimates are often of ultimate concern, we give greatest weight to those estimates when evaluating results from the different calibrations. In the current study, there were anchor tests in the concurrent calibration, while the separate calibrations were based on the assumption of equivalent groups, which might not be true. Therefore, in the remainder of this report the results from the concurrent calibration are reported and used.

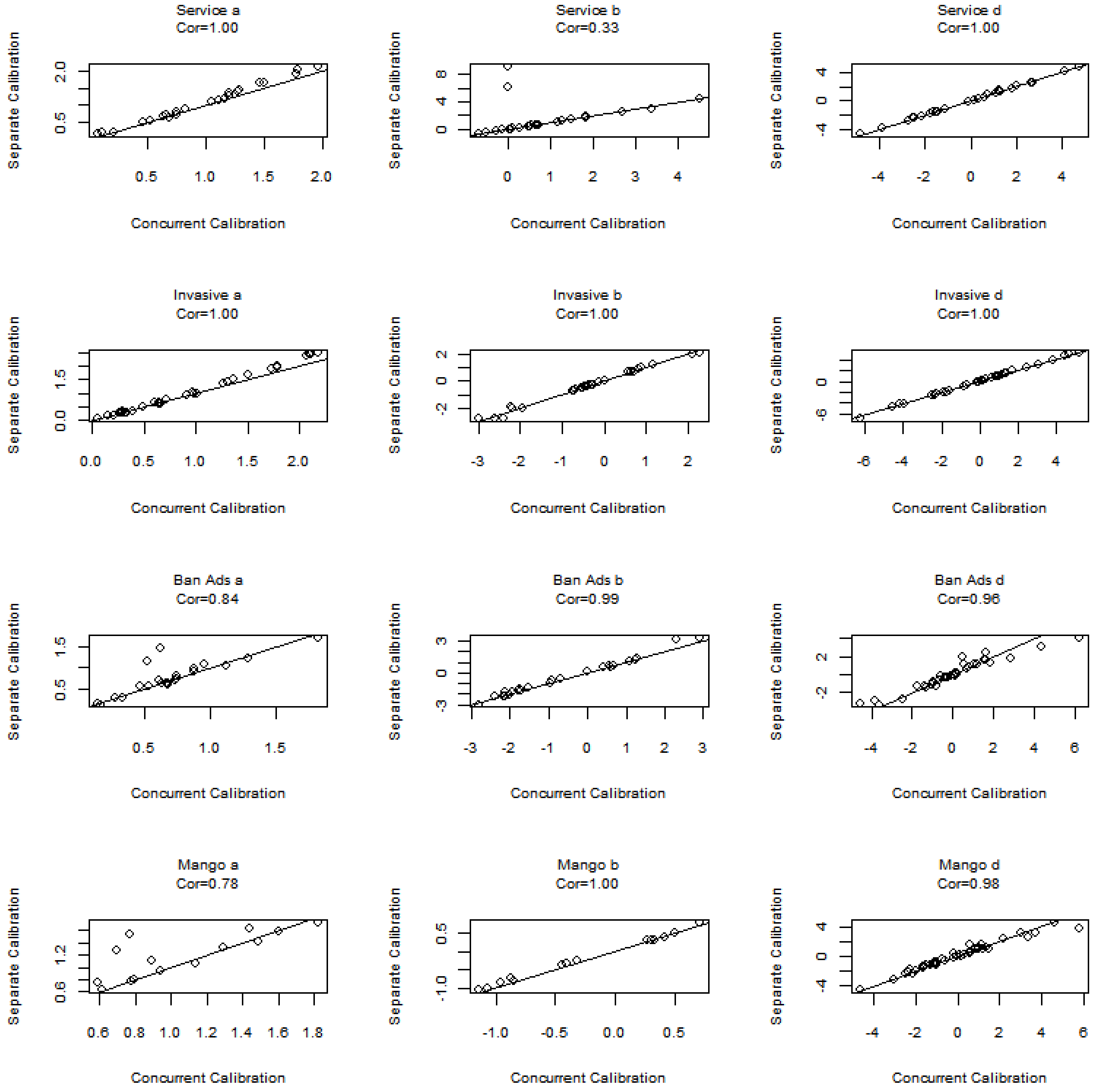

Figure 5**. Comparison of item parameters between concurrent calibrations and separate calibrations; the lines in the plots are diagonal lines.**

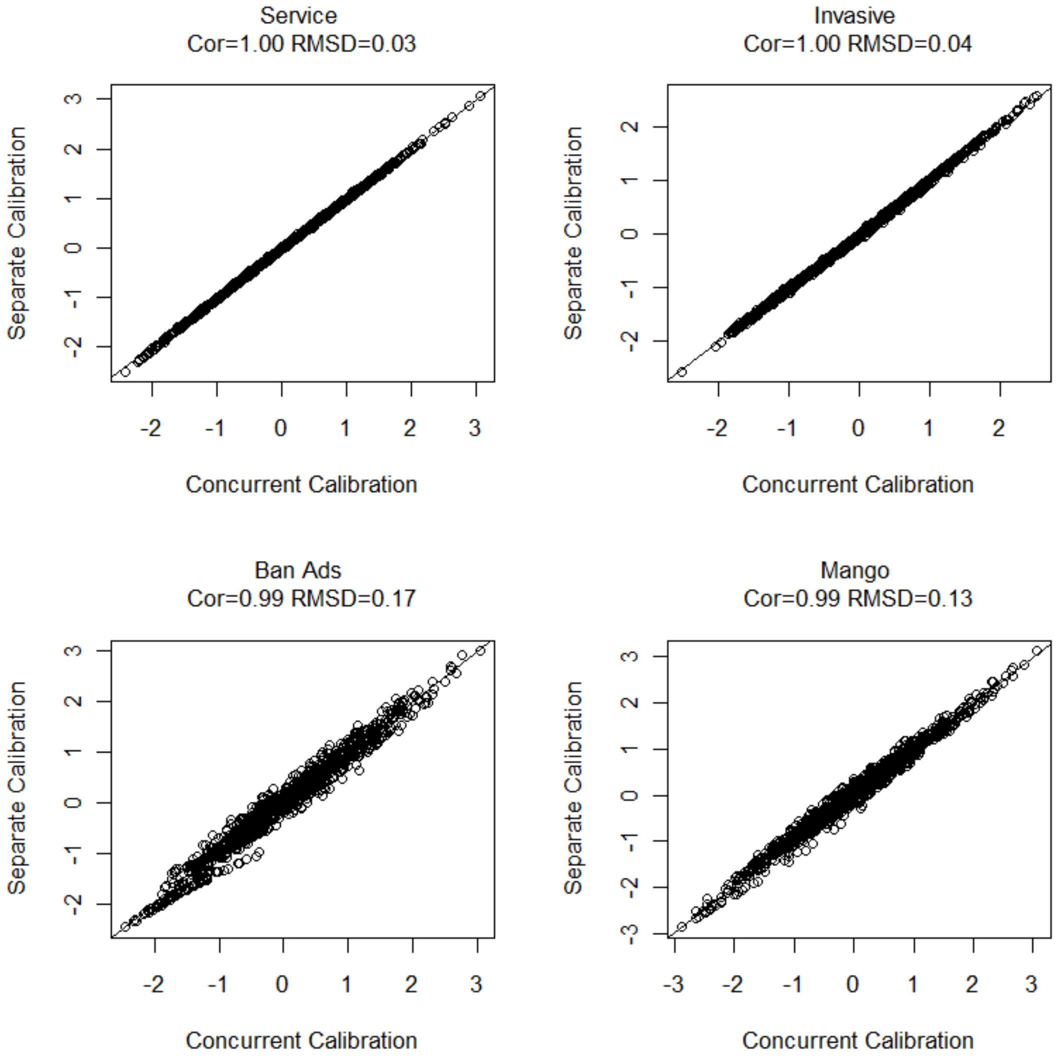

***Figure 6*. Comparison of EAP thetas between concurrent calibrations and separate calibrations; the lines in the plots are diagonal lines.**

Item Parameter and Theta Estimates

Tables 27 to 30 list the item parameter estimates, standard errors, and significance levels of item chi-square fit statistics for the four PAAs. Across the four PAAs, the item discrimination parameter estimates (a_i) were between .06 and 2.18, the item location parameter estimates (b_i) were between -2.99 and 4.51, and the item step parameter estimates (d_{ih}) were in the range from-6.28 to 6.25.

