ELPA21

English Language Proficiency Assessment for the 21st Century

Item Analysis and Calibration

October 2017

Document Change History

Date	Version	Change made by	Description
October 2017	1.0	---	---

Table of Contents

List of Tables

List of Figures

Acknowledgements

The English Language Proficiency Assessment for the 21st Century (ELPA21) and the National Center for Research on Standards, Evaluation, and Student Testing (CRESST) would like to acknowledge the contributions of many organizations and individuals who participated in the processes that CRESST employed to determine item parameters and calibrate the item bank. ELPA21 is built on the foundation of thoughtful collaboration with experts and educators across the nation, specifically, the agency staff and educators of ELPA21 states, task management team leads, and the Council of Chief State School Officers (CCSSO), all of whom participated in data review, differential item functioning review (DIF review), and other aspects of the item bank calibration process. We wish to acknowledge and thank these individuals for their contributions.

DIF Review
DIF review meeting was held in July 2016 and involved the ELPA21 Administration, Accessibility, and Accommodations (AAA) Task Management Team (TMT). The review was facilitated by Martha Thurlow and Vitaliy Shyyan from the National Center on Educational Outcomes (NCEO). Participants included:

State representatives
- Andrew Hinkle (Ohio Department of Education)
- Brad Lenhardt (Oregon Department of Education)
- Nancy Rowch (Nebraska Department of Education)
- Brooke David (Nebraska Department of Education)

CCSSO
- Kristin Baddour
- Margaret Ho
- Cathryn Still

CRESST
- Mark Hansen

NCEO
- Deb Albus
- Linda Goldstone

Data Review
The data review meeting was held in July 2016 and involved the ELPA21 Assessment Design and Scaling Task Management Team. The review was facilitated by Bill Auty from Education Measurement Consulting. Participants included:

State representatives
- Kurt Taube (Ohio Department of Education)
- Steve Slater (Oregon Department of Education)

CCSSO
- Kristin Baddour
- Cathryn Still
- Margaret Ho

CRESST
- Mark Hansen

ELPA21 Item Acquisition and Development Task Management Team
- Phoebe Winter

The report was prepared by Mark Hansen with help from Seth Leon (CRESST). It was reviewed by Li Cai and Cathryn Still (CRESST) and the ELPA21 Research & Evaluation Committee, including the following state representatives: Nancy Rowch (Nebraska Department of Education), Steve Slater (Oregon Department of Education), Lucas Snider (Washington Office of Superintendent of Public Instruction), and Kurt Taube (Ohio Department of Education).

English Language Proficiency for the 21st Century (ELPA21): Item Analysis and Calibration

Introduction

This document describes the procedures used in evaluating the performance of items in the ELPA21 pool and in obtaining parameters to be used in operational scoring of the ELPA21 assessments.

Part 1 of this report describes the data used for these analyses. All data were collected in the 2015-16 summative test administration. A preliminary (early-return) sample was used for initial item analyses, and a larger sample was used for the final item calibrations.

Part 2 describes item analyses that were conducted prior to performing final calibrations. Items were first examined with respect to several descriptive statistics, including the proportion of respondents in each score point, the average item score, and the item-total correlation, among others. Initial item response theory (IRT) calibrations provided preliminary parameter estimates to complement these descriptive statistics. Items were also evaluated for differences in functioning for students with disabilities, compared to those without. Items with classical statistics falling outside an acceptable range or showing evidence of bias were reviewed.

A total of 261 items were flagged based on their descriptive statistics. Members of the ELPA21 Assessment Design and Scaling (ADS) and Item Acquisition and Development (IAD) Task Management Teams reviewed these items. One item was determined to have been scored incorrectly. Upon correction of the scoring rule, this item was accepted. Two other items were rejected from the pool, based on low or negative item-total correlations. All other items flagged on the basis of their descriptive statistics were deemed acceptable.

The analysis of differential item functioning for students with disabilities resulted in 67 items being flagged, with a similar number of items seemingly favoring students with disabilities (37 items) as favoring students without disabilities (30 items). These 67 items were reviewed by the ELPA21 Administration, Accommodations, and Accessibility (AAA) Task Management Team. All were judged to be acceptable.

All items with acceptable statistical results or approved for use following individual review were included in the final IRT models. Part 3 of this report describes the estimation of these models. Calibrations were performed separately for each of ELPA21's six gradebands (Kindergarten, 1, 2-3, 4-5, 6-8, and 9-12). The resulting parameter estimates allow for the computation of IRT scaled scores representing skills in each of four domains of language use (Listening, Reading, Speaking, and Writing), overall language proficiency (a composite scale score based on performance across all four domains), and language comprehension (a composite scale score based on performance in the Listening and Reading domains). Scoring parameters are provided for all calibrated items.

Part 4 describes additional IRT analyses that were used to examine the relationships of student performance in language domains across gradebands. These analyses utilized a small number of non-operational items administered in the gradeband above their intended level (e.g., Kindergarten items embedded in the Grade 1 test forms). The domain scores were generally found to correlate strongly across gradebands (above 0.93, on average). In addition, performance of students in the higher gradeband was generally found to be higher on average than the performance of students in the lower gradeband. These differences were largest in the lowest gradebands.

Part 1: Data Collection

All data used for item analysis and calibration were collected within the 2015-16 ELPA21 summative test administration. In order to maximize the number of test items analyzed, five to six alternate forms were constructed for each domain and gradeband. Forms were then randomly assigned to students. As a result, each alternate form was used with similar frequency, and item response data not collected for a particular student are missing by design. Hence, they could be treated as Missing Completely At Random (Rubin, 1987), which facilitates subsequent statistical analysis.

