Listening. Learning. Leading.®

Research Report

ETS RR-13-10

Dimensionality Analysis of *CBAL*™ Writing Tests

Jianbin Fu

Seunghee Chung

Maxwell Wise

May 2013

ETS Research Report 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 Report 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 Report 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.

Dimensionality Analysis of *CBAL*$^{\text{TM}}$ Writing Tests

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

May 2013

Action Editor: James Carlson

Reviewers: Richard Schwartz and Shameem Gaj

Abstract

The Cognitively Based Assessment of, for, and as Learning (*CBAL*^{TM}) research initiative is aimed at developing an innovative approach to K–12 assessment based on cognitive competency models. Because the choice of scoring and equating approaches depends on test dimensionality, the dimensional structure of CBAL tests must be understood. The purpose of this study was to investigate the dimensionality of 4 CBAL Grade 8 writing tests. Each of the 4 tests focused on one of the following writing genres: persuasive letter/memorandum, research-based report (pamphlet), persuasive essay, and literature essay. Dimensionality was investigated using exploratory and confirmatory factor analyses. The results show multidimensionality and support subscore structures and bifactor task models for all 4 tests.

Key words: CBAL, writing assessment, test dimensionality, factor analysis

Acknowledgments

Thanks due to Randy Bennett, Jim Carlson, Rick Morgan, Yanmei Li, Brad Moulder, Christine Mills, and Frank Rijmen for their helpful suggestions and edits on earlier versions of this paper. We are also grateful to Kim Fryer for her editorial assistance.

Table of Contents

List of Tables

The Cognitively Based Assessment *of*, *for,* and *as* Learning (*CBAL*$^{\mathrm{TM}}$) research initiative is intended to develop an innovative K–12 assessment model that measures what students have learned (of learning), provides timely feedback for planning instruction (for learning), and is a worthwhile educational experience in and of itself (as learning; Bennett, 2010). The CBAL system model includes four components: conceptual, accountability assessment, formative assessment, and professional support. The model and its prototype assessments are based on the integration of cognitive research and curriculum standards. This paper deals with dimensionality of summative tests in the accountability assessment component.

In the accountability assessment component, multiple summative tests, referred to as periodic accountability assessments (PAAs), are administrated throughout a school year. Scores for each administration are reported to provide prompt interim information for teachers to use for directing formative follow-up. In addition, aggregate scores from these multiple PAAs are used for accountability purposes because such scores should be, in principle, more reliable than the individual PAA scores. All tests are developed based on underlying cognitive competency models and curriculum standards that describe skills that students need to learn, the skills' relationships to one another, and how these skills might be ordered as learning progressions for purposes of guiding assessment and instruction (Deane, 2011; Graf, 2009; O'Reilly & Sheehan, 2009a, 2009b). All tests are administered online and include innovative technology-enhanced items organized under a common scenario, which is different from standard assessments. The tasks comprising a scenario attempt to gauge higher order skills as well as more fundamental competencies. Because CBAL summative assessments follow a distributed model, more testing time is available in the aggregate so that assessment tasks can be more complex and more integrative than might be possible in a single annual administration.

The purpose of this study was to investigate the dimensionality—using exploratory and confirmatory factor analyses—of four CBAL Grade 8 writing PAAs administered in fall 2009. Understanding the dimensional structure within and across PAAs is important for the choice of appropriate scoring and equating approaches. Obviously, different equating and scoring methods (including score aggregation methods) should be used for unidimensional and multidimensional tests. For multidimensional tests, multidimensional models should be used for test equating and scoring, and scores on each dimension as well as overall scores aggregated from all dimensions may be reported separately. On the other hand, test equating and scoring for unidimensional tests are simpler, and only scores on the whole test need to be reported. In addition, dimensionality analysis can provide validity evidence with respect to the underlying competency model.

The subsequent sections are organized as follows: The content design of the four writing PAAs are presented first, including the underlying CBAL writing competency model and the content structure of each test. Second, the factor analysis method for assessing test dimensionality is introduced. Third, the fall 2009 CBAL writing test administration and data structure and the analysis design for assessing dimensionality are described, which is followed by the dimensionality assessment results. Finally, the results are summarized and discussed.

Content Design of CBAL Writing Tests

All CBAL writing tests were developed based on the writing competency model (Deane, 2011). A key theme of this model is the integration of writing and literacy skills. The basic idea is that writing and literacy skills are inseparable: Those skills for effective reading and critical thinking are also essential for effective writing. Therefore, this competency model is actually a literacy model and contains three basic modes of thought during literacy activities: interpretation, expression, and reflection and deliberation. Each thought then has five levels of cognitive representation that represent five basic types of information that enter into literate thought. These cognitive representations are the social model, the conceptual model, the textual model, the verbal model, and the lexical/orthographic model. Skills classified under this three-by-five structure cover those skills needed for reading, writing, and critical thinking. In order to accomplish an activity or goal in literacy, and especially in writing, multiple skills at different levels must be functioned concurrently or in rapid succession in a coordinated way; individual skills in isolation are usually meaningless for literacy activities. Therefore, the competency model emphasizes how these literacy skills are coordinated and interwoven to achieve specific purposes.

A direct implication for writing instruction and assessment is that writing activities should be organized in terms of genres and strategies, that is, types of writing that serve specific communicative goals. Different writing genres require different sets of literacy skills and strategies to control and organize component skills. Deane (2011) and Deane et al. (2009) provided a list of writing genres and critical thinking strategy families.

Deane, Fowles, Baldwin, and Persky (2011) and Deane et al. (2009) identified four writing genres consistent with the writing curriculum in Grade 8 (persuasive letter/memorandum, research-based report [pamphlet], persuasive essay, and literature essay) and developed four writing PAAs targeted for the four writing genres, respectively. Each PAA embodies a realistic scenario, and the items associated with that scenario are organized under four tasks. Table 1 lists the scenario, genre, and critical thinking strategy for each PAA. The first three tasks are the lead-in tasks providing a scaffold to build reading and critical thinking skills necessary for an effective writer of the genre specific to that PAA, and the fourth task is the culminating task of writing that essay. The first three tasks contain selected-response (SR) items, which have dichotomous scores, and constructed-response (CR) items, which have dichotomous or polytomous scores.[1] Essays are assigned two scores on (a) content and critical thinking and (b) organization, development, and phrasing. The first criterion is to assess the control of critical thinking skills needed to accomplish the writing task so that students can develop and explain their ideas effectively and accurately. And the second criterion is about the control of word choice, sentence structure, and written conventions that students employ in writing up their ideas.