Table 27

Service Learning (PAA 1): Item Parameter Estimates and Standard Errors

Item	Slope		Location		Step 1		Step 2		Step 3		Step 4		Step 5		Step 6		Step 7		Step 8		Step 9		Step 10		Sig.
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	
W_SERVLEARN01	1.20	.09	.71	.07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN02	.64	.08	2.69	.33	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN03	.46	.08	3.40	.57	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN04	.74	.07	.53	.09	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN05	1.20	.10	1.28	.09	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN06	.82	.09	1.85	.17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN07	1.28	.09	.30	.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN08H	.66	.07	1.50	.17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN08G	1.96	.11	.04	.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN09H	.08	.06	.00	.74	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**
W_SERVLEARN09G	1.11	.08	.11	.06	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN10H	.11	.06	.00	.54	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**
W_SERVLEARN10G	.69	.08	1.85	.20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN11H	.21	.07	4.51	1.41	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN11G	1.79	.11	-.13	.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN12H	1.05	.08	.72	.07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN12G	1.28	.10	1.18	.08	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN13H	1.46	.09	-.27	.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*
W_SERVLEARN13G	1.50	.10	-.53	.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN14H	1.77	.11	.09	.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN14G	1.16	.09	-.68	.07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN16	.74	.03	.58	.03	1.25	.19	2.00	.18	1.09	.15	1.29	.13	-1.10	.14	-.09	.17	-2.72	.28	-1.73	.43	-	-	-	-	-
W_SERVLEARN17_I	.52	.02	.48	.03	.71	.40	4.13	.39	1.81	.28	2.67	.24	.58	.20	.36	.20	-1.46	.24	-1.49	.30	-2.46	.39	-4.85	.76	-
W_SERVLEARN17_III	.52	.02	.67	.03	1.28	.37	4.78	.34	1.28	.24	2.72	.22	.18	.19	-.08	.21	-1.62	.27	-2.51	.40	-3.88	.73	-2.16	.94	-
W_INVASIVE_02_14	1.79	.11	-.71	.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_15	1.79	.11	-.26	.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Item	Slope		Location		Step 1		Step 2		Step 3		Step 4		Step 5		Step 6		Step 7		Step 8		Step 9		Step 10		Sig.
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	
W_INVASIVE_02_16	1.74	.11	-.67	.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*
W_INVASIVE_03_01	.39	.01	.88	.04	-6.28	.56	4.47	.62	1.42	.40	1.76	.35	.93	.31	-.03	.31	-.66	.35	-1.59	.44	-	-	-	-	-
W_INVASIVE_04_02_I	.20	.01	.83	.04	2.47	.58	3.11	.59	.85	.63	3.86	.61	.21	.57	.34	.61	-3.96	.79	-4.58	1.19	-2.29	1.52	-	-	-
W_INVASIVE_04_02_III	.32	.02	1.17	.04	5.27	.36	4.73	.29	.63	.29	.11	.38	-.64	.48	-1.58	.64	-1.89	.83	-2.47	1.07	-4.16	1.54	-	-	-

Note. Est = estimate, Sig = significance of chi squared goodness of fit statistic.

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

Table 28

Invasive Plant Species (PAA 2): Item Parameter Estimates and Standard Errors

Item	Slope		Location		Step 1		Step 2		Step 3		Step 4		Step 5		Step 6		Step 7		Step 8		Step 9		Step 10		Sig.
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	
W_INVASIVE_01_01	.66	.03	.56	.04	-1.73	.28	1.02	.35	1.33	.29	1.20	.23	.94	.18	.31	.17	.15	.16	-.09	.17	-.80	.20	-2.33	.35	-
W_INVASIVE_01_02	.91	.07	-.31	.07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_03	.29	.06	.65	.24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_04	1.01	.11	-1.96	.17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_05	.15	.06	2.28	1.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_06	.26	.06	-2.23	.60	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_07	.97	.11	-2.19	.21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_08	.64	.09	-2.60	.34	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_09	.61	.07	-.39	.11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_10	.33	.07	-2.40	.52	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_11	.49	.08	-2.99	.49	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_13	.29	.06	-.13	.22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_01	2.12	.12	-.54	.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Item	Slope		Location		Step 1		Step 2		Step 3		Step 4		Step 5		Step 6		Step 7		Step 8		Step 9		Step 10		Sig.
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	
W_INVASIVE_02_02	.98	.08	-.42	.07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_03	.64	.07	.72	.12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_04	.72	.07	-.44	.09	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_05	.06	.06	.00	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**
W_INVASIVE_02_06	1.27	.09	-.46	.06	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_07	2.18	.13	-.49	.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**
W_INVASIVE_02_08	2.10	.13	-.72	.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_09	1.30	.09	-.28	.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_10	1.51	.09	-.36	.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_11	1.36	.09	-.76	.06	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_12	.27	.06	2.13	.52	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_13	2.08	.12	-.41	.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_14	1.79	.11	-.71	.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_15	1.79	.11	-.26	.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_16	1.74	.11	-.67	.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*
W_INVASIVE_03_01	.39	.01	.88	.04	-6.28	.56	4.47	.62	1.42	.40	1.76	.35	.93	.31	-.03	.31	-.66	.35	-1.59	.44	-	-	-	-	-
W_INVASIVE_04_02_I	.20	.01	.83	.04	2.47	.58	3.11	.59	.85	.63	3.86	.61	.21	.57	.34	.61	-3.96	.79	-4.58	1.19	-2.29	1.52	-	-	-
W_INVASIVE_04_02_III	.32	.02	1.17	.04	5.27	.36	4.73	.29	.63	.29	.11	.38	-.64	.48	-1.58	.64	-1.89	.83	-2.47	1.07	-4.16	1.54	-	-	-

Note. Est = estimate, Sig = significance of chi squared goodness of fit statistic.

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

Table 29

Ban Ads (PAA 3): Item Parameter Estimates and Standard Errors

Item	Slope		Location		Step 1		Step 2		Step 3		Step 4		Step 5		Step 6		Step 7		Step 8		Step 9		Step 10		Sig.
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	
W_BANADS_01A_02	.67	.08	-2.03	.22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_01A_03	.54	.07	-.70	.13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**
W_BANADS_01A_04	.17	.06	2.29	.90	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_01A_05	.66	.08	-2.17	.24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_01B	1.82	.09	1.09	.04	.02	.05	.69	.06	-.31	.07	-.40	.09	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_01C	1.29	.06	.57	.04	-.27	.08	1.07	.08	-.97	.09	.18	.11	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02AX_A	.75	.08	-1.72	.17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02AX_B	.68	.08	-2.15	.24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**
W_BANADS_02AX_C	.28	.06	-1.93	.45	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02AX_D	.96	.09	-1.80	.15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02AX_E	.88	.09	-1.76	.15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02AX_F	.75	.09	-2.40	.25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02AX_G	.47	.07	-2.12	.32	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02AX_H	1.12	.11	-2.13	.17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02AX_I	.88	.08	-1.52	.13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02AX_J	.61	.07	-.90	.13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02BX_A	.15	.06	-2.80	1.15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02BX_B	.28	.06	1.25	.35	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02BX_C	.68	.07	.42	.10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02BX_D	.64	.07	-.97	.13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02BX_E	.34	.07	2.90	.59	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_02BX_F	.73	.07	.01	.09	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_03	.63	.03	1.28	.04	-.13	.16	1.57	.19	.75	.20	1.17	.19	-1.06	.24	-.27	.32	-1.37	.43	-.66	.52	-	-	-	-	**
W_BANADS_04_I	.63	.03	.61	.02	1.63	.30	4.37	.28	1.65	.19	1.83	.18	.09	.18	.20	.20	-.94	.23	-1.35	.27	-3.86	.48	-3.62	.84	-
W_BANADS_04_III	.52	.02	.69	.03	.51	.42	6.25	.39	.53	.23	2.82	.22	-.58	.22	.09	.25	-1.76	.32	-.82	.38	-4.57	.67	-2.47	.96	

Note. Est = estimate, Sig = significance of chi squared goodness of fit statistic.