Test Forms

Table 1 illustrates the general approach used to construct alternate forms for a given gradeband and domain. The guiding insight should be understood as aiming for overall balance. Within each task type, the available tasks were distributed across the test forms, based on the number required, per the operational test blueprints (Appendix A). In the example shown in Table 1, there are six tasks of type A, three tasks of type B, two tasks of type C, and four tasks of type D. Suppose the test blueprint requires one task for types A, B, and C and two tasks of type D. Each type A task would be used once, each type B task would be used twice, and each type C task would be used three times. Each type D task would also be used three times, and the combinations of type D tasks would vary over the forms. Note that tasks may consist of a single item or multiple items. When a task consists of multiple items, those items were always administered as a set (i.e., tasks were kept intact). Appendix B summarizes the number of operational items (as well as the number of scores and score points) for each test form administered.

Table 1. Sample Distribution of Tasks Across Alternate Test Forms.

Task	Test Form #						Type	Status
	1	2	3	4	5	6		
A1	x						A	Operational
A2		x					A	Operational
A3			x				A	Operational
A4				x			A	Operational
A5					x		A	Operational
A6						x	A	Operational
B1	x	x					B	Operational
B2			x	x			B	Operational
B3					x	x	B	Operational
C1	x	x	x				C	Operational
C2				x	x	x	C	Operational
D1	x	x	x				D	Operational
D2	x			x	x		D	Operational
D3		x		x		x	D	Operational
D4			x		x	x	D	Operational
E1	x						E	Off-grade
F1		x					F	Off-grade
G1			x				G	Off-grade
H1				x			H	Field Test
I1					x		I	Field Test
J1						x	J	Field Test

In addition to the operational tasks, the test forms included a small number of non-operational tasks (i.e., tasks not ultimately contributing to students' scores). The majority of these were tasks developed for the gradeband just below that of the test form. For example, Grade 1 forms included a small number of tasks from the Kindergarten test forms. The embedding of these off-gradeband tasks facilitated comparisons of performance across gradebands, as described in Part 4 of this report. The remaining non-operational tasks

were developed specifically for the ELPA21 screener forms and are not required by the summative test blueprints. These tasks occupied what might be considered field test slots. Like the operational tasks, off-gradeband and field test tasks were also distributed across the alternate forms. Unlike the operational tasks, it was not essential that the number of non-operational tasks per task type be equal across the alternate forms. For example, only Form 1 in Table 1 has a type D task; only Form 2 has a type E task.

Of course, the available tasks didn't always divide quite as neatly across the alternate test forms as is illustrated in Table 1. A general principal guiding test form construction was to ensure that, within task type, the difference in the minimum and maximum number of forms on which a task appeared was never greater than one. For example, it would be acceptable for one task to appear on two forms and another task of the same type to appear on three forms. On the other hand, it would be unacceptable for one task to appear on two forms and another appear on four forms. The content of the operational portion of each test form followed the test blueprint. To that, non-operational tasks (field test and/or below-gradeband tasks) were added. Estimates of item difficulty and response time from the 2015 Field Test (Questar Assessment, 2016) were used to ensure that the resulting forms would be similar in overall test length and difficulty.

Samples

Initial item analyses were performed using a preliminary (early return) sample of cases with item scores available shortly after the close of the 2015-16 summative test administration window. This sample was the basis for computation of classical item statistics and differential item functioning statistics. Final item calibrations were performed using a more complete sample. The sample sizes for the preliminary and final samples by gradeband are shown in Table 2. The preliminary sample represented approximately 38% of the final sample.

Table 2. Sample Sizes for Preliminary and Final Item Analyses

Gradeband	Preliminary	Final
K	11,922	37,305
1	12,287	37,705
2-3	22,990	70,984
4-5	19,611	55,189
6-8	24,068	55,602
9-12	22,559	49,285
TOTAL	113,437	306,070

Both the preliminary and final samples used in item analyses and calibration were limited to online tests using the alternate fixed forms developed by the consortium and in which all domains were administered. Tests labeled as incomplete or invalidated were excluded, as were tests from students with multiple records.

Table 3 provides some demographic information about the final sample (used in item calibrations). For some variables, information was not available for a rather high proportion of students. When information is missing, the status of the student for that variable is counted as "Unknown." The proportions reported for students with information should thus be understood as lower bounds. For example, it appears that at least 16.1% of students in the sample were economically disadvantaged, but there is another 82.0% for whom status is unknown; 10.1% of students were identified as having a disability, but another 23.6% have unknown status. Within this sample, Spanish was by far the most common first language, accounting for at least 37.2% of students across all gradebands.

Table 3. Characteristics of the Final Calibration Sample.