Table 1

Genres and Critical Thinking Strategies of Grade 8 Writing PAAs

PAA no.	PAA scenario	Genre	Rhetorical purpose of the genre	Critical thinking strategies developed[a]
1	Classroom Service Learning Projects	Persuasive letter/memorandum	Make a recommendation based on explicit criteria	Standard setting
2	Invasive Plant Species	Research-based report (pamphlet)	Present information on a topic in accessible, easy-to-digest form	Guiding questions
3	Whether to Ban Ads to Children Under Twelve	Persuasive essay	Make a claim and support it effectively	Argument building
4	The House on Mango Street	Literary essay	Justify an interpretation of a text	Close reading

Note. This table is adapted from Table 3 in *Progress on Designing the CBAL Summative Writing Assessment: Design Principles and Results,* by P. Deane et al., 2009, unpublished manuscript. CR = constructed response, PAA = periodic accountability assessment, SR = selected response.
[a]See Deane et al. (2009, p. 20) for the classification and description of critical thinking strategies.

Tables 2 through 5 provide task and subscore information for the four writing PAAs, respectively, including task and subscore description, numbers of SR and CR items, and maximum score points for each task and subscore. For all PAAs, the tasks and subscores totally overlap except for Task 1 in PAA 2 (Invasive Plant Species),and Tasks 1 and 2 in PAA3 (Ban Ads), where each task was separated into two subscores. For a detailed description of the test designs of the four writing PAAs, readers are referred to Deane et al. (2011) and Deane et al. (2009).

Unlike traditional writing tests that contain only essay writing, these tests also include lead-in tasks that try to guide students to the basic skills and integrative strategy for a certain writing genre. In this way, CBAL writing tests provide incentives for teachers to focus their teaching on these skills and strategy. The literacy skills measured in the four PAAs for the eighth grade do not follow an apparent learning progression and are actually parallel. Thus, these PAAs are self-contained and can be administered in any order within a school year to match curricular requirements.

Table 2

Test Structure: PAA 1 (Service Learning)

Task (subscore) no.	Task (subscore)	No. of SR items	No. of CR items	Max score points
1	Give feedback on a peer’s letter with respect to a rubric	7	0	7 (1 per item)
2	Evaluate and compare two activities by deciding how well they fit prespecified goals	14[a]	0	14 (1 per item)
3	Briefly explain findings to another student	0	1	8
4	Write a persuasive memorandum to a decision-maker–scored	0	2	30 (15 per item)
Total		21	3	59

Note. This table is adapted from Table 4 in *Progress on Designing the CBAL Summative Writing Assessment: Design Principles and Results,* by P. Deane et al., 2009, unpublished manuscript. CR = constructed response, PAA = periodic accountability assessment, SR = selected response.

[a]One item (SERVLEARN15) in Task 2 is excluded from this table and the subsequent analyses because of its zero item-total correlation.

Table 3

Test Structure: PAA 2 (Invasive Plant Species)

Task no.	Subscore no.	Task (subscore)	No. of SR items	No. of CR items	Max score points
1		Gather and evaluate information for a pamphlet	11	1	16
	1	Read an article and generate guiding questions	0	1	5
	2	Evaluate sources for research	11[a]	0	11 (1 per item)
2	3	Use guiding questions to organize information for the pamphlet	16	0	16 (1 per item)
3	4	Review and revise sections of the pamphlet	0	1	8
4	5	Write headings for two sections of the pamphlet; write two sections of the pamphlet	0	2	40 (20 per item)
Total			27	4	80

Note. This table is adapted from Table 5 in *Progress on Designing the CBAL Summative Writing Assessment: Design Principles and Results*, by P. Deane et al., 2009, unpublished manuscript. CR = constructed response, PAA = periodic accountability assessment, SR = selected response.

[a]One item (INVASIVE_01_12) in Subscore 2 is excluded from this table and the subsequent analyses because of its negative item-total correlation.

Table 4

Test Structure: PAA 3 (Ban Ads)

Task no.	Subscore no.	Task (subscore)	No. of SR items	No. of CR items	Max score points
1		Read and summarize arguments: select feedback for Anna’s summary	4	2	8
	1	Write two to three sentences summarizing an easy article	4[a]	0	4 (1 per item)
	2	Write two to three sentences summarizing a more complex article	0	2	4 (2 per item)
2		Analyze arguments	16	0	16
	3	Consider arguments for/against: classify reasons ban or allow	10	0	10 (1 per item)

Task no.	Subscore no.	Task (subscore)	No. of SR items	No. of CR items	Max score points
	4	Consider evidence from articles: supports, weakens, or irrelevant	6	0	6 (1 per item)
3	5	Help classmates critique the arguments in a letter to the editor	0	1	8
4	6	Write a persuasive essay	0	2	30 (15 per item)
Total			20	5	62

Note. This table is adapted from Table 6 in *Progress on Designing the CBAL Summative Writing Assessment: Design Principles and Results*, by P. Deane et al., 2009, unpublished manuscript. CR = constructed response, PAA = periodic accountability assessment, SR = selected response.

[a] One item (BANADS_01A_01) in Subscore 1 is excluded from this table and the subsequent analyses because of its negative item-total correlation.

Table 5

Test Structure: PAA 4 (Mango Street)

Task (subscore) no.	Task (subscore)	No. of SR items	No. of CR items	Max score points
1	Support interpretive statements about a story with details from the text	5	0	5 (1 per item)
2	Explain whether a character's attitude changes (respond to conflicting interpretations)	0	1	8
3	Help another student interpret the story (plausibility of explanations given the text)	5	1	8 (1 per SR item; 3 per CR item)
4	Write an essay about the story	0	2	20 (10 per item)
Total		10	4	41

Note. This table is adapted from Table 7 in *Progress on Designing the CBAL Summative Writing Assessment: Design Principles and Results*, by P. Deane et al., 2009, unpublished manuscript. CR = constructed response, PAA = periodic accountability assessment, SR = selected response.

Exploratory and Confirmatory Factor Analyses

Test dimensionality is an important issue in education measurement, and many psychometric procedures have been proposed to assess test structure (De Champlain & Gessaroli, 1998; Hattie, 1984, 1985; Zwick, 1987). Levy and Svetina (2010) classified test dimensionality assessment methods and computer programs in terms of their confirmatory or exploratory nature; parametric or nonparametric assumptions; and applicability to dichotomous, polytomous, and missing data. Among them, factor analysis (including parallel analysis [PA]) is widely used in practice (Levy & Svetina, 2010; Tate, 2003).

Factor analysis is a parametric modeling approach used to find a small number of factors that represents the underlying structure of a large number of correlated variables. Factor analysis can be classified as exploratory or confirmatory based on whether there is a prior hypothesis as to the factor structure: exploratory factor analysis (EFA) is used to explore the factor structure of a dataset without a prior hypothesis, while confirmatory factor analysis (CFA) is used to verify a hypothesized structure that is based on theory.

For the factor analysis of categorical data, polychoric correlations are preferred to Pearson correlations. The reasons is that polychoric correlations, which assume that two ordinal variables have an underlying bivariate normal distribution, more closely reflect the actual relationships between the two ordinal variables, while Pearson correlations tend to underestimate those relationships (Garson, 2011).