** *p* < .01.

Table 30

Mango Street (PAA 4): Item Parameter Estimates and Standard Errors

Item	Slope		Location		Step 1		Step 2		Step 3		Step 4		Step 5		Step 6		Step 7		Step 8		Step 9		Step 10		Sig.
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	
W_MANGO_01_01	1.13	.05	.71	.07	-2.17	.14	2.17	.15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_MANGO_01_02	1.60	.07	-.40	.05	-1.02	.07	1.02	.07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_MANGO_01_03	.62	.03	-1.14	.12	-4.66	.31	4.66	.30	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_MANGO_01_04	1.49	.08	-.84	.06	-3.01	.30	3.01	.30	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_MANGO_01_05	1.82	.08	.27	.04	-1.03	.07	1.03	.08	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_MANGO_02_01	.59	.02	.31	.04	.86	.32	3.69	.27	.65	.17	1.33	.15	-1.16	.16	-1.30	.22	-1.60	.29	-2.47	.41	-	-	-	-	*
W_MANGO_03_01	.79	.07	-1.07	.11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_MANGO_03_02	1.44	.10	-.87	.06	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_MANGO_03_03	.94	.07	-.32	.07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_MANGO_03_04	1.29	.09	-.96	.07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_MANGO_03_05	.78	.07	-.44	.08	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_MANGO_03_06	.89	.05	.33	.04	.10	.12	.62	.13	.30	.12	.55	.11	-.99	.12	-.57	.16	-	-	-	-	-	-	-	-	*
W_MANGO_04_I	.70	.03	.42	.03	1.16	.29	3.37	.25	1.01	.16	1.19	.16	-.06	.16	.16	.17	-1.11	.19	-1.10	.24	-2.29	.34	-2.35	.51	-
W_MANGO_04_III	.77	.03	.51	.03	.61	.44	5.83	.38	-.20	.14	1.50	.14	-.74	.15	-.19	.17	-1.49	.22	-1.68	.32	-1.67	.41	-1.98	.51	-

Note. Est = estimate, Sig = significance of chi squared goodness of fit statistic.

* p < .05.

The mean item location parameter estimates were .87, -.40, -.64, and -.25 for the four PAAs. Therefore, PAA 1 was more difficult than PAAs 2 to 4, which is consistent with the result from the *p*+ values. The item fit tests indicate that the model did not fit 11 items very well. Note that the two items, W_SERVLEARN09H and W_SERVLEARN10H, having large difference on the location parameters between the two calibrations, were poorly fitted items. Appendix D lists the item fit statistics for all items.

Figures 7 and 8 show the test information curves and test characteristic curves, respectively, for the four PAAs based on the EAP theta estimates and EAP true score estimates. For all PAAs the test information curves had the same shape; however, they had different modes and spreads.

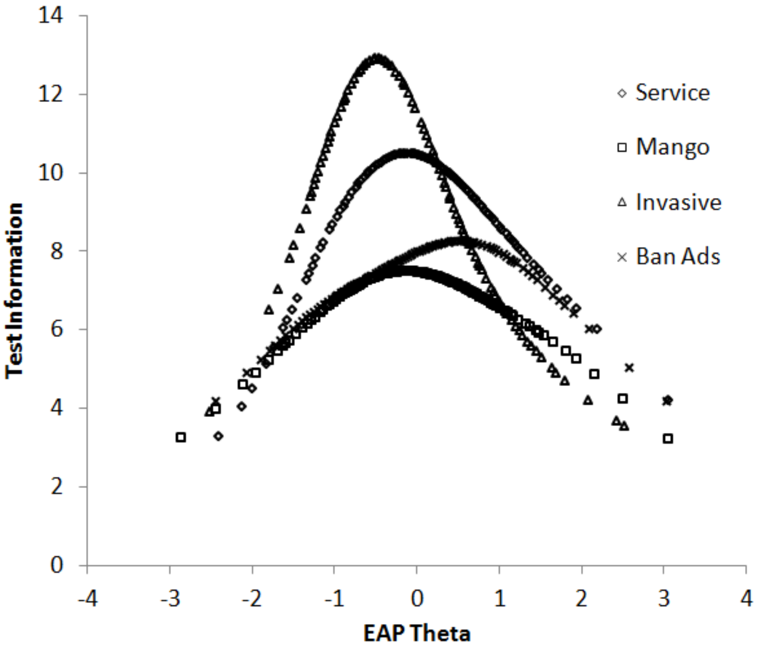

***Figure 7*. Test information curves based on EAP theta estimates.**

Figure 9 includes the histograms of the distributions of the EAP theta estimates for the four PAAs. The four PAAs had theta means of .05, .01, -.03, and -.02, and standard deviations of .97, .97, .99, and 1.01, respectively. Figure 9 also shows that the theta reliability estimates for the four PAAs were between .87 and .89. The theta reliability for a test was estimated by the formula (Haberman & Sinharay, 2010):

$$\hat{R} = 1 - \frac{N^{-1}\sum_{j=1}^{N}\overset{\wedge}{Var}(\theta_j)}{\overset{\wedge}{Var}(\boldsymbol{\theta})},$$

where $\overset{\wedge}{Var}(\theta_j)$ is the estimated posterior variance of examinee *j*'s theta, $\overset{\wedge}{Var}(\boldsymbol{\theta})$ is the estimated posterior population variance of theta, and *N* is the total number of examinees.

Analyses of Factors Affecting Test Scores

The effects of PAA, test order, and demographic groups on test scores were evaluated using *t*-tests, one-way analysis of variance (ANOVA), multiple comparisons, and mixed models.

Subgroup Comparison

Table 31 provides *t*-test results as well as means and standard deviations of raw scores and theta estimates on each PAA for gender and (SES). Statistically significant differences were found for gender and SES groups across the four PAAs. The male and the economically disadvantage groups had significantly lower test scores than their respective comparison groups across the four PAAs.