Status	Grade K		Grade 1		Grades 2-3		Grades 4-5		Grades 6-8		Grades 9-12		All Grades	
	N	P	N	P	N	P	N	P	N	P	N	P	N	P
Gender														
Female	15,702	.421	15,901	.422	28,544	.402	20,976	.380	19,160	.345	16,814	.341	117,097	.383
Male	17,046	.457	17,042	.452	32,046	.451	24,494	.444	23,607	.425	21,060	.427	135,295	.442
Unknown	4,557	.122	4,762	.126	10,394	.146	9,719	.176	12,835	.231	11,411	.232	53,678	.175
Disability (IEP and/or 504 Plan)														
Yes	2,371	.064	2,611	.069	5,917	.083	6,421	.116	7,672	.138	5,961	.121	30,953	.101
No	28,385	.761	28,385	.753	49,942	.704	35,908	.651	31,876	.573	28,422	.577	202,918	.663
Unknown	6,549	.176	6,709	.178	15,125	.213	12,860	.233	16,054	.289	14,902	.302	72,199	.236
Economic Disadvantage														
Yes	5,863	.157	6,184	.164	10,615	.150	8,298	.150	9,403	.169	8,814	.179	49,177	.161
No	780	.021	824	.022	1,210	.017	894	.016	939	.017	1,292	.026	5,939	.019
Unknown	30,662	.822	30,697	.814	59,159	.833	45,997	.833	45,260	.814	39,179	.795	250,954	.820
Hispanic Ethnicity														
Yes	19,366	.519	20,090	.533	37,658	.531	28,884	.523	27,264	.490	21,899	.444	155,161	.507
No	13,362	.358	12,800	.339	22,399	.316	16,109	.292	14,947	.269	15,338	.311	94,955	.310
Unknown	4,577	.123	4,815	.128	10,927	.154	10,196	.185	13,391	.241	12,048	.244	55,954	.183
American Indian or Alaskan Native														
Yes	1,240	.033	1,322	.035	2,561	.036	2,315	.042	1,960	.035	1,444	.029	10,842	.035
No	26,373	.707	26,139	.693	45,574	.642	34,251	.621	32,256	.580	28,365	.576	192,958	.630
Unknown	9,692	.260	10,244	.272	22,849	.322	18,623	.337	21,386	.385	19,476	.395	102,270	.334
Asian														
Yes	4,141	.111	3,635	.096	5,510	.078	3,718	.067	3,675	.066	4,413	.090	25,092	.082
No	23,472	.629	23,826	.632	42,625	.600	32,848	.595	30,541	.549	25,396	.515	178,708	.584
Unknown	9,692	.260	10,244	.272	22,849	.322	18,623	.337	21,386	.385	19,476	.395	102,270	.334
Native Hawaiian or Other Pacific Islander														
Yes	348	.009	323	.009	607	.009	527	.010	638	.011	660	.013	3,103	.010
No	11,824	.317	12,338	.327	23,517	.331	17,397	.315	18,116	.326	17,402	.353	100,594	.329
Unknown	25,133	.674	25,044	.664	46,860	.660	37,265	.675	36,848	.663	31,223	.634	202,373	.661
First Language														
English	700	.019	697	.018	1,271	.018	1,084	.020	1,097	.020	655	.013	5,504	.018
Russian	899	.024	836	.022	1,296	.018	869	.016	593	.011	369	.007	4,862	.016
Somali	422	.011	443	.012	666	.009	552	.010	494	.009	592	.012	3,169	.010
Spanish	13,213	.354	14,995	.398	28,215	.397	21,668	.393	19,916	.358	15,773	.320	113,780	.372
Vietnamese	70	.002	70	.002	120	.002	94	.002	113	.002	151	.003	618	.002
Other	8,526	.229	6,839	.181	10,434	.147	7,054	.128	7,032	.126	7,850	.159	47,735	.156
Unknown	13,475	.361	13,825	.367	28,982	.408	23,868	.432	26,357	.474	23,895	.485	130,402	.426
All Students														
	37,305		37,705		70,984		55,189		55,602		49,285		306,070	

Part 2: Item Analysis

Prior to performing final calibrations, tasks were examined with respect to several descriptive statistics and evaluated for differential item functioning based on disability status.

Analysis of Descriptive (Classical) Item Statistics

The initial analyses included 2,080 items and utilized the preliminary sample described in Part 1. Not included in this initial analysis were 36 items from the Kindergarten and Grade 1 writing supplement paper forms.[1] The distribution of these items across domains and gradebands is shown in Table 4.

Table 4. Items Included in Initial Analysis, by Domain and Gradeband.

Gradeband	Listening	Reading	Speaking	Writing	All Domains
K	163	132	58	60	413
1	113	146	51	53	363
2-3	113	123	46	68	350
4-5	103	113	65	63	344
6-8	107	94	53	37	291
9-12	102	127	53	37	319
All Grades	701	735	326	318	2,080

Various descriptive statistics were computed from the item-level data of students in the preliminary sample. Among the statistics were descriptions of the distribution of students across response options and score points, the average item score, and relationships between item response or score and total score (including correlation between the item score and total score, correlations between distractors and the total score, and the and the average total score by response option). For these analyses, the criterion/total score was the summed scored for the domain test on which the item appeared.

Once the various descriptive statistics were computed, their values were compared with the threshold values specified in Table 5. Items with descriptive statistics outside of the acceptable range were flagged for further review. Because items could appear on multiple forms and total scores from different forms might not be directly comparable, the descriptive stats were computed separately by form. If any item was flagged based on statistics obtained from any form on which it appeared, the item was included in the subsequent review.

Table 5. Item Flags: Criteria for Evaluating Descriptive Item Statistics.

Flag	Description
1	average item score (divided by maximum possible score) < .10
2	average item score (divided by maximum possible score) > .95
3	minimum proportion achieving item score < .03
4	proportion of invalid responses (skipped, omitted, not reached, unscoreable) > .20
5	item-total biserial/polyserial correlation < .20
6	average total score increases with each score point
7	high ability students (above 80th percentile on criterion score) prefer a distractor over the key
8	mean criterion score for students selecting distractor is higher than mean for students selecting key
9	positive distractor biserial correlation

Note: Flags 7-9 applied to multiple-choice single-select (MCSS) items only.

[1] Scores for these items were not available at the time of this analysis. However, descriptive stats for these items were computed subsequently, and these items were included in the DIF analyses described in the next section.

It should be noted that all items included in this analysis had been previously examined in the 2015 Field Test and deemed acceptable for inclusion on operational test forms. One benefit of again evaluating these items is that in the operational administration, each item was administered to a much larger number of students than had participated in the field test. A second benefit is that there would be less reason for concern about student motivation in the operational test than in a non-operational field test study (i.e., the operational data may provide a more accurate characterization of how items can be expected to function).