To determine the optimal number of factors in EFA, scree plots and PA have been broadly utilized (e.g., Zwick & Velicer, 1986). However, the criteria used in scree plots are somewhat subjective, while PA is considered as a more accurate method for determining optimal numbers of factors in the exploratory mode (O’Conner, 2000). Therefore, PA was employed in the current study. In PA, simulated datasets with the same number of variables and examinees as those in the real dataset were first generated based on the standard multivariate normal distribution with no correlations among variables. Then, eigenvalues were calculated for each simulated dataset and the 95th percentile cut point for the eigenvalue of each factor was determined across all the simulated datasets. The generally accepted rule is that the number of factors to extract in a real dataset is equal to the maximum number of factors where the eigenvalue of the real dataset is in the 95th percentile of eigenvalues from the simulated datasets (O’Conner, 2000).

Method

Participants

Schools were recruited nationwide to participate in the CBAL writing summative field test administered in fall 2009. The schools that volunteered were compensated monetarily in return for the participation of their students. The resulting convenience sample included 2,580 Grade 8 students from 21 schools in 12 states. See Table 6 for the sample distribution by state, English language learner (ELL) status, gender, race and socioeconomic status (SES).

Procedure

Students took CBAL PAAs on their school computers linked to ETS's testing server via the Internet. Each student took two (of four) PAAs that were randomly assigned at the school level to 1 of 12 possible test sequences. Ninety-three percent of students completed the second PAA within 1 month of taking the first PAA.

Data Analysis

Because most students took two PAAs within 1 month, equivalent groups were assumed between the two test occasions, and thus test data from the two test occasions were combined for the dimensionality analyses. The sample sizes used in the factor analyses for the four PAAs were 1,057, 912, 1,025, and 1,067 examinees, respectively. Note that, for all analyses, an omitted response was treated as zero while not reached was treated as missing. For any PAA, students with any missing value were excluded from the factor analyses on the test form (i.e., listwise deletion). The sample sizes for the six PAA pairs were between 200 and 375, while the number of items for the PAA pairs ranged from 38 to 56. With so many items and such small sample sizes, the dimensionality analysis results on PAA pairs were not stable and thus were not reported here.[2]

Both EFA and CFA were conducted on an interitem polychoric correlation matrix[3] using the computer program LISREL (Joreskog & Sorbom, 1996a), and PAs were carried out using an SAS macro developed in O'Conner (2000). For CFA, the following hypothesized factor models related to item type (SR vs. CR items), subscore, and task, as well as bifactor model, were compared:

- 1 factor
- 2 item-type factors: SR versus CR items

- 4 task factors
- Bifactor task model: one general factor (it has loadings on all items) and task-specific factors (each has loadings only on the items in the associated task)
- 4-6 subscore factors

Table 6

Test Sample Distribution by Demographic

Demographic	*N*	Percentage
State		
Alabama	122	5
Arizona	573	22
Arkansas	290	11
California	64	2
Florida	41	2
Georgia	201	8
Kentucky	61	2
Louisiana	110	4
Massachusetts	106	4
Mississippi	99	4
Ohio	192	7
Texas	204	8
Unreported	517	20
ELL status		
No	1,077	42
Yes	52	2
Unreported	1,451	56
Gender		
Male	1,051	41
Female	1,010	39
Unreported	519	20
Race		
African American	374	14
Asian/Pacific Islander	58	2
Hispanic	196	8
Native American	10	0
White	1,032	40
Unreported	910	35
Low SES status		
No	701	27
Yes	705	27
Unreported	1,174	46

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

These structures were evaluated in this study for three reasons. First, a lead-in task/subscore measures a specific reading or critical thinking skill, and the essay task/subscore examines the ability of integrating and coordinating relevant literacy skills for writing with a specific purpose. Thus, it is important to check if the data support the skill structures. Second, there have been numerous debates and studies on the question of construct equivalence of multiple-choice (MC) items and CR items since the 1980s. The literature on the question whether MC and CR items represent different dimensions is equivocal (see a brief review in Rauch & Hartig, 2010, and the references therein). Therefore, whether item format affects test dimensions should be checked. Third, from the perspectives of both validity and scoring, it is meaningful to check whether a general factor related to all items representing overall writing ability specific to a writing genre, and factors specific to tasks, can be extracted from the data. A bifactor model is particularly useful to generate overall scores for a multidimensional test.

For EFA, the emphasis was to determine whether a PAA deviated from unidimensionality, but finding an optimal factor pattern was not the interest here because we already had hypothesized factor structures.

Results

Subscore and Total Score Summary and Correlations

Tables 7–10 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 contained 1 to 16 items (see Tables 2–5), and their reliabilities (standardized Cronbach’s alpha) ranged from .24 to .92. For each PAA, the subscore computed from the essay had the highest reliability. As mentioned above, each essay subscore contained two scores measuring different aspects of the same essay. The correlations between subscores and total scores ranged from .43 to .93. The inter-subscore correlations were low to intermediate, between .18 and .64, indicating possibly well-defined different dimensions existing in each PAA.

Table 7

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

Score[a]	*N*	Mean	SD	Standardized alpha[b]	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 1, S2 = Subscore 2, S3 = Subscore 3, S4 = Subscore 4.

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

Table 8

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

Score[a]	*N*	Mean	SD	Standardized alpha[b]	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 1, S2 = Subscore 2, S3 = Subscore 3, S4 = Subscore 4, S5 = Subscore 5.

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

Table 9

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

Score[a]	*N*	Mean	SD	Standardized alpha[b]	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 1, S2 = Subscore 2, S3 = Subscore 3, S4 = Subscore 4, S5 = Subscore 5, S6 = Subscore 6.

[a]See Table 4 for subscore information. [b]Reliability was not calculated for a subscore with one item.

Table 10

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

Score[a]	*N*	Mean	SD	Standardized alpha[b]	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 1, S2 = Subscore 2, S3 = Subscore 3, S4 = Subscore 4.

[a]See Table 5 for subscore information. [b]Reliability was not calculated for a subscore with one item.

Exploratory Factor Analysis

Tables 11–14 show the PA results for the four PAAs; these tables contain for each factor the 95% cut point of eigenvalues from the 10,000 generated datasets, as well as eigenvalues, and proportion and cumulative proportion of variance explained from the real data. The extracted factors suggested by PA were those factors whose eigenvalues were larger than the 95% cut points of random eigenvalues. Based on the PA, the numbers of optimal factors for the four PAAs were 3, 5, 5, and 1, respectively.