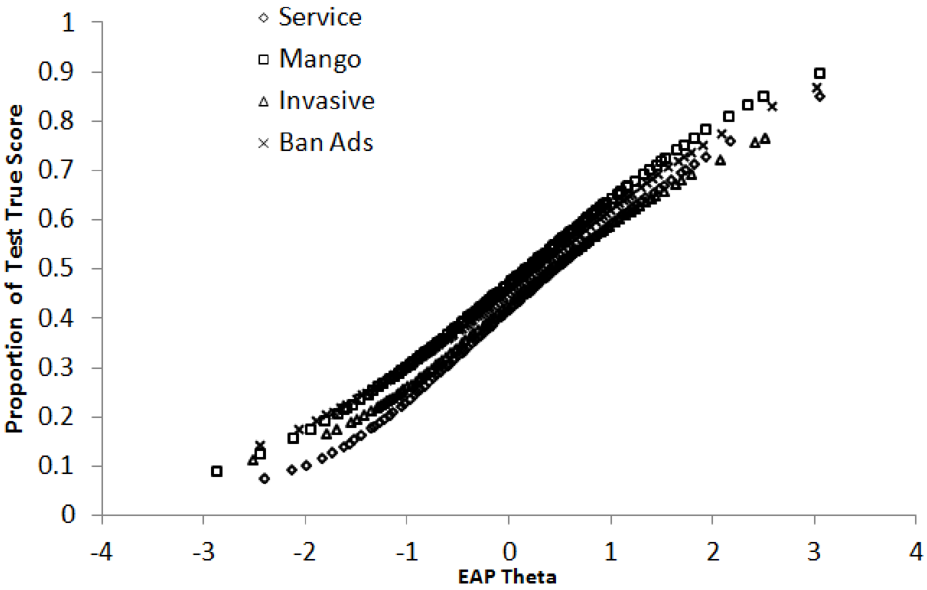

***Figure 8*. Test characteristic curves based on EAP theta and EAP true score estimates.**

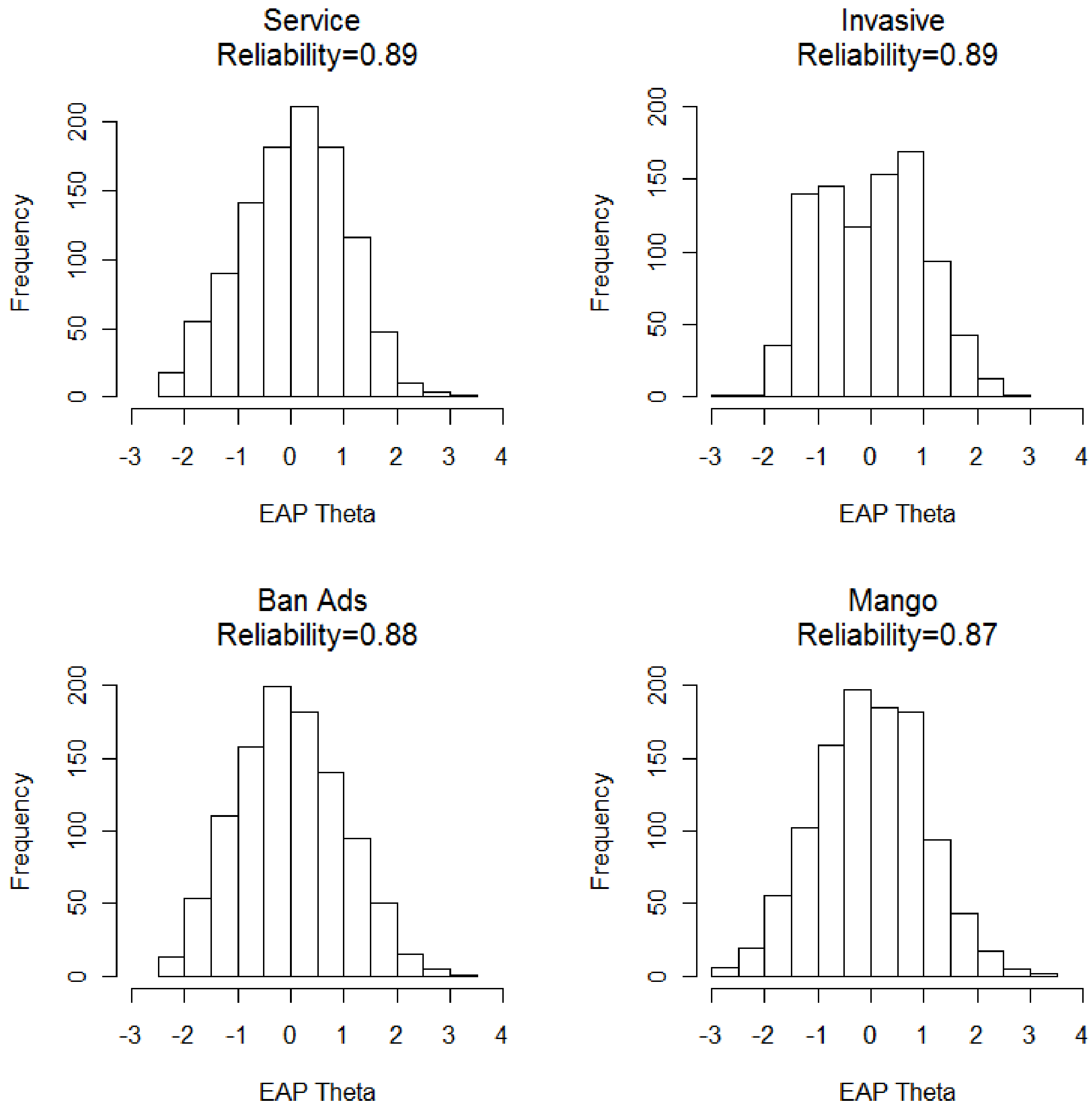

Figure 9. **EAP theta estimate distributions.**

Because the race subgroup had four subgroups, one-way ANOVAs on each PAA were first carried out on ethnic groups for theta estimates and raw scores. As shown in Table 32, all the one-way ANOVA tests were significant. Therefore, multiple comparisons (Tukey HSD test) were conducted on all pairs of racial/ethnic groups, and the group pairs having significant differences are shown in Table 32. Table 32 also provides the means and standard deviations of the theta estimates and raw scores for each racial/ethnic group in each PAA. One can see that, across the four PAAs, the order of the test scores of the four racial/ethnic groups from high to low was Asian/Pacific Islander, White, Hispanic, and African American, and most of the score differences between racial/ethnic groups were statistically significant. In the following section, besides the demographics, the school, PAA, and test-order effects on test scores were examined.

Mixed Model

Mixed models were used to check the school, PAA, and test-order effects on test scores. The dependent variable was students' theta estimates on each PAA from the GPCM IRT calibrations.

Table 31

Subgroup Comparison on Each Periodic Accountability Assessment (PAA)

Subgroup	Category	*N*	Theta				Raw score			
			Mean	SD	*t* value	*p* value	Mean	SD	*t* value	*p* value
Service Learning										
Gender	M	441	-.16	.98	-5.98	.00**	21.95	11.14	-6.88	.00**
	F	441	.23	.95			27.00	10.68		
Low SES	N	338	.14	1.00	5.20	.00**	25.89	11.46	5.56	.00**
	Y	271	-.26	.91			20.91	10.36		
Invasive Plant Species										
Gender	M	380	-.14	.94	-5.06	.00**	31.45	13.61	-5.83	.00**
	F	399	.20	.97			37.26	14.16		
Low SES	N	273	.19	1.00	6.35	.00**	36.50	14.96	6.28	.00**
	Y	266	-.32	.85			29.10	12.28		
Ban Ads										
Gender	M	436	-.20	.92	-6.94	.00**	26.17	10.05	-7.05	.00**
	F	435	.26	1.03			31.32	11.47		
Low SES	N	271	.09	.99	4.88	.00**	29.41	10.97	4.90	.00**
	Y	292	-.30	.92			25.06	10.12		
Mango Street										
Gender	M	446	-.24	.97	-8.80	.00**	17.36	7.46	-9.50	.00**
	F	412	.34	.98			22.25	7.60		
Low SES	N	287	.25	1.05	6.11	.00**	21.41	8.17	6.19	.00**
	Y	297	-.25	.90			17.51	7.00		

Note. SES = socioeconomic status.

** $p < .01$.

Table 32

Race Subgroup Comparison on Each Periodic Accountability Assessment (PAA)

Race	*N*	Theta				Theta: multiple comparison[a]			Raw score				Raw score: multiple comparison[a]		
		Mean	SD	*F* value	*p* value	1	2	3	Mean	SD	*F* value	*p* value	1	2	3
Service Learning															
1	29	.67	1.08	32.46	.00**	-	-	-	32.26	11.98	33.97	.00**	-	-	-
2	141	-.51	.86			*	-	-	17.99	10.02			*	-	-
3	478	.27	.89			-	*	-	27.02	10.24			*	*	
4	88	-.11	.93			*	*	*	22.68	10.72			*	*	*
Invasive Plant Species															
1	23	1.11	.88	22.60	.00**	-	-	-	51.22	12.90	26.12	.00**	-	-	-
2	127	-.46	.80			*	-	-	26.54	11.51			*	-	-
3	399	.06	.97			*	*	-	35.02	14.16			*	*	-
4	75	-.18	.91			*	-	-	31.90	13.92			*	*	-
Ban Ads															
1	28	1.26	.82	44.66	.00**	-	-	-	42.04	9.23	41.78	.00**	-	-	-
2	163	-.56	.88			*	-	-	22.43	9.73			*	-	-
3	406	.18	.94			*	*	-	30.35	10.47			*	*	-
4	78	-.13	.85			*	*	*	26.65	9.18			*	*	*
Mango Street															
1	26	.97	.90	29.50	.00**	-	-	-	26.98	7.45	29.06	.00**	-	-	-
2	166	-.49	.94			*	-	-	15.65	7.19			*	-	-
3	422	.22	1.00			*	*	-	21.15	7.86			*	*	-
4	82	-.08	.95			*	*	-	-	-			*	*	-

Note. 1 = Asian/Pacific Islander, 2 = African American, 3 = White, 4 = Hispanic.