Table 6 summarizes the number of items flagged based on each criterion. In total, 261 items were flagged. By far, the most frequent causes of an item being flagged were that the item was extremely easy (flag 2, average item score > .95), that at least one score point was achieved by fewer than 3% of examinees (flag 3). Many items (66) were flagged based on both of these criteria.

Table 6. Number of Items Flagged, by Criterion.

Flag	Description	Unflagged		Flagged	
		N	P	N	P
1	average item score < .10	2079	>.999	1	<.001
2	average item score > .95	1924	.925	156	.075
3	minimum proportion achieving item score < .03	1928	.927	152	.073
4	proportion of invalid responses (skipped, omitted, not reached, unscoreable) > .20	2080	1.000	0	.000
5	item-total biserial/polyserial correlation < .20	2074	.997	6	.003
6	average total score increases with each score point	2073	.997	7	.003
7	high ability students (above 80th percentile on criterion score) prefer a distractor over the key	2080	1.000	0	.000
8	mean criterion score for students selecting distractor is higher than mean for students selecting key	2079	>.999	1	<.001
9	positive distractor biserial correlation	2065	.993	15	.007
	any flag	1819	.875	261	.125

Note: MCSS = multiple-choice single-select

Table 7 shows how the flagged items were distributed across the domains and gradebands. In general, Listening and Writing items were flagged at a higher rate than Reading and Speaking items. Items in the 2-3, 4-5, and 6-8 gradebands were flagged at a higher rather than other bands.

Table 7. Items Flagged Based on Descriptive Statistics, by Domain and Gradeband.

Gradeband	Listening			Reading			Speaking			Writing			All Domains		
	flagged		total	flagged		total	flagged		total	flagged		total	flagged		total
	N	P		N	P		N	P		N	P		N	P	
K	7	.04	163	10	.08	132	0	.00	58	0	.00	60	17	.04	413
1	15	.13	113	4	.03	146	10	.20	51	2	.04	53	31	.09	363
2-3	20	.18	113	27	.22	123	1	.02	46	3	.04	68	51	.15	350
4-5	22	.21	103	9	.08	113	13	.20	65	14	.22	63	58	.17	344
6-8	39	.36	107	1	.01	94	9	.17	53	22	.59	37	71	.24	291
9-12	9	.09	102	11	.09	127	0	.00	53	13	.35	37	33	.10	319
All Grades	112	.16	701	62	.08	735	33	.10	326	54	.17	318	261	.13	2080

All items flagged on the basis of their descriptive statistics were forwarded to a review committee comprised of members of the ELPA21 Assessment Design and Scaling (ADS) and Item Acquisition and Development (IAD) Task Management Teams for further review. The team was provided the descriptive statistics for each

item, as well as item parameter estimates from preliminary calibrations using unidimensional item response theory (IRT) models fitted to the items from a given domain and gradeband. The models were estimated using flexMIRT® version 3 (Cai, 2016). Additional details concerning the IRT calibrations are provided in Part 3.

The review committee met via web conference on July 8, 2016 to review the item analysis results. The results of this discussion and subsequent follow-up discussion are summarized below:

Criterion 1 (average item score < .10). Inspection of the one item flagged due to its low average score (.019) led the team to suspect that the item had been incorrectly scored. This was subsequently confirmed. Upon rescoring, the average score for the item was found to be in an acceptable range (.769), and the item was retained.

Criterion 2 (average item score > .95). The review team recommended that items should not be removed due to a high average item score alone. The rationale for this recommendation was that while such items may have limited value (in terms of marginal reliability, for example) when administered to the full population of students taking the summative assessment, there are some students (at lowest levels of language ability) for whom these items would be informative. If future tests were to be administered adaptively, it would be desirable to have calibrated items with levels of difficulty that span the range of student ability. These items might also be appropriate for administration within the ELPA21 screener (for which the population of examinees could have greater variability). Accordingly, all 156 items flagged based on this criterion 2 were retained.

Criterion 3 (minimum proportion achieving item score < .03). The primary decision to be made for items flagged based on criterion 3 was whether adjacent score points should be combined (collapsed). If collapsing is deemed necessary, a second question is what kind of collapsing is most appropriate. After reviewing descriptive statistics and preliminary item parameters for the 152 flagged items, the review team recommended that item score points should not be collapsed. Despite these items having one or more score point achieved with very low frequency, there was evidence that these score points were otherwise functioning as expected. For example, evaluation of the average total scores by item score point showed that the score points were properly ordered (i.e., the items were not flagged based on criterion 6). Similarly, the item-total correlations were in an acceptable range (criterion 5). Finally, preliminary IRT calibrations showed that these item's parameters could be estimated with reasonable certainty (small standard errors). Consequently, all 152 items were retained without collapsing.

Criterion 4 (proportion of invalid responses > .20). No items were flagged on the basis of this criterion.

Criterion 5 (item-total biserial/polyserial correlation < .20). Six items were flagged based on criterion 5. Each of these items were inspected by members of the review team. One of the six was the item flagged based on criterion 1, which was found to have been incorrectly scored. This item was retained after correcting the scoring rule. Among the remaining five items, three were retained and two were rejected. The three retained items had item-total correlations of .174, .187, .194 (and initial slope parameter estimates of .138, .218, and .252). The two that were rejected had item-total correlations of .045 and .106 (and initial slope parameter estimates of -.077 and -.055).