Table 11

Parallel Analysis: PAA1 (Service Learning)

Factor number	Simulated 95th percentile eigenvalue cut point	Real data eigenvalue	Real data proportion of variance	Real data cumulative proportion of variance
1	1.32	8.62	.36	.36
2	1.27	1.76	.07	.43
3[a]	1.23	1.25	.05	.48
4	1.20	1.15	.05	.53
5	1.17	1.14	.05	.58
6	1.15	1.03	.04	.62
7	1.12	.90	.04	.66
8	1.10	.83	.03	.69
9	1.08	.80	.03	.73
10	1.06	.78	.03	.76
11	1.04	.76	.03	.79
12	1.02	.60	.03	.82
13	1.00	.59	.02	.84
14	.98	.56	.02	.86
15	.96	.51	.02	.89
16	.94	.42	.02	.90
17	.93	.40	.02	.92
18	.91	.37	.02	.94
19	.89	.33	.01	.95
20	.87	.32	.01	.96
21	.85	.29	.01	.98
22	.83	.25	.01	.99
23	.80	.22	.01	.99
24	.78	.14	.01	1.00

Note. PAA = periodic accountability assessment.

[a] The shaded row for Factor Number 3 shows the number of factors suggested by the parallel analysis.

Table 12

Parallel Analysis: PAA 2 (Invasive Plant Species)

Factor number	Simulated 95th percentile eigenvalue cut point	Real data eigenvalue	Real data proportion of variance	Real data cumulative proportion of variance
1	1.40	10.31	.33	.33
2	1.34	1.79	.06	.39
3	1.30	1.56	.05	.44
4	1.27	1.46	.05	.49
5[a]	1.24	1.30	.04	.53
6	1.22	1.21	.04	.57
7	1.19	1.09	.04	.60
8	1.17	1.04	.03	.64
9	1.15	.96	.03	.67
10	1.12	.90	.03	.70
11	1.10	.80	.03	.72
12	1.08	.78	.03	.75
13	1.06	.73	.02	.77
14	1.04	.67	.02	.79
15	1.02	.66	.02	.82
16	1.01	.58	.02	.83
17	.99	.58	.02	.85
18	.97	.54	.02	.87
19	.95	.52	.02	.89
20	.93	.46	.02	.90
21	.92	.44	.01	.92
22	.90	.39	.01	.93
23	.88	.38	.01	.94
24	.86	.35	.01	.95
25	.85	.28	.01	.96
26	.83	.26	.01	.97
27	.81	.24	.01	.98
28	.79	.22	.01	.98
29	.77	.20	.01	.99
30	.75	.16	.01	1.00
31	.72	.12	.00	1.00

Note. PAA = periodic accountability assessment.

[a] The shaded row for Factor Number 5 shows the number of factors suggested by the parallel analysis.

Table 13

Parallel Analysis: PAA 3 (Ban Ads)

Factor number	Simulated 95th percentile eigenvalue cut point	Real data eigenvalue	Real data proportion of variance	Real data cumulative proportion of variance
1	1.34	6.61	.26	.26
2	1.28	1.92	.08	.34
3	1.24	1.58	.06	.40
4	1.21	1.34	.05	.46
5[a]	1.18	1.22	.05	.51
6	1.16	1.12	.04	.55
7	1.13	1.07	.04	.59
8	1.11	1.01	.04	.64
9	1.09	.99	.04	.68
10	1.07	.88	.04	.71
11	1.05	.81	.03	.74
12	1.03	.76	.03	.77
13	1.01	.72	.03	.80
14	.99	.69	.03	.83
15	.97	.61	.02	.85
16	.95	.55	.02	.88
17	.93	.53	.02	.90
18	.92	.48	.02	.92
19	.90	.45	.02	.93
20	.88	.43	.02	.95
21	.86	.33	.01	.96
22	.84	.33	.01	.98
23	.82	.26	.01	.99
24	.80	.18	.01	1.00
25	.77	.11	.00	1.00

Note. PAA = periodic accountability assessment.

[a] The shaded row for Factor Number 5 shows the number of factors suggested by the parallel analysis.

Table 14

Parallel Analysis: PAA 4 (Mango Street)

Factor number	Simulated 95th percentile eigenvalue cut point	Real data eigenvalue	Real data proportion of variance	Real data cumulative proportion of variance
1[a]	1.24	6.37	.46	.46
2	1.18	1.00	.07	.53
3	1.14	.99	.07	.60
4	1.11	.82	.06	.66
5	1.08	.77	.05	.71
6	1.05	.71	.05	.76
7	1.03	.61	.04	.80
8	1.00	.53	.04	.84
9	.98	.49	.03	.88
10	.96	.47	.03	.91
11	.93	.42	.03	.94
12	.91	.35	.03	.97
13	.88	.32	.02	.99
14	.85	.15	.01	1.00

Note. PAA = periodic accountability assessment.

[a] The shaded row for Factor Number 1 shows the number of factors suggested by the parallel analysis.

EFA models with maximum likelihood estimation were analyzed in LISREL to determine the number of factors extracted for each PAA using chi-squared tests. For each PAA, the chi-squared tests showed multiple factors up to that for which the model failed to converge, that is, 11, 10, 6, and 5 factors for the four PAAs, respectively.

Confirmatory Factor Analysis

For analyzing a correlation matrix using CFA, the recommended estimation method is the generally weighted least-squares (WLS; Joreskog & Sorbom, 1996a, pp. 21–23; Joreskog, Sorbom, du Toit, & du Toit, 2000, pp. 209–214). To assess model fit, Kline (1998, pp. 130–131) recommended reporting at least the following fit indexes: chi-squared test (chi-square statistic, degrees of freedom, and probability), goodness of fit index (GFI), normed fit index (NFI), nonnormed fit index (NNFI), and standardized root mean square residual (SRMR). The chi-squared test is very sensitive to sample size, and with a large sample size, the test is significant even if the model fits the data well. In place of the chi-squared test, for large sample sizes the ratio of the chi-squared statistic to degree of freedom is recommended, and values smaller than 3 are considered favorable.

To indicate adequate fit, GFI, NFI, and NNFI should be greater than .90, and SRMR should be smaller than .10. In addition, the root mean square error of approximation (RMSEA) is commonly reported, and a favorable value is .5 or less (Bollen & Long, 1993, p. 144). In the following, the CFA results for the four PAAs were presented. Note that in PAA 1 (Service Learning) and PAA 4 (Mango Street) subscores and tasks totally overlapped; thus, for these two PAAs, subscores could be also referred to as tasks.