[a] Tukey HSD test.

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

In the full model, the random effects were school and student-within-school, and the fixed effects were PAA (A or B), test order (Test 1 or Test 2), and their interaction effect. Because the interaction was not significant (*p* = .70), it was dropped from the full model. The model comparisons show that school and student-within-school were significant random effects (both *p*s = .00). The final model estimates are shown in Table 33, which indicates that both PAA and test-order effects were significant. Table 34 shows that students performed better on the first PAA than the second PAA no matter which PAA they took first and that the theta means were different across the four PAAs. We also added the demographic variables to the final model to compare subgroup performance, and the results are shown in Table 35. The demographics (gender, SES, and race/ethnicity) had statistically significant effects on theta estimates, and test order was still significant; however, PAA was not statistically significant for this model once the demographic variables were taken into account. Note that gender, SES, and race/ethnicity were also significant in the above *t*-tests and one-way ANOVAs.

Table 33

Mixed Model for Periodic Accountability Assessment (PAA) and Test Order Effects

Fixed effect	Numerator *df*	Denominator *df*	*F* value	*p* value	Random effect	Variance
Order	3	1,435	113.98	.00	School	.26
PAA	1	1,435	4.73	.00	Student nested in school	.52
-	-	-	-	-	Residual	.25

Note. N = 3,394.

Table 34

Mean and Standard Deviation of Theta Estimates by Test Order and Periodic Accountability Assessment (PAA)

Test order	Service Learning		Invasive Plant Species		Ban Ads		Mango Street		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	.21	.88	.06	.91	.09	.91	.05	1.06	.10	.95
2	-.13	1.03	-.09	1.06	-.14	1.05	-.11	.94	-.12	1.02
Total	.05	.97	.01	.97	-.03	.99	-.02	1.01	.00	.99

Table 35

Mixed Model With Subgroup Comparisons

Fixed effect	Numerator *df*	Denominator *df*	*F* value	*p* value	Random effect	Variance
Order	1	815	72.01	.00	School	.20
PAA	3	815	1.80	.15	Student nested in school	.41
Gender	1	815	53.67	.00	Residual	.21
SES	1	815	44.13	.00	-	-
Race	3	815	25.55	.00	-	-

Note. N = 1,963. PAA = periodic accountability assessment, SES = socioeconomic status.

Summary

The psychometric properties of the fall 2009 CBAL Writing PAAs were studied under both classical test theory and IRT models. Classical item statistics and IRT item parameter estimates were reported. The summary statistics and reliabilities of raw subscores and total scores, and IRT theta scores were presented. In addition, in the report we explored the effects of various factors (such as school, PAA, test order, task, item, student, and demographic characteristics) on item and test scores. The main findings are as follows:

1. The classical item statistics and IRT item parameter estimates using the GPCM show all items performed reasonably well except for the three items, W_SERVLEARN15, W_INVASIVE_01_12, and W_BANADS_01A_01, which had zero or negative correlations with the total test scores and were removed from the test analyses. For the human-scored items, the weighted kappa coefficients showed good to very good rater agreement. The missing response rates were smaller than 5%, except for most items in Task 1 in PAA 2 (Invasive Plant Species) and W_BANADS_01C. Only two items (W_SERVLEARN16, and W_BANADS_01C) had Category C DIF.

2. The total raw scores of the four PAAs had reliabilities (standardized Cronbach alpha) between .79 and .86, and they 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 four PAAs. PAA 1 (Service Learning) was more difficult than PAAs 2 to 4, which had similar levels of difficulty. The correlations among the four PAAs were between .66 and .76. For all PAAs the inter-subscore correlations were between .18 and .64, and most were intermediate.

3. The total raw scores for all PAAs had intermediate correlations with some state tests on ELA, math, reading and writing, which provides some evidence to support the construct validity of the PAAs. The intermediate correlations with the state math tests may indicate the involvement of reading and writing skills in the math tests to some degree. The relatively low correlations with the state writing tests signify the difference between the CBAL writing tests and the state writing tests: In the authors' opinion, the state writing tests measured writing skills very narrowly, while the CBAL writing PAAs also measured reading skills in addition to writing skills (i.e., writing from reading).

4. Within each PAA, each lead-in subscore monotonically increased with essay score; however, the correlations among the total scores of the lead-in tasks and the essay scores within and across PAAs did not reveal a consistent relationship among the lead-in tasks and essays within and across PAAs.

5. Test order, school, student, gender, SES, and race/ethnicity had significant effects on test scores. Students performed better on the first test than the second test no matter which PAA they took. This test order effect may be due to test motivation: Because the tests had no stakes attached, students might not have been motivated to take these tests, especially the second one.

6. There were some differences in item parameter estimates between the separate and concurrent calibrations, especially for Ban Ads, Mango Street, and Service Learning. As for the EAP theta estimates, both calibrations produced very similar estimates for Service Learning and Invasive Plant Species, and highly correlated but somewhat different estimates for Ban Ads and Mango Street. Both calibrations involved assumptions: Separate calibration assumed equivalent groups taking the four PAAs, which might not be true; and concurrent calibration assumed no change on students' abilities across the two test occasions, which was not the case as the mixed models showed students performed better on the first test than the second test. Therefore, a more deliberate equating design is needed for the Writing PAAs.

7. The IRT results from the concurrent calibration were reported. Most items had reasonable parameter estimates; however, the IRT model did not fit 11 items well, and some item parameter estimates had extreme values. The reliabilities of the EAP theta estimates ranged from .87 to .89.

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.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Cook, R. D., & Weisberg, S. (1999). *Applied regression including computing and graphics*. New York, NY: Wiley.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

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

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

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

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.

Fu, J., Wise, M. D., & Chung, S. (2011). *Dimensionality analysis of CBAL Writing tests*. Manuscript submitted for publication, Educational Testing Service, Princeton, NJ.

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

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75,* 209–227.

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.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16,* 159–176.

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

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

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

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 four PAAs, item scores had various score ranges and thus their score variances varied considerably.

List of Appendices

Page

Appendix A

Item Score Frequency Tables

Table A1

Service Learning: Item Score Frequency

Item score ID	Total		Score																																							
			0		1		1.5		2		3		4		4.5		5		6		7		7.5		8		9		10.5		12		13.5		15		SE		OM		NR	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
W_SERVLEARN01	1,193	100	787	66	406	34	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_SERVLEARN02	1,193	100	988	83	202	17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	0
W_SERVLEARN03	1,193	100	966	81	220	18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7	1
W_SERVLEARN04	1,193	100	692	58	492	41	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9	1
W_SERVLEARN05	1,193	100	902	76	272	23	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	19	2
W_SERVLEARN06	1,193	100	913	77	243	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37	3
W_SERVLEARN07	1,193	100	640	54	497	42	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	56	5
W_SERVLEARN08G	1,193	100	610	51	582	49	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0
W_SERVLEARN08H	1,193	100	848	71	344	29	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0
W_SERVLEARN09G	1,193	100	623	52	568	48	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_SERVLEARN09H	1,193	100	979	82	212	18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_SERVLEARN10G	1,193	100	904	76	286	24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	0
W_SERVLEARN10H	1,193	100	948	79	242	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	0
W_SERVLEARN11G	1,193	100	549	46	640	54	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	0
W_SERVLEARN11H	1,193	100	859	72	330	28	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	0
W_SERVLEARN12G	1,193	100	902	76	284	24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0	-	-	6	1
W_SERVLEARN12H	1,193	100	770	65	416	35	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0	-	-	6	1
W_SERVLEARN13G	1,193	100	427	36	760	64	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	1
W_SERVLEARN13H	1,193	100	506	42	681	57	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	1
W_SERVLEARN14G	1,193	100	411	34	776	65	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	1
W_SERVLEARN14H	1,193	100	621	52	566	47	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	1
W_SERVLEARN15	1,193	100	1020	86	163	14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8	1
W_SERVLEARN16	1,193	100	117	10	93	8	-	-	162	14	186	16	324	27	-	-	126	11	137	11	28	2	-	-	16	1	-	-	-	-	-	-	-	-	-	-	-	-	2	0	2	0
W_SERVLEARN17_I	1,117	100	75	7	-	-	33	3	-	-	93	8	-	-	98	9	-	-	205	18	-	-	186	17	-	-	195	17	103	9	70	6	38	3	8	1	-	-	11	1	2	0
W_SERVLEARN17_III	1,117	100	76	7	-	-	41	4	-	-	152	14	-	-	117	10	-	-	248	22	-	-	184	16	-	-	157	14	79	7	33	3	9	1	8	1	-	-	11	1	2	0