Criterion 6 (average total score increases with each score point). Seven items were flagged due to the average total score not increasing with each score point. In each of the seven cases, the average total score for examinees with a score of 0 points on the item was slightly higher than the average total score for examinees with a score of 1 point. In contrast, the average item score for examinees with item scores of 2 or more points was consistently higher (and for items with additional score points above 2, averages of total

scores were ordered as expected). The results suggest that the students awarded 0 points or 1 point are similar in their overall ability. As such, collapsing these scores was considered. However, it was noted that for each of the seven items, the average total score for students awarded 1 point was based on a rather small number of cases; across the items, the number of examinees awarded 1 point ranged from 52 to 226 students (six of the seven items were also flagged based on criterion 3). Thus, the reversals in ordering could be due to sampling variability. (Three of the seven flagged items appeared on two test forms. For each of these three items, one instance was flagged based on this criterion and the other instance was not flagged. That is, in the second instance, average total score increased with item score.) These items were found to have strong item-total correlations (all polyserial correlations above .610), and initial calibrations produced stable parameter estimates. Based on these results, the review team judged these items to be acceptable without any collapsing of score points.

Criterion 7 (high ability students (above 80th percentile on criterion score) prefer a distractor over the key). No items were flagged on the basis of this criterion.

Criterion 8 (mean criterion score for students selecting distractor is higher than mean for students selecting key). One item was flagged based on this criterion. The same item was also flagged base on its low item-total correlation (criterion 5) and positive correlation between the total score and one of the distractor options (criterion 9). The review committee rejected this item.

Criterion 9 (positive distractor biserial correlation). Fifteen multiple-choice single-select items were flagged as a result of having at least one distractor with a positive correlation with the total score. The review team noted that the correlations, while positive, were quite close to zero and—with only one exception—much smaller than the correlation between the correct response and the total score. The exception was an item with a low item-total correlation (.045) and thus also flagged based on criteria 5 and 8. This item was rejected, while the other fourteen items flagged based on criterion 9 were accepted.

In summary, among the 2,080 items examined, 261 items were flagged based on one of more of the criteria described in Table 5. Among these 261 items, 258 items were accepted as-is, one item was accepted following correction of its scoring rule, and two items were rejected. Both of the rejected items had been flagged as a result of having low item-total correlations (criterion 5). One of the two rejected items was also flagged based on criteria 8 and 9.[2] A third item was subsequently rejected—not due to item statistics but because the item was found to contain a factual error.

Analysis of Differential Item Functioning (DIF) According to Disability Status

Data from the preliminary sample described above were used to compare the functioning of items across students with and without a disability. Note that differential functioning across other subgroups (gender, ethnicity, economic status, and English learner status) was previously examined in conjunction with the 2015 field test (Questar Assessment, 2016). Items were analyzed by computing standard DIF statistics for dichotomous and polytomous items (e.g., Zwick, Donoghue, & Grima, 1993).

In order to compute the DIF statistics, the distribution of item scores was obtained within five levels of language for the two student subgroups (defined by disability status). The five levels were based on estimated scale scores obtained by applying scoring parameters from the initial item calibrations. The observed scale

[2] As noted above, 36 Kindergarten and Grade 1 writing items were not included in the analysis because the item scores were not available at the time of this analysis and review. All 36 items were subsequently evaluated the criteria 1-6 (criteria 7-9 only applicable to MCSS items). Of these, 11 items were flagged: five items were flagged due to criterion 2 only, and six items were flagged due to criteria 2 and 3. All 36 items were retained, for the same reasons that other items flagged for these two criteria were retained: the items showed strong item-total correlations and the preliminary item calibrations produced stable estimates.

among the students administered a particular item was divided into five bins of equal width. For dichotomous items, we computed the Mantel-Haenszel (1959) chi-square statistic, Holland and Thayer's (1988) Mantel-Haenszel δ difference (MH D-DIF) and its standard error, and an effect size based on the standardized mean difference in item scores across groups (Dorans & Kulick, 1986; Zwick, Donoghue, & Grima, 1993). For polytomous items, we computed the Mantel-Haenszel chi-square and effect size statistics.

DIF statistics obtained for a given item would be expected to vary over samples of examinees. In addition, very small differences in item functioning may be of little consequence or concern. Accordingly, DIF results should be interpreted in terms of both statistical and practical significance. Table 8 summarizes one widely accepted scheme (see, e.g., Michaelides, 2008) for categorizing DIF as negligible (category A), intermediate (category B), or large (category C).

Table 8. Criteria for Interpreting DIF results.

Category	Dichotomous Items	Polytomous Items
A	|MH D-DIF| < 1 *- or -* MH D-DIF is not significantly different from zero ($p \geq .05$)	MH chi-square is not significantly different from zero ($p \geq .05$) *- or -* |ES| ≤ .17
B	|MH D-DIF| is significantly different from zero ($p < .05$) *- and -* |MH D-DIF| > 1 *- and either -* |MH D-DIF| < 1.5 *- or -* |MH D-DIF| is not significantly different from one ($p \geq .05$)	MH chi-square is significantly different from zero ($p < .05$) *- and -* .17 < |ES| ≤ .25
C	|MH D-DIF| is significantly greater than one ($p < .05$) *- and -* |MH D-DIF| > 1.5	MH chi-square is significantly different from zero ($p < .05$) *- and -* |ES| > .25

Note: A = negligible DIF, B = Intermediate DIF, C = Large DIF. Criteria from review/summary article by Michaelides (2008).

In these analyses, students with disabilities were treated as the focal group, while students with no disability were the reference group. Thus, positive effect sizes indicate higher item scores for students with disabilities (conditioning on ability), and negative effect sizes indicate lower item scores for students with disabilities. This direction of bias can be indicated by adding a plus (+) or minus (–) sign to the DIF categories described above.