Tables 15–18 list the factor loadings, model fit statistics mentioned above, factor correlations, and chi-squared tests for model comparisons for those confirmatory models described previously for PAA 1 (Service Learning). SERVLEARN09H had the smallest factor loadings in all models, and except for the two-factor item type model, its factor loadings were not significant, which indicates that this item measured something different from the intended skill. In the bifactor model, nine items in Subscore 2 had nonsignificant factor loadings on the subscore specific factor, and thus their loadings on the subscore specific factor were removed from the bifactor model. This indicates that for these nine items, once the general factor was taken into account, there was no significant variance explained by the subscore specific factor. The fit statistics show that all the models fitted reasonably well, although the χ^2 / df was rather large (3.51) for the one-factor model, and SRMRs were rather high for both the one-factor and two-factor models (.13 and .11, respectively). However, the fit statistics did show some gradual improvement in the following models in order: the one-factor model, the two-factor model, the four-factor subscore model, and the bifactor model. The one-factor and the two-factor models, and the one-factor and the bifactor models were nested models, and so were the two-factor item type and the four-factor subscore models because the SR comprised Subscores 1 and 2 and the CR comprised Subscores 3 and 4. The chi-squared tests for model comparisons were significant for all the nested models. The two item-type factors had an estimated correlation of .80; however, because the item types overlapped with subscore, this correlation might be confounded with the subscore correlations. The estimated subscore factor correlations were between .60 and .89 with the highest correlation between Subscores 1 and 2.

Table 15

Confirmatory Factor Analysis Results for PAA 1 (Service Learning): Factor Loading

Score ID	1 factor	2 factors (item type)		4 factors (subscore/task)[a]				Bifactor (general+ subscore/task)[a,b]			
	General	SR	CR	S(T) 1	S(T) 2	S(T) 3	S(T) 4	General	S(T) 1	S(T) 2	S(T) 4
SERVLEARN01	.71	.72		.74				.65	.27		
SERVLEARN02	.51	.55		.63				.32	.73		
SERVLEARN03	.37	.44		.54				.28	.70		
SERVLEARN04	.51	.53		.55				.48	.29		
SERVLEARN05	.72	.74		.76				.65	.31		
SERVLEARN06	.58	.56		.56				.55	-.11		
SERVLEARN07	.74	.76		.78				.73	.10		
SERVLEARN08H	.44	.46			.45			.44			
SERVLEARN08G	.86	.86			.84			.85			
SERVLEARN09H	.05[c]	.08			.07[c]			.05 [c]			
SERVLEARN09G	.67	.67			.66			.68			
SERVLEARN10H	.09	.14			.14			.11			
SERVLEARN10G	.39	.39			.38			.42			
SERVLEARN11H	.15	.15			.14			.12			
SERVLEARN11G	.86	.87			.87			.84		.20	
SERVLEARN12H	.69	.70			.69			.57		.41	
SERVLEARN12G	.77	.80			.80			.65		.66	
SERVLEARN13H	.80	.79			.79			.78		.20	
SERVLEARN13G	.78	.80			.81			.74		.26	
SERVLEARN14H	.88	.87			.88			.84		.20	
SERVLEARN14G	.69	.70			.69			.67			
SERVLEARN16	.80		.79			1.00		.82			
SERVLEARN17_I	.93		.94				.95	.81			.48
SERVLEARN17_III	.91		.91				.90	.78			.48

Note. CR = constructed response, PAA = periodic accountability assessment, SR = selected response, S(T)1 = Subscore (Task) 1, S(T)2 = Subscore (Task) 2, S(T)3 = Subscore (Task) 3, S(T)4 = Subscore (Task) 4.

[a]See Table 2 for task and subscore information. [b]For a bifactor model, factor independence was assumed. The nonsignificant loadings were removed, and the subscore (task) with only one item was not treated as a specific factor in the bifactor model. [c]Nonsignificant factor loadings at .05 level (-1.96 < *t* < 1.96).

Table 16

Confirmatory Factor Analysis Results for PAA 1 (Service Learning): Fit Statistics

Fit statistic	1 factor	2 factors (item type)	4 factors (subscore/task)[a]	Bifactor (general + subscore/task)[a,b]
χ^2	885.56	712.20	616.40	478.73
df	252	251	247	238.00
p	.00	.00	.00	.00
χ^2 / *df*	3.51	2.84	2.50	2.01
RMSEA	.05	.04	.04	.03
NFI	.91	.93	.94	.95
NNFI	.93	.95	.96	.97
GFI	.98	.98	.98	.99
SRMR	.13	.11	.10	.08

Note. GFI = goodness of fit index, NFI = normed fit index, NNFI = nonnormed fit index, PAA = periodic accountability assessment, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual.

[a]See Table 2 for task and subscore information. [b]For a bifactor model, factor independence was assumed. The nonsignificant loadings were removed, and the subscore (task) with only one item was not treated as a specific factor in the bifactor model.

Table 17

Confirmatory Factor Analysis Results for PAA 1 (Service Learning): Factor Correlation

Factor	2 factors (item type)		4 factors (subscore/task)[a]			
	SR	CR	S(T) 1	S(T) 2	S(T) 3	S(T) 4
2	.80		.89			
3			.60	.68		
4			.66	.73	.72	

Note. CR = constructed response, PAA = periodic accountability assessment, SR = selected response, S(T) = subscore(task).

[a]See Table 2 for task and subscore information.

Table 18

Confirmatory Factor Analysis Results for PAA 1 (Service Learning): Model Comparison

Chi-square test	1 vs. 2 factors	2 vs. 4 factors	1 vs. bifactor
χ^2 difference	173.36	95.80	406.83
df difference	1	4	14
p	.00	.00	.00

Note. PAA = periodic accountability assessment.

Tables 19–22 show the results of the CFA models for PAA 2 (Invasive Plant Species). INVASIVE_02_05 had negative factor loadings on all CFA models, indicating that it might not measure the construct that other items measured; therefore, this item should be further reviewed and revised by test developers. In the bifactor task model, this item's loading on the Task 2 specific factor was removed because otherwise this item became a Heywood case (i.e., estimated error variance of an item is negative), and some other items' loadings on the Task 1 or 2 specific factors were removed because of nonsignificant loadings. Note that there were positive and negative loadings on the Task 1 specific factor, indicating that the items in Task 1 had opposite relationships with this factor. The fit statistics show that the four-factor task model, the five-factor subscore model, and the bifactor task model might have adequate fits. The chi-squared comparisons of nested models were all significant, showing that the models with more factors fitted better in the following model pairs: the two-factor item type model versus the one-factor model, the five-factor subscore model versus the two-factor item type model, the five-factor subscore model versus the four-factor task model, and the bifactor task model versus the one-factor model. The estimated correlation between SR and CR factors was .85; however, this correlation might be confounded with subscore correlations as SR comprised Subscores 2 and 3, and CR comprised Subscores 1, 4, and 5. The estimated correlations ranged from .56 to .95 between task factors with the highest correlation between Tasks 1 and 2, and from .52 to .93 between subscore factors with the highest correlation between Subscores 2 and 3.