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

Table A2

Invasive Plant Species: Item Score Frequency for Scores 0 to 4.75

Item score ID	Total		Score																																							
			0		.25		.5		.75		1		1.25		1.5		1.75		2		2.25		2.5		2.75		3		3.25		3.5		3.75		4		4.25		4.5		4.75	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
W_INVASIVE_01_01	1,212	100	166	14	16	1	18	1	16	1	27	2	30	2	41	3	49	4	65	5	90	7	75	6	7	6	92	8	81	7	87	7	80	7	74	6	48	4	47	4	11	1
W_INVASIVE_01_02	1,212	100	501	41	-	-	-	-	-	-	652	54	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_03	1,212	100	615	51	-	-	-	-	-	-	520	43	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_04	1,212	100	174	14	-	-	-	-	-	-	957	79	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_05	1,212	100	649	54	-	-	-	-	-	-	470	39	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_06	1,212	100	401	33	-	-	-	-	-	-	711	59	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_07	1,212	100	143	12	-	-	-	-	-	-	957	79	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_08	1,212	100	175	14	-	-	-	-	-	-	907	75	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_09	1,212	100	473	39	-	-	-	-	-	-	598	49	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_10	1,212	100	317	26	-	-	-	-	-	-	731	60	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_11	1,212	100	195	16	-	-	-	-	-	-	841	69	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_12	1,212	100	719	59	-	-	-	-	-	-	295	24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_01_13	1,212	100	477	39	-	-	-	-	-	-	516	43	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_01	1,212	100	432	36	-	-	-	-	-	-	776	64	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_02	1,212	100	512	42	-	-	-	-	-	-	696	57	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_03	1,212	100	737	61	-	-	-	-	-	-	470	39	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_04	1,212	100	528	44	-	-	-	-	-	-	679	56	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_05	1,212	100	977	81	-	-	-	-	-	-	229	19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_06	1,212	100	485	40	-	-	-	-	-	-	721	59	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_07	1,212	100	444	37	-	-	-	-	-	-	761	63	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_08	1,212	100	368	30	-	-	-	-	-	-	837	69	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_09	1,212	100	536	44	-	-	-	-	-	-	668	55	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_10	1,212	100	504	42	-	-	-	-	-	-	699	58	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_11	1,212	100	396	33	-	-	-	-	-	-	808	67	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_12	1,212	100	773	64	-	-	-	-	-	-	432	36	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_13	1,212	100	477	39	-	-	-	-	-	-	727	60	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_14	1,212	100	383	32	-	-	-	-	-	-	821	68	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_15	1,212	100	534	44	-	-	-	-	-	-	671	55	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_02_16	1,212	100	398	33	-	-	-	-	-	-	807	67	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_INVASIVE_03_01	1,212	100	461	38	-	-	-	-	-	-	22	2	-	-	-	-	-	-	77	6	-	-	-	-	-	-	89	7	-	-	-	-	-	-	129	11	-	-	-	-	-	-
W_INVASIVE_04_02_I	1,103	100	158	14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	13	12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	148	13	-	-	-	-	-	-
W_INVASIVE_04_02_III	1,103	100	127	12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	19	18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	327	30	-	-	-	-	-	-

Table A3

Invasive Plant Species: Item Score Frequency for Scores 5 to 20

Item score ID	Score																									
	5		6		7		8		10		12		14		16		18		20		SE		OM		NR	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
W_INVASIVE_01_01	12	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	1	-	-
W_INVASIVE_01_02	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	20	2	39	3
W_INVASIVE_01_03	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	11	1	65	5
W_INVASIVE_01_04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	0	76	6
W_INVASIVE_01_05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	1	82	7
W_INVASIVE_01_06	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	0	96	8
W_INVASIVE_01_07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	1	102	8
W_INVASIVE_01_08	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18	1	112	9
W_INVASIVE_01_09	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	1	131	11
W_INVASIVE_01_10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	19	2	145	12
W_INVASIVE_01_11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	12	1	164	14
W_INVASIVE_01_12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	19	2	178	15
W_INVASIVE_01_13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	20	2	199	16
W_INVASIVE_02_01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	0
W_INVASIVE_02_02	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0	3	0
W_INVASIVE_02_03	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	0
W_INVASIVE_02_04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	0
W_INVASIVE_02_05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0	5	0
W_INVASIVE_02_06	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	1
W_INVASIVE_02_07	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0	6	1
W_INVASIVE_02_08	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7	1
W_INVASIVE_02_09	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0	-	-	7	1
W_INVASIVE_02_10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0	-	-	7	1
W_INVASIVE_02_11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0	-	-	7	1
W_INVASIVE_02_12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7	1
W_INVASIVE_02_13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0	-	-	7	1
W_INVASIVE_02_14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0	-	-	7	1
W_INVASIVE_02_15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7	1
W_INVASIVE_02_16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7	1
W_INVASIVE_03_01	148	12	128	11	95	8	54	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	0	5	0
W_INVASIVE_04_02_I	-	-	112	10	-	-	172	16	141	13	132	12	58	5	25	2	13	1	6	1	-	-	-	-	1	0
W_INVASIVE_04_02_III	-	-	186	17	-	-	114	10	69	6	38	3	23	2	14	1	3	0	3	0	-	-	-	-	1	0

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

Table A4

Ban Ads: Item Score Frequency

Item score ID	Total		Score																																									
			0		.5		1		1.5		2		3		4		4.5		5		6		7		7.5		8		9		10.5		12		13.5		15		SE		OM		NR	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
W_BANADS_01A_01	1,160	100	441	38	-	-	719	62	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_01A_02	1,160	100	266	23	-	-	893	77	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0	-	-	-	-
W_BANADS_01A_03	1,160	100	485	42	-	-	675	58	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_01A_04	1,160	100	691	60	-	-	469	40	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_01A_05	1,160	100	249	21	-	-	910	78	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_BANADS_01B	1,160	100	570	49	164	14	259	22	105	9	57	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	0	1	0
W_BANADS_01C	1,160	100	360	31	138	12	368	32	93	8	138	12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	27	2	36	3
W_BANADS_02AX_A	1,160	100	285	25	-	-	873	75	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_BANADS_02AX_B	1,160	100	245	21	-	-	913	79	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_BANADS_02AX_C	1,160	100	430	37	-	-	728	63	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_BANADS_02AX_D	1,160	100	224	19	-	-	934	81	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_BANADS_02AX_E	1,160	100	246	21	-	-	912	79	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_BANADS_02AX_F	1,160	100	196	17	-	-	962	83	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_BANADS_02AX_G	1,160	100	330	28	-	-	828	71	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_BANADS_02AX_H	1,160	100	148	13	-	-	1,010	87	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_BANADS_02AX_I	1,160	100	285	25	-	-	873	75	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_BANADS_02AX_J	1,160	100	442	38	-	-	716	62	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0
W_BANADS_02BX_A	1,160	100	457	39	-	-	698	60	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0	3	0
W_BANADS_02BX_B	1,160	100	676	58	-	-	478	41	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	0	3	0
W_BANADS_02BX_C	1,160	100	658	57	-	-	496	43	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	0	3	0
W_BANADS_02BX_D	1,160	100	424	37	-	-	730	63	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	0	3	0
W_BANADS_02BX_E	1,160	100	834	72	-	-	319	28	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	0	3	0
W_BANADS_02BX_F	1,160	100	585	50	-	-	568	49	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	0	3	0
W_BANADS_03	1,160	100	460	40	-	-	136	12	-	-	148	13	119	10	153	13	-	-	58	5	44	4	20	2	-	-	17	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	0
W_BANADS_04_I	1,059	100	87	8	-	-	-	-	43	4	-	-	157	15	-	-	142	13	-	-	199	19	-	-	130	12	-	-	127	12	86	8	64	6	14	1	5	0	-	-	2	0	3	0
W_BANADS_04_III	1,059	100	90	9	-	-	-	-	28	3	-	-	216	20	-	-	111	10	-	-	246	23	-	-	123	12	-	-	115	11	54	5	54	5	10	1	7	1	-	-	2	0	3	0