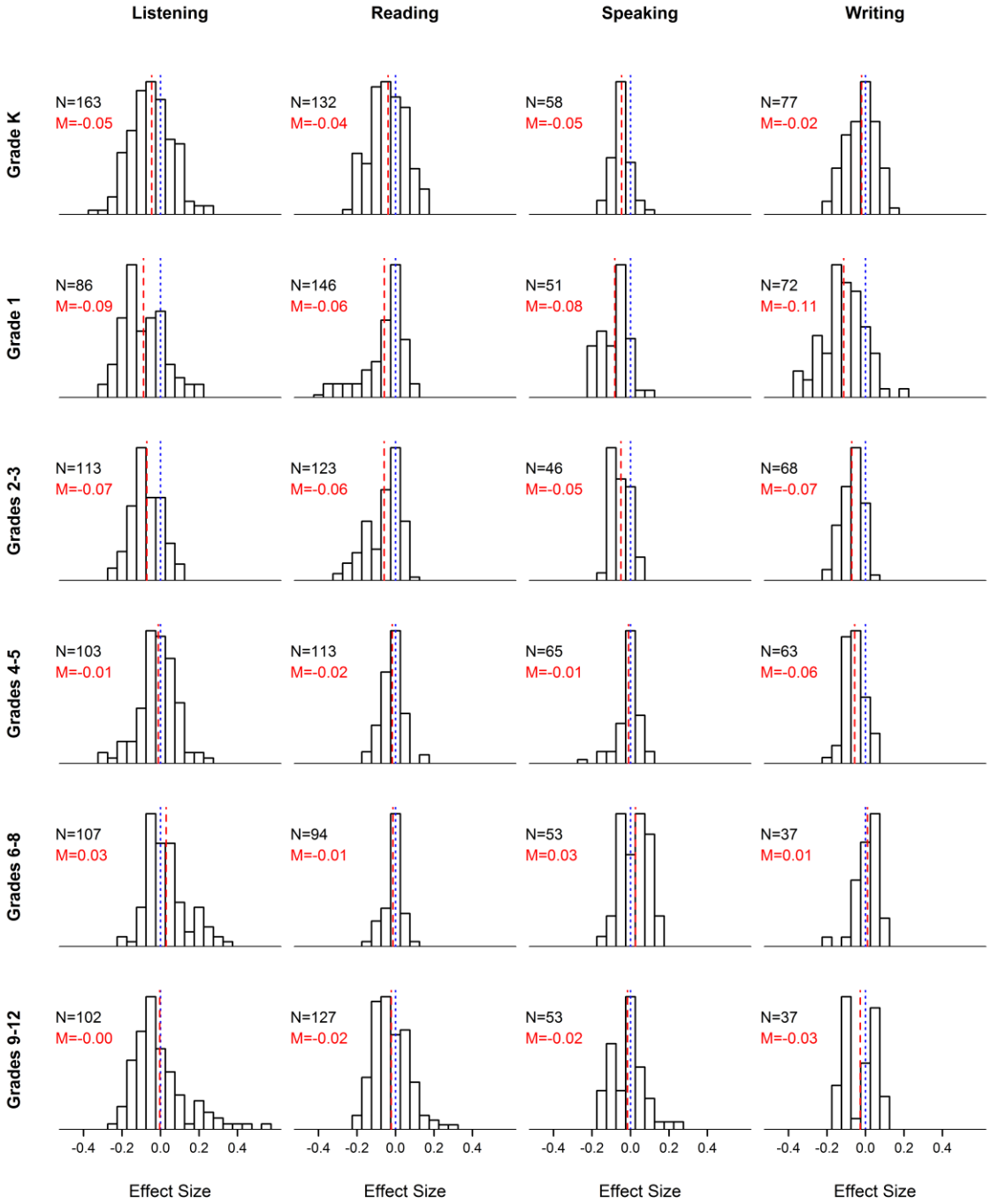
Table 9 summarizes the results of the DIF analyses. A total of 2,116 items were examined, including the 2,080 for which descriptive analyses had been obtained, as well as an additional 36 Kindergarten and Grade 1 writing supplement items. Among the 2,116 items examined, 1,857 (87.8%) were assigned to DIF category A, 165 (7.8%) were assigned to category B (6.1% B–; 1.7% B+), and 67 (3.2%) were assigned to category C (1.4% C–; 1.7% C+). The distribution of effect size estimates for each item evaluated are shown in Figure 1. The mean effect sizes were generally quite close to zero, with no clear pattern relating between the magnitude or direction to gradeband or domain.

Table 9. Analyses of DIF According to Disability Status: DIF Category by Domain and Gradeband.