Table 19

Confirmatory Factor Analysis Results for PAA 2 (Invasive Plant Species): Factor Loading

Score ID	1 factor	2 factors (item type)		4 factors (task)[a]				5 factors (subscore)[a]					Bifactor (general+ task)[a,b]			
	General	SR	CR	T1	T2	T3	T4	S1	S2	S3	S4	S5	General	T1	T2	T4
INVASIVE_01_01	.65		.75	.62				1.00					.63	.06		
INVASIVE_01_02	.65	.68		.67					.69				.66			
INVASIVE_01_03	.24	.24		.22					.25				.23	-.43		
INVASIVE_01_04	.52	.53		.53					.55				.54			
INVASIVE_01_05	.20	.21		.20					.24				.19	-.34		
INVASIVE_01_06	.23	.23		.20					.22				.22	.17		
INVASIVE_01_07	.61	.63		.68					.70				.67	.38		
INVASIVE_01_08	.37	.39		.41					.42				.44			
INVASIVE_01_09	.58	.60		.55					.58				.52	-.46		
INVASIVE_01_10	.21	.26		.25					.28				.23	.28		
INVASIVE_01_11	.48	.44		.44					.44				.45	.35		
INVASIVE_01_13	.24	.22		.19					.20				.21	-.41		
INVASIVE_02_01	.90	.90			.92					.91			.91		.16	
INVASIVE_02_02	.62	.64			.64					.63			.64			
INVASIVE_02_03	.51	.55			.48					.49			.48			
INVASIVE_02_04	.49	.51			.52					.51			.41		.49	
INVASIVE_02_05	-.09	-.12			-.11					-.11			-.10			
INVASIVE_02_06	.69	.72			.70					.71			.72			
INVASIVE_02_07	.94	.94			.92					.92			.88		.27	
INVASIVE_02_08	.85	.88			.87					.87			.85		.11	
INVASIVE_02_09	.69	.71			.69					.70			.70		.09	
INVASIVE_02_10	.85	.85			.84					.83			.82		.21	
INVASIVE_02_11	.85	.84			.85					.85			.74		.48	
INVASIVE_02_12	.20	.19			.17					.17			.12		.32	
INVASIVE_02_13	.90	.90			.90					.90			.84		.33	
INVASIVE_02_14	.82	.83			.85					.85			.78		.35	
INVASIVE_02_15	.84	.85			.87					.87			.80		.30	

Score ID	1 factor	2 factors (item type)		4 factors (task)[a]				5 factors (subscore)[a]					Bifactor (general+ task)[a,b]			
	General	SR	CR	T1	T2	T3	T4	S1	S2	S3	S4	S5	General	T1	T2	T4
INVASIVE_02_16	.86	.86			.85					.85			.84		.13	
INVASIVE_03_01	.76		.79			1.00					1.00		.75			
INVASIVE_04_02_I	.90		.91				.91					.91	.64			.65
INVASIVE_04_02_III	.90		.92				.92					.92	.65			.65

Note. CR = constructed response, PAA = periodic accountability assessment, S1 = Subscore 1, S2 = Subscore 2, S3 = Subscore 3, S4 = Subscore 4, S5 = Subscore 5, SR = selected response, T1 = Task 1, T2 = Task 2, T3 = Task 3, T4 = Task 4.

[a] See Table 3 for task and subscore information. [b] For a bifactor model, factor independence was assumed. The nonsignificant loadings were removed, and the task with only one item was not treated as a specific factor in the bifactor model.

Table 20

Confirmatory Factor Analysis Results for PAA 2 (Invasive Plant Species): Fit Statistics

Fit statistic	1 factor	2 factors (item type)	4 factors (task)[a]	5 factors (subscore)[a]	Bifactor (general+ task)[a,b]
χ^2	1,927.25	1,736.06	1,556.92	1,533.76	1,247.19
df	434	433	429	426	412
p	.00	.00	.00	.00	.00
χ^2 / *df*	4.44	4.01	3.63	3.60	3.03
RMSEA	.06	.06	.05	.05	.05
NFI	.87	.89	.90	.90	.92
NNFI	.89	.91	.92	.92	.94
GFI	.96	.96	.96	.97	.97
SRMR	.14	.14	.11	.12	.11

Note. GFI = goodness of fit index, NFI = normed fit index, NNFI = nonnormed fit index, PAA = periodic accountability assessment, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual.

[a] See Table 3 for task and subscore information. [b] For a bifactor model, factor independence was assumed. The nonsignificant loadings were removed, and the task with only one item was not treated as a specific factor in the bifactor model.

Table 21

Confirmatory Factor Analysis Results for PAA 2 (Invasive Plant Species): Factor Correlation

Factor	2 factors (item type)		4 factors (task)[a]				5 factors (subscore)[a]				
	SR	CR	T1	T2	T3	T4	S1	S2	S3	S4	S5
2	.85		.95				.60				
3			.75	.70			.61	.93			
4			.69	.70	.56		.56	.72	.71		
5							.52	.67	.70	.58	

Note. CR = constructed response, PAA = periodic accountability assessment, S1 = Subscore 1, S2 = Subscore 2, S3 = Subscore 3, S4 = Subscore 4, S5 = Subscore 5, SR = selected response, T1 = Task 1, T2 = Task 2, T3 = Task 3, T4 = Task 4.

[a] See Table 3 for task and subscore information.

Table 22

Confirmatory Factor Analysis Results for PAA 2 (Invasive Plant Species): Model Comparison

Chi-square test	1 vs. 2 factors	2 vs. 5 factors	4 vs. 5 factors	1 vs. bifactor
χ^2 difference	191.19	202.30	23.15	680.06
df difference	1.00	7.00	3.00	22.00
p	.00	.00	.00	.00

Note. PAA = periodic accountability assessment.

The CFA results for PAA 3 (Ban Ads) are shown in Tables 23–26. In the bifactor task model, two items (BANADS_02AX_J and BANADS_02BX_F) had nonsignificant loadings on the specific factor on Task 2, and thus these two loadings were removed from the model. The fit statistics show that the four-factor task model, the six-factor subscore model, and the bifactor task model appeared to have adequate data-model fit. All the comparisons between nested models favored more complicated models: the one-factor model versus the two-factor item type model and the bifactor task model, the two-factor item type model versus the six-factor subscore model, and the four-factor task model versus the six-factor subscore model. Again, because SR comprised Subscores 1, 3, and 4, and CR comprised Subscores 2, 5, and 6, the estimated correlation between the SR and CR factors (.89) might be confounded with subscore correlations. The estimated correlations ranged from .69 to .84 between task factors with the highest correlation between Tasks 1 and 2, and from .56 to .79 between subscore factors with the highest correlation between Subscores 2 and 4.