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

Table A5

Mango Street: Item Score Frequency

Item score ID	Total		Score																																	
			0		.5		1		1.5		2		2.5		3		4		5		6		7		8		9		10		OM		NR	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
W_MANGO_01_01	1,213	100	781	64	41	3	391	32	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W_MANGO_01_02	1,213	100	457	38	85	7	663	55	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8	1
W_MANGO_01_03	1,213	100	408	34	30	2	757	62	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18	1
W_MANGO_01_04	1,213	100	354	29	5	0	819	68	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	35	3
W_MANGO_01_05	1,213	100	648	53	63	5	454	37	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	48	4
W_MANGO_02_01	1,213	100	69	6	-	-	48	4	-	-	218	18	-	-	199	16	340	28	165	14	93	8	55	5	24	2	-	-	-	-	-	-	2	0
W_MANGO_03_01	1,213	100	401	33	-	-	809	67	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	0
W_MANGO_03_02	1,213	100	367	30	-	-	842	69	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	0
W_MANGO_03_03	1,213	100	543	45	-	-	666	55	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	0
W_MANGO_03_04	1,213	100	357	29	-	-	852	70	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	0
W_MANGO_03_05	1,213	100	525	43	-	-	683	56	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	0	4	0
W_MANGO_03_06	1,213	100	212	17	143	12	174	14	183	15	278	23	124	10	93	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	0
W_MANGO_04_I	1,116	100	56	5	-	-	41	4	-	-	161	14	-	-	152	14	206	18	148	13	157	14	88	8	63	6	25	2	12	1	4	0	3	0
W_MANGO_04_III	1,116	100	27	2	-	-	12	1	-	-	348	31	-	-	132	12	255	23	118	11	112	10	51	5	26	2	17	2	11	1	4	0	3	0

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

Appendix B

Differential Item Functioning (DIF) Results

Table B1

Service Learning: Item Differential Item Functioning (DIF) Categories

Item score ID	Male (*N* = 441) vs. female (*N* = 441)	White (*N* = 478) vs. Black (*N* = 141)	White (*N* = 478) vs. combination[a] (*N* = 120)	Low SES: no (*N* = 338) vs. yes (*N* = 271)	Number of C DIF (if not 0)
W_SERVLEARN01	A	A	A	A	-
W_SERVLEARN02	A	A	A	A	-
W_SERVLEARN03	A	A	A	A	-
W_SERVLEARN04	A	A	A	A	-
W_SERVLEARN05	A	A	A	A	-
W_SERVLEARN06	A	B-	A	A	-
W_SERVLEARN07	A	A	A	A	-
W_SERVLEARN08H	A	B+	A	A	-
W_SERVLEARN08G	A	A	A	A	-
W_SERVLEARN09H	A	A	A	A	-
W_SERVLEARN09G	A	A	A	A	-
W_SERVLEARN10H	A	A	A	A	-
W_SERVLEARN10G	A	A	A	A	-
W_SERVLEARN11H	A	A	A	A	-
W_SERVLEARN11G	A	B-	A	A	-
W_SERVLEARN12H	A	A	A	A	-
W_SERVLEARN12G	B-	A	A	A	-
W_SERVLEARN13H	A	A	A	A	-
W_SERVLEARN13G	A	A	A	A	-
W_SERVLEARN14H	A	B-	A	A	-
W_SERVLEARN14G	A	A	A	A	-
W_SERVLEARN16	C+	A	A	A	1
W_SERVLEARN17_I	A	A	A	A	-
W_SERVLEARN17_III	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.

[a] Combination of Native American, Asian/Pacific Islander, and Hispanic.

Table B2

Invasive Plant Species: Item Differential Item Functioning (DIF) Categories

Item score ID	Male (*N* = 380) vs. female (*N* = 399)	White (*N* = 399) vs. Black (*N* = 127)	White (*N* = 399) vs. combination[a] (*N* = 101)	Low SES: no (*N* = 273) vs. yes (*N* = 266)	Number of C DIF (if not 0)
W_INVASIVE_01_01	A	A	A	A	-
W_INVASIVE_01_02	A	A	A	A	-
W_INVASIVE_01_03	A	A	A	A	-
W_INVASIVE_01_04	A	A	A	A	-
W_INVASIVE_01_05	A	A	A	A	-
W_INVASIVE_01_06	B-	A	A	A	-
W_INVASIVE_01_07	A	A	A	A	-
W_INVASIVE_01_08	B+	A	A	A	-
W_INVASIVE_01_09	B-	A	A	A	-
W_INVASIVE_01_10	A	A	A	A	-
W_INVASIVE_01_11	A	A	A	A	-
W_INVASIVE_01_13	A	A	A	A	-
W_INVASIVE_02_01	A	A	A	A	-
W_INVASIVE_02_02	A	A	A	A	-
W_INVASIVE_02_03	A	A	A	A	-
W_INVASIVE_02_04	A	A	A	A	-
W_INVASIVE_02_05	A	A	A	A	-
W_INVASIVE_02_06	A	A	A	A	-
W_INVASIVE_02_07	A	A	A	A	-
W_INVASIVE_02_08	A	A	A	B-	-
W_INVASIVE_02_09	A	A	A	A	-
W_INVASIVE_02_10	A	A	A	A	-
W_INVASIVE_02_11	A	A	A	A	-
W_INVASIVE_02_12	A	A	A	A	-
W_INVASIVE_02_13	A	A	A	A	-
W_INVASIVE_02_14	A	A	A	A	-
W_INVASIVE_02_15	A	A	A	A	-
W_INVASIVE_02_16	A	A	A	A	-
W_INVASIVE_03_01	B+	A	A	A	-
W_INVASIVE_04_02_I	A	A	A	A	-
W_INVASIVE_04_02_III	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

[a] Combination of Native American, Asian/Pacific Islander, and Hispanic.

Table B3

Ban Ads: Item Differential Item Functioning (DIF) Categories

Item score ID	Male (*N* = 436) vs. female (*N* = 435)	White (*N* = 406) vs. Black (*N* = 163)	White (*N* = 406) vs. combination [a] (*N* = 111)	Low SES: no (*N* = 271) vs. yes (*N* = 292)	Number of C DIF (if not 0)
W_BANADS_01A_02	A	A	A	A	-
W_BANADS_01A_03	A	A	A	A	-
W_BANADS_01A_04	A	A	A	A	-
W_BANADS_01A_05	A	A	A	A	-
W_BANADS_01B	A	A	A	A	-
W_BANADS_01C	A	C-	A	A	1
W_BANADS_02AX_A	A	A	A	A	-
W_BANADS_02AX_B	A	A	A	A	-
W_BANADS_02AX_C	A	A	A	A	-
W_BANADS_02AX_D	A	A	A	A	-
W_BANADS_02AX_E	A	B-	A	A	-
W_BANADS_02AX_F	A	A	B+	A	-
W_BANADS_02AX_G	A	A	A	A	-
W_BANADS_02AX_H	A	A	A	A	-
W_BANADS_02AX_I	A	A	A	A	-
W_BANADS_02AX_J	A	A	A	A	-
W_BANADS_02BX_A	A	A	A	A	-
W_BANADS_02BX_B	A	A	B-	A	-
W_BANADS_02BX_C	A	A	A	A	-
W_BANADS_02BX_D	A	A	A	B-	-
W_BANADS_02BX_E	A	A	A	A	-
W_BANADS_02BX_F	B-	A	A	A	-
W_BANADS_03	A	A	A	A	-
W_BANADS_04_I	A	A	A	A	-
W_BANADS_04_III	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

[a] Combination of Native American, Asian/Pacific Islander, and Hispanic.