Domain	# Items	C–		B–		A		B+		C+		Excluded		Effect Size	
		N	P	N	P	N	P	N	P	N	P	N	P	M	SD
Grade K															
Listening	163	2	.012	19	.117	132	.810	8	.049	2	.012	0	.000	-.046	.107
Reading	132	0	.000	8	.061	118	.894	6	.045	0	.000	0	.000	-.038	.092
Speaking	58	0	.000	0	.000	58	1.000	0	.000	0	.000	0	.000	-.046	.048
Writing	77	0	.000	4	.052	73	.948	0	.000	0	.000	0	.000	-.020	.078
All Domains	**430**	**2**	**.005**	**31**	**.072**	**381**	**.886**	**14**	**.033**	**2**	**.005**	**0**	**.000**	**-.039**	**.092**
Grade 1															
Listening	113	3	.027	15	.133	65	.575	3	.027	0	.000	27	.239	-.088	.111
Reading	146	8	.055	12	.082	126	.863	0	.000	0	.000	0	.000	-.058	.111
Speaking	51	0	.000	8	.157	43	.843	0	.000	0	.000	0	.000	-.082	.073
Writing	72	7	.097	13	.181	52	.722	0	.000	0	.000	0	.000	-.113	.107
All Domains	**382**	**18**	**.047**	**48**	**.126**	**286**	**.749**	**3**	**.008**	**0**	**.000**	**27**	**.071**	**-.080**	**.107**
Grades 2-3															
Listening	113	0	.000	12	.106	101	.894	0	.000	0	.000	0	.000	-.071	.078
Reading	123	5	.041	16	.130	102	.829	0	.000	0	.000	0	.000	-.059	.089
Speaking	46	0	.000	0	.000	46	1.000	0	.000	0	.000	0	.000	-.049	.047
Writing	68	0	.000	2	.029	66	.971	0	.000	0	.000	0	.000	-.071	.055
All Domains	**350**	**5**	**.014**	**30**	**.086**	**315**	**.900**	**0**	**.000**	**0**	**.000**	**0**	**.000**	**-.064**	**.075**
Grades 4-5															
Listening	103	3	.029	4	.039	88	.854	6	.058	2	.019	0	.000	-.011	.098
Reading	113	0	.000	1	.009	112	.991	0	.000	0	.000	0	.000	-.017	.056
Speaking	65	1	.015	0	.000	64	.985	0	.000	0	.000	0	.000	-.010	.060
Writing	63	0	.000	1	.016	62	.984	0	.000	0	.000	0	.000	-.056	.051
All Domains	**344**	**4**	**.012**	**6**	**.017**	**326**	**.948**	**6**	**.017**	**2**	**.006**	**0**	**.000**	**-.021**	**.073**
Grade 6-8															
Listening	107	0	.000	3	.028	83	.776	5	.047	16	.150	0	.000	.029	.107
Reading	94	0	.000	0	.000	94	1.000	0	.000	0	.000	0	.000	-.013	.045
Speaking	53	0	.000	0	.000	52	.981	1	.019	0	.000	0	.000	.026	.068
Writing	37	0	.000	1	.027	36	.973	0	.000	0	.000	0	.000	.010	.056
All Domains	**291**	**0**	**.000**	**4**	**.014**	**265**	**.911**	**6**	**.021**	**16**	**.055**	**0**	**.000**	**.012**	**.080**
Grades 9-12															
Listening	102	1	.010	7	.069	77	.755	3	.029	14	.137	0	.000	-.004	.146
Reading	127	0	.000	3	.024	120	.945	1	.008	3	.024	0	.000	-.022	.092
Speaking	53	0	.000	1	.019	50	.943	2	.038	0	.000	0	.000	-.016	.090
Writing	37	0	.000	0	.000	37	1.000	0	.000	0	.000	0	.000	-.027	.080
All Domains	**319**	**1**	**.003**	**11**	**.034**	**284**	**.890**	**6**	**.019**	**17**	**.053**	**0**	**.000**	**-.016**	**.111**
All Grades (K-12)															
Listening	701	9	.013	60	.086	546	.779	25	.036	34	.049	27	.039	-.032	.115
Reading	735	13	.018	40	.054	672	.914	7	.010	3	.004	0	.000	-.036	.088
Speaking	326	1	.003	9	.028	313	.960	3	.009	0	.000	0	.000	-.028	.074
Writing	354	7	.020	21	.059	326	.921	0	.000	0	.000	0	.000	-.053	.084
All Domains	**2116**	**30**	**.014**	**130**	**.061**	**1857**	**.878**	**35**	**.017**	**37**	**.017**	**27**	**.013**	**-.037**	**.096**

Note: Students with disability comprise focal group, no disability is reference. DIF categories with negative sign (i.e., C- or B-) indicate lower item scores for students with disabilities (conditioning on ability); categories with positive sign (i.e., B+ or C+) indicate higher item scores for students with disabilities. N is the number of items assigned to a particular DIF category; P is the proportion of the total number of items. M and SD are the mean and standard deviation of the effect size estimates, respectively.

Figure 1. DIF Effect Size Estimates by Domain and Gradeband.



Note: Students with disabilities were treated as the focal group, while students with no disability were the reference group. Negative effect sizes indicate lower item scores for students with disabilities (conditioning on ability), and positive effect sizes indicate higher item scores for students with disabilities. N is the number of items for which effect sizes were computed (number of items represented in the histogram). M is the mean effect size for items in the gradeband and domain, which is also indicated by the vertical red line (dashed). The vertical blue line (dotted) corresponds to an effect size of 0.

The 67 items assigned to category C (30 C–; 37 C+) were flagged for further investigation by the ELPA21 Administration, Accommodations, and Accessibility (AAA) Task Management Team. This committee convened via web conference on July 26, 2016 to review each item individually and provide a judgement as

to whether any items flagged in this analysis on the basis of the DIF statistics were likely to put students at a disadvantage due to their disability status. A separate report (ELPA21, 2017) describes the details of this review.

In their inspection of the 67 flagged items, the review committee discussed concerns related to the clarity of graphics, completeness of directions, and font size. However, the committee concluded that none of the issues they identified were likely to differentially affect students based on disability status. As a result, all items flagged for the review on the basis of DIF statistics were included in the subsequent calibration.

Part 3: Item Calibrations

Based on the initial item analyses and evaluation of differential item functioning (and the exclusion of one item following identification of a factual error), 2,113 of the 2,116 on-gradeband items administered on the online test forms were included in the item response theory (IRT) calibrations.

Model Estimation

All calibrations were performed using FlexMIRT® version 3 (Cai, 2016). Model parameters were estimated using full-information maximum marginal likelihood estimation via Bock and Aitkin's (1981) expectation maximization (EM) algorithm. The M-step convergence criterion was 1×10^{-6}. E-step cycles terminated when the maximum absolute difference in parameters fell below 1×10^{-4} for adjacent E-step cycles. We used 31 quadrature points equally spaced from –6 to +6 (on the logit scale) per dimension. Analytical dimension reduction (see Gibbons et al., 2007; Rijmen, 2009; Cai, Yang, & Hansen, 2012) implemented in FlexMIRT was used in order to improve the efficiency of estimation of each multidimensional model. Standard errors were calculated by FlexMIRT® via the Richardson extrapolation method (Houts & Cai, 2015; Jamshidian & Jennrich, 2000). All calibrations met the termination criterion and were found to have converged to a stable solution (first- and second-order tests; see Houts & Cai, 2016).