Table 23

Confirmatory Factor Analysis Results for PAA 3 (Ban Ads): Factor Loading

Score ID	1 factor	2 factors (item type)		4 factors (task)[a]				6 factors (subscore)[a]						Bifactor (general+ task)[a,b]			
	General	SR	CR	T1	T2	T3	T4	S1	S2	S3	S4	S5	S6	General	T1	T2	T4
BANADS_01A_02	.38	.40		.40				.48						.38	.24		
BANADS_01A_03	.34	.35		.39				.50						.35	.19		
BANADS_01A_04	.17	.17		.17				.23						.13	.15		
BANADS_01A_05	.44	.43		.42				.53						.38	.21		
BANADS_01B	.79		.79	.83					.84					.77	.24		
BANADS_01C	.73		.72	.76					.76					.70	.38		
BANADS_02AX_A	.62	.65			.67					.71				.49		.54	
BANADS_02AX_B	.55	.56			.58					.61				.47		.47	
BANADS_02AX_C	.17	.14			.11					.10				.15		.12	
BANADS_02AX_D	.73	.74			.75					.76				.59		.45	
BANADS_02AX_E	.69	.71			.74					.78				.54		.58	
BANADS_02AX_F	.59	.60			.60					.63				.41		.61	
BANADS_02AX_G	.43	.45			.48					.55				.37		.43	
BANADS_02AX_H	.66	.67			.65					.67				.51		.47	
BANADS_02AX_I	.71	.73			.73					.74				.60		.28	
BANADS_02AX_J	.50	.52			.50					.49				.50			
BANADS_02BX_A	.08	.08			.08						.11			.08		.11	
BANADS_02BX_B	.30	.30			.32						.41			.31		.10	
BANADS_02BX_C	.45	.46			.46						.59			.52		.29	
BANADS_02BX_D	.45	.44			.47						.56			.51		.19	
BANADS_02BX_E	.18	.17			.20						.29			.26		.13	
BANADS_02BX_F	.54	.55			.56						.60			.59			

Score ID	1 factor	2 factors (item type)		4 factors (task)[a]				6 factors (subscore)[a]						Bifactor (general+ task)[a,b]			
	General	SR	CR	T1	T2	T3	T4	S1	S2	S3	S4	S5	S6	General	T1	T2	T4
BANADS_03	.80		.79			1.00						1.00		.81			
BANADS_04_I	.95		.95				.97						.97	.82			.51
BANADS_04_III	.92		.92				.91						.90	.76			.51

Note. CR = constructed response, PAA = periodic accountability assessment, S1 = Subscore 1, S2 = Subscore 2, S3 = Subscore 3, S4 = Subscore 4, S5 = Subscore 5, S6 = Subscore 6, SR = selected response, T1 = Task 1, T2 = Task 2, T3 = Task 3, T4 = Task 4.

[a]See Table 4 for task and subscore information. [b]For a bifactor model, factor independence was assumed. The nonsignificant loadings were removed, and the task with only one item was not treated as a specific factor in the bifactor model.

Table 24

Confirmatory Factor Analysis Results for PAA 3 (Ban Ads): Fit Statistics

Fit statistic	1 factor	2 factors (item type)	4 factors (task)[a]	6 factors (subscore)[a]	Bifactor (general+ task)[a,b]
χ^2	1,144.61	1,097.45	901.98	777.81	706.14
df	275	274	270	261	254
p	.00	.00	.00	.00	.00
χ^2 / *df*	4.16	4.01	3.34	2.98	2.78
RMSEA	.06	.05	.05	.04	.04
NFI	.93	.93	.94	.95	.96
NNFI	.94	.94	.96	.96	.97
GFI	.97	.97	.98	.98	.98
SRMR	.12	.11	.10	.09	.09

Note. GFI = goodness of fit index, NFI = normed fit index, NNFI = nonnormed fit index, PAA = periodic accountability assessment, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual.

[a]See Table 4 for task and subscore information. [b]For a bifactor model, factor independence was assumed. The nonsignificant loadings were removed, and the task with only one item was not treated as a specific factor in the bifactor model.

Table 25

Confirmatory Factor Analysis Results for PAA 3 (Ban Ads):

Factor Correlation

Factor	2 factors (item type)		4 factors (task)[a]				6 factors (subscore)[a]				
	SR	CR	T1	T2	T3	T4	S1	S2	S3	S4	S5
2	.89		.84				.79				
3			.72	.71			.57	.72			
4			.77	.73	.69		.56	.79	.59		
5							.61	.71	.60	.67	
6							.63	.77	.66	.62	.68

Note. PAA = periodic accountability assessment, S1 = Subscore 1, S2 = Subscore 2, S3 = Subscore 3, S4 = Subscore 4, S5 = Subscore 5, SR = selected response, T1 = Task 1, T2 = Task 2, T3 = Task 3, T4 = Task 4.

[a]See Table 4 for task and subscore information.

Table 26

Confirmatory Factor Analysis Results for PAA 3 (Ban Ads):

Model Comparison

Chi-square test	1 vs. 2 factors	2 vs. 6 factors	4 vs. 6 factors	1 vs. bifactor
χ^2 difference	47.16	319.64	124.17	438.48
df difference	1	13	9	21
p	.00	.00	.00	.00

Note. PAA = periodic accountability assessment.

Tables 27–30 list the CFA results for PAA 4 (Mango Street). All items in Task 3 had nonsignificant loadings on the specific factor on Task 3 in the bifactor task model; therefore, this factor was removed from the model. The fit statistics indicate all the models had adequate fit; however, as in the three other PAAs, the comparisons between nested models were all significant and favored complicated models: the one-factor model versus the two-factor item type model and the bifactor subscore model, and the two-factor item type model versus the four-factor subscore model. The estimated correlation between the SR and CR factors was .87, and again, this correlation might be confounded with subscore correlations because SR comprised Subscores 1 and 3, and CR comprised Subscores 2 and 4. The estimated subscore factor correlations were between .64 and .93 with the highest correlation between Subscores 1 and 3.

Table 27

Confirmatory Factor Analysis Results for PAA 4 (Mango Street): Factor Loading

Score ID	1 factor	2 factors (item type)		4 factors (subscore/task)[a]				Bifactor (general + subscore/task)[a,b]		
	General	SR	CR	S(T) 1	S(T) 2	S(T) 3	S(T) 4	General	S(T) 1	S(T) 4
MANGO_01_01	.65	.67		.67				.55	.47	
MANGO_01_02	.78	.79		.80				.71	.40	
MANGO_01_03	.38	.38		.38				.35	.18	
MANGO_01_04	.75	.77		.75				.70	.17	
MANGO_01_05	.81	.83		.82				.73	.36	
MANGO_02_01	.77		.76		1.00			.76		
MANGO_03_01	.50	.52				.52		.51		
MANGO_03_02	.79	.80				.78		.79		
MANGO_03_03	.57	.58				.57		.57		
MANGO_03_04	.70	.70				.71		.70		
MANGO_03_05	.48	.48				.49		.48		
MANGO_03_06	.71		.70			.70		.70		
MANGO_04_I	.92		.92				.93	.80		.47
MANGO_04_III	.92		.92				.92	.79		.47

Note. CR = constructed response, SR = selected response, S(T) = subscore (task).