Table B4

Mango Street: Item Differential Item Functioning (DIF) Categories

Item score ID	Male (*N* = 446) vs. female (*N* = 412)	White (*N* = 422) vs. Black (*N* = 166)	White (*N* = 422) vs. combination [a] (*N* = 114)	Low SES: no (*N* = 287) vs. yes (*N* = 297)	Number of C DIF (if not 0)
W_MANGO_01_01	B-	A	A	A	-
W_MANGO_01_02	A	A	A	A	-
W_MANGO_01_03	A	A	A	A	-
W_MANGO_01_04	A	A	A	A	-
W_MANGO_01_05	A	B-	A	A	-
W_MANGO_02_01	A	A	A	A	-
W_MANGO_03_01	A	A	A	A	-
W_MANGO_03_02	A	A	A	A	-
W_MANGO_03_03	A	A	A	A	-
W_MANGO_03_04	A	A	A	A	-
W_MANGO_03_05	A	A	B+	A	-
W_MANGO_03_06	A	B+	A	A	-
W_MANGO_04_I	A	A	A	A	-
W_MANGO_04_III	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

[a] Combination of Native American, Asian/Pacific Islander, and Hispanic.

Appendix C

Cronbach's Alphas on Task and Item Raw Scores

PAA	Alpha	
	Task	Item
Service Learning (PAA 1)	.70	.78
Invasive Plant Species (PAA 2)	.69	.76
Ban Ads (PAA 3)	.69	.76
Mango Street (PAA 4)	.77	.81

Note. PAA = periodic accountability assessment.

Appendix D

Item Response Theory (IRT) Item Fit Statistics

Table D1

Service Learning: Item Fit Statistics

Item	Chi-square	*df*	*p* value	Sig.
W_SERVLEARN01	33.76	35	.53	-
W_SERVLEARN02	48.57	39	.14	-
W_SERVLEARN03	36.90	39	.57	-
W_SERVLEARN04	27.27	36	.85	-
W_SERVLEARN05	46.45	38	.16	-
W_SERVLEARN06	27.69	38	.89	-
W_SERVLEARN07	35.63	32	.30	-
W_SERVLEARN08H	34.84	38	.62	-
W_SERVLEARN08G	29.21	30	.51	-
W_SERVLEARN09H	579.45	38	.00	**
W_SERVLEARN09G	34.99	32	.33	-
W_SERVLEARN10H	470.22	38	.00	**
W_SERVLEARN10G	43.59	38	.25	-
W_SERVLEARN11H	28.83	39	.88	-
W_SERVLEARN11G	39.37	29	.10	-
W_SERVLEARN12H	34.45	36	.54	-
W_SERVLEARN12G	43.56	38	.25	-
W_SERVLEARN13H	46.81	29	.02	*
W_SERVLEARN13G	21.54	26	.71	-
W_SERVLEARN14H	25.61	31	.74	-
W_SERVLEARN14G	36.52	29	.16	-
W_SERVLEARN16	150.19	133	.15	-
W_SERVLEARN17_I	123.35	126	.55	-
W_SERVLEARN17_III	118.26	123	.60	-

Note. Sig. = significance.

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

Table D2

Invasive Plant Species: Item Fit Statistics

Item	Chi-square	*df*	*p* value	Sig.
W_INVASIVE_01_01	149.45	150	.50	-
W_INVASIVE_01_02	34.95	32	.33	-
W_INVASIVE_01_03	43.60	35	.15	-
W_INVASIVE_01_04	11.48	13	.57	-
W_INVASIVE_01_05	26.36	36	.88	-
W_INVASIVE_01_06	23.31	32	.87	-
W_INVASIVE_01_07	6.48	11	.84	-
W_INVASIVE_01_08	22.30	14	.07	-
W_INVASIVE_01_09	42.43	32	.10	-
W_INVASIVE_01_10	23.46	31	.83	-
W_INVASIVE_01_11	24.28	18	.15	-
W_INVASIVE_01_13	43.42	34	.13	-
W_INVASIVE_02_01	28.84	23	.19	-
W_INVASIVE_02_02	31.93	31	.42	-
W_INVASIVE_02_03	32.38	37	.69	-
W_INVASIVE_02_04	33.49	34	.49	-
W_INVASIVE_02_05	519.25	36	.00	**
W_INVASIVE_02_06	28.66	29	.48	-
W_INVASIVE_02_07	44.52	23	.01	**
W_INVASIVE_02_08	31.83	21	.06	-
W_INVASIVE_02_09	36.59	29	.16	-
W_INVASIVE_02_10	28.14	28	.46	-
W_INVASIVE_02_11	23.09	25	.57	-
W_INVASIVE_02_12	39.92	38	.39	-
W_INVASIVE_02_13	27.14	24	.30	-
W_INVASIVE_02_14	25.73	23	.31	-
W_INVASIVE_02_15	37.22	28	.11	-
W_INVASIVE_02_16	35.65	23	.05	*
W_INVASIVE_03_01	135.58	121	.17	-
W_INVASIVE_04_02_I	148.28	132	.16	-
W_INVASIVE_04_02_III	121.91	122	.49	-

Note. Sig. = significance.

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

Table D3

Ban Ads: Item Fit Statistics

Item	Chi-square	*df*	*p* value	Sig.
W_BANADS_01A_02	20.70	24	.66	-
W_BANADS_01A_03	61.35	34	.00	*
W_BANADS_01A_04	45.00	38	.20	-
W_BANADS_01A_05	16.54	23	.83	-
W_BANADS_01B	93.58	84	.22	-
W_BANADS_01C	72.54	76	.59	-
W_BANADS_02AX_A	26.48	25	.38	-
W_BANADS_02AX_B	43.22	23	.01	*
W_BANADS_02AX_C	39.70	34	.23	-
W_BANADS_02AX_D	27.02	22	.21	-
W_BANADS_02AX_E	17.07	22	.76	-
W_BANADS_02AX_F	24.87	20	.21	-
W_BANADS_02AX_G	35.87	29	.18	-
W_BANADS_02AX_H	4.22	11	.96	-
W_BANADS_02AX_I	20.89	24	.65	-
W_BANADS_02AX_J	37.57	32	.23	-
W_BANADS_02BX_A	21.73	35	.96	-
W_BANADS_02BX_B	40.11	38	.38	-
W_BANADS_02BX_C	35.98	36	.47	-
W_BANADS_02BX_D	15.82	31	.99	-
W_BANADS_02BX_E	35.69	38	.58	-
W_BANADS_02BX_F	43.11	34	.14	-
W_BANADS_03	157.30	113	.00	*
W_BANADS_04_I	146.55	123	.07	-
W_BANADS_04_III	111.21	116	.61	-

Note. Sig. = significance.

* *p* < .01.

Table D4

Mango Street: Item Fit Statistics

Item	Chi-square	*df*	*p* value	Sig.
W_MANGO_01_01	37.12	36	.42	-
W_MANGO_01_02	38.33	31	.17	-
W_MANGO_01_03	30.66	32	.53	-
W_MANGO_01_04	21.99	23	.52	-
W_MANGO_01_05	33.01	31	.37	-
W_MANGO_02_01	157.68	125	.03	*
W_MANGO_03_01	33.77	29	.25	-
W_MANGO_03_02	24.91	23	.36	-
W_MANGO_03_03	34.19	30	.27	-
W_MANGO_03_04	32.92	24	.11	-
W_MANGO_03_05	34.77	33	.38	-
W_MANGO_03_06	164.18	127	.02	*
W_MANGO_04_I	149.85	132	.14	-
W_MANGO_04_III	130.51	108	.07	-

Note. Sig. = significance.

* *p* < .05.