[a]See Table 5 for task and subscore information. [b]For a bifactor model, factor independence was assumed. The nonsignificant loadings were removed, and the subscore (task) with only one item was not treated as a specific factor in the bifactor model.

Table 28

Confirmatory Factor Analysis Results for PAA 4 (Mango Street): Fit Statistics

Fit statistic	1 factor	2 factors (item type)	4 factors (subscore/task)[a]	Bifactor (general + subscore/task)[a,b]
χ^2	285.36	204.63	143.95	157.48
df	77	76	72	71
p	.00	.00	.00	.00
χ^2 / *df*	3.71	2.69	2.00	2.22
RMSEA	.05	.04	.03	.03
NFI	.96	.97	.98	.98
NNFI	.97	.98	.99	.99
GFI	.99	.99	.99	.99
SRMR	.10	.07	.05	.05

Note. GFI = goodness of fit index, NFI = normed fit index, NNFI = nonnormed fit index, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual.

[a]See Table 5 for task and subscore information. [b]For a bifactor model, factor independence was assumed. The nonsignificant loadings were removed, and the subscore (task) with only one item was not treated as a specific factor in the bifactor model.

Table 29

Confirmatory Factor Analysis Results for PAA 4 (Mango Street): Factor Correlation

Factor	2 factors (item type)		4 factors (subscore/task)[a]			
	SR	CR	S(T) 1	S(T) 2	S(T) 3	S(T) 4
2			.64			
3	.87		.93	.73		
4			.75	.69	.81	

Note. PAA = periodic accountability assessment.

[a]See Table 5 for task and subscore information. [b]For a bifactor model, factor independence was assumed. The nonsignificant loadings were removed, and the subscore (task) with only one item was not treated as a specific factor in the bifactor model.

Table 30

Confirmatory Factor Analysis Results for PAA 4 (Mango Street): Model Comparison

Chi-square test	1 vs. 2 factors	2 vs. 4 factors	1 vs. bifactor
χ^2 difference	80.73	60.68	127.88
df difference	1	4	6
p	.00	.00	.00

Note. PAA = periodic accountability assessment.

Conclusions

In this study, the dimensional structures of the four CBAL writing PAAs were investigated using EFA and CFA. The main findings are as follows:

- The parallel analyses suggest the optimal numbers of factors for PAAs 1–4 were 3, 5, 5, and 1, respectively.
- The comparisons of EFA models with the maximum likelihood estimation suggest large numbers of factors for all the four PAAs.
- The comparisons of nested CFA models as well as model fit indexes indicate that models with more factors fitted data better than simple models in all cases across the four PAAs. The four-factor task models, the four to six-factor subscore models, and the bifactor task models in all four PAAs showed adequate model-data fit and had better fit than the one-factor models and the two-factor item type models. The two-factor item type models fitted better than the one-factor models. In addition, for PAAs 2 and 3, separating two subscores in one or two tasks led to the better fitted subscore models when compared to the task models.

All the results suggest the four writing tests were multidimensional except that the PA recommends one dimension for PAA 4. Furthermore, the CFA results support the subscore structures as well as the bifactor models in the four PAAs. These findings suggest that the skills represented by the subscores in each PAA are well-defined and distinguishable, and within each of these subscores or tasks a common writing ability is assessed. The results provide evidence to support the construct validity of the four PAAs and the competency model underlying the test designs. The implication for test equating and scoring is that a multidimensional model should be used for test equating and scoring, and subscores should be reported. If desired, overall writing scores can be reported using the bifactor model.

The fact that the two-factor item type models performed better than the one-factor models may indicate that item type was an influential factor on item performance. However, in all four PAAs, item type was confounded with task or subscore, and moreover, the task or subscore models had better fit than the item type models. Therefore, under the current test designs, the results shed little light on the question of whether SR and CR items represent different test dimensions.

One limitation of this study is that a convenience sample was used in the analyses, and the sample size is relatively small for analyzing large-scale assessments such as CBAL tests. Hence, the conclusions regarding the dimensionality of these tests from the current study should be considered as preliminary. As more test data from representative samples become available for these writing tests in the future, the dimensionality of these tests should be further examined. In addition, other dimensionality assessment methods as recommended in De Champlain and Gessaroli (1998), Levy and Svetina (2010), and Hattie (1984, 1985) can be used to analyze these tests (i.e., the methods under item response theory).

References

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.

Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models.* Newbury Park, CA: SAGE.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation, 10*(7). Retrieved from http://pareonline.net/getvn.asp?v=10&n=7

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., . . . Wagner, M. (2009). *Progress on designing the CBAL summative writing assessment: Design principles and results.* Unpublished manuscript.

De Champlain, A., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education, 11,* 231–253.

Garson, G. D. (2011). *Correlation. Statnotes: Topics in multivariate analysis*. Retrieved from http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm

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.

Hattie, J. A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19,* 49–78.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Joreskog, K. G., & Sorbom, D. (1996a). *LISREL 8 user's reference guide*. Chicago, IL: Scientific Software International.

Joreskog, K. G., & Sorbom, D. (1996b). *PRELIS 2 user's reference guide*. Chicago, IL: Scientific Software International.

Joreskog, K. G., Sorbom, D., du Toit, S., & du Toit, M. (2000). *LISREL 8: New statistical features*. Chicago, IL: Scientific Software International.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York, NY: Guilford.

Levy, R., & Svetina, D. (2010, May). *A framework for dimensionality assessment for multidimensional item response models.* Paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.

O'Conner, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers, 32,* 396–402.

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.

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Testing and Assessment Modeling, 52,* 354–379.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27,* 159–203.

Zwick, W. R. (1987). *Assessment of dimensionality of NAEP Year 15 reading data* (Research Report No. RR-86-04). Princeton, NJ: Educational Testing Service.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432–442.

Notes

[1] *Item* in this paper was defined as *final score unit*. The score for a multipart item was based on the scores of all the items in the set; however, the multipart item was counted as one item. On the other hand, two scores were given for each essay and counted as two items.

[2] Costello and Osborne (2005) showed that even with a 20:1 subject to item ratio the misclassification rate of factor structures in EFA was well above 5%. For analyzing a correlation matrix in CFA using the generally WLS estimation method as we did in this paper, the minimum sample size requirement is $k(k-1)/2$, where k is the number of variables. The sample size requirement is necessary for a stable estimate of the asymptotic variance-covariance matrix of the polychoric correlations, which is needed for WLS to calculate the weight matrix in the fit function (Joreskog & Sorbom, 1996a, pp. 21–23; Joreskog & Sorbom, 1996b, pp. 167–171).

[3] Because INVASIVE_01_01 (Item 1 in Invasive Plant Species) had 21 score categories, this score was treated as a continuous variable, and its interitem correlations were polyserial correlations.