Item Factor Analysis Models

Separate calibrations were performed within each gradeband for three item factor analysis models. These models differed with respect to the primary dimensions predicting item scores. The first model was specified with four correlated dimensions, with each dimension corresponding to one of the four domains subtests: Listening, Reading, Speaking, and Writing. The second model was a restricted hierarchical item factor analysis model with a single dimension underlying performance across the full test (i.e., overall English language proficiency). Four additional dimensions represented the domain-specific variation not explained by the general/overall factor. The third model was also a restricted hierarchical item factor analysis but was fitted to the items from the Listening and Reading subtests only. In this model, the primary dimension represents English language comprehension (or reception) skills. Two additional dimensions capture the listening- and reading-specific variation not explained by the comprehension factor.

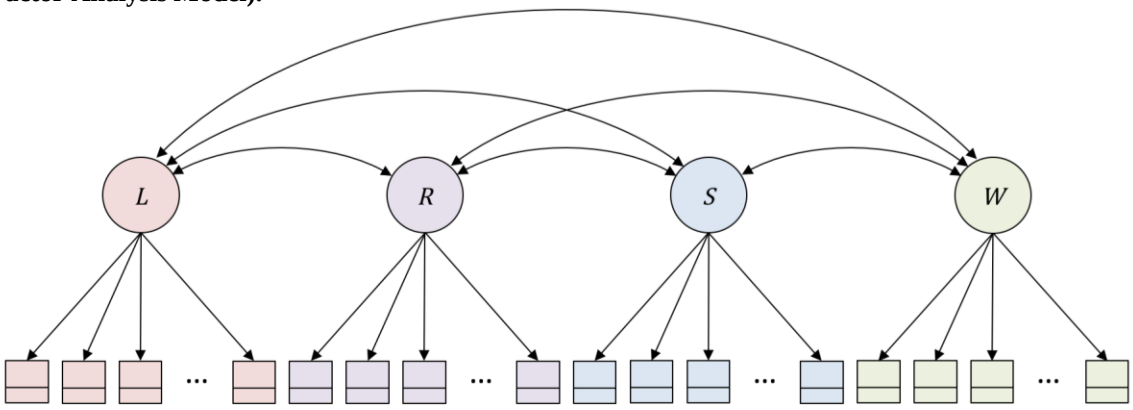
Calibration of Model 1: An Independent Clusters Item Factor Analysis Model for Proficiency in Four Language Domains

Figure 2 illustrates the structure of the four-dimensional model used to obtain parameters relating item performance to standing on the four language domains (Listening, Reading, Speaking, and Writing). This is the primary scoring model for obtaining estimates of performance within domains (the basis of classification of students into performance levels and overall proficiency determinations). For the current pool of operational items, each item was assumed to load on exactly one domain (that is, items have an "independent cluster" structure; McDonald, 2000), and the dimensions of the model are correlated.

Due to the large number of students (Table 1) and items (Table 4) within each gradeband, this analysis was performed in two steps. First, we calibrated the items with respect to their respective domains (i.e., fitting a unidimensional model to the item data for each domain subtest). This provided estimates of items' slope and threshold parameters. Second, we used a restricted hierarchical model (specifically, a testlet response model;

Wainer, Bradlow, & Wang, 2007) in order to estimate the correlations between the four domains (Thissen, 2013).

Figure 2. Path Diagram for Calibrating Items with Respect to Domains (Independent Clusters Item Factor Analysis Model).



Note: Circles represent latent variables (factors), squares represent manifest variables (test items), single-headed arrows represent regression paths (of items onto the latent variables), and double-headed arrows indicate covariances between latent variables.

The probability of student i achieving item score $c = \{0,1,\dots,C-1\}$ for item $j = \{1,2,\dots,J\}$, given the student's standing on the latent dimension underlying performance in assessed domain $d = \{\text{Listening, Reading, Speaking, Writing}\}$, is represented by $\mathrm{P}(y_{ijd} = c|\theta_d)$. Items were skipped, omitted, or not reached were assigned the minimum item score ($y_{ijd} = 0$), per ELPA21 scoring rules.

For all items, the probability $\mathrm{P}(y_{ijd} = c|\theta_d)$ was modeled using the logistic graded response model (Samejima, 1969). If an item has *C* categories, then the possible item scores are $c = \{0,1,\dots,C-1\}$. The following logistic functions describe the cumulative probability of achieving an item score of c or greater (that is, $y_{ijd} \geq c$):

$$\mathrm{P}(y_{ijd} \geq 0|\theta_d) = 1$$

$$\mathrm{P}(y_{ijd} \geq 1|\theta_d) = \frac{1}{1 + exp(-[a_{jd}\theta_d + d_{j1}])}$$
$$\vdots$$
$$\mathrm{P}(y_{ijd} \geq c|\theta_d) = \frac{1}{1 + exp(-[a_{jd}\theta_d + d_{jc}])}$$
$$\vdots$$
$$\mathrm{P}(y_{ijd} \geq C-1|\theta_d) = \frac{1}{1 + exp(-[a_{jd}\theta_d + d_{j(C-1)}])},$$

where a_{jd} is a slope (or discrimination) parameter $d_{j1},\dots,d_{jc},\dots,d_{j(C-1)}$ comprise a set of ordered intercept parameters. The probability of achieving score can then be defined as the difference of adjacent cumulative probabilities:

$$\mathrm{P}(y_{ijd} = c|\theta_d) = \mathrm{P}(y_{ijd} \geq c|\theta_d) - \mathrm{P}(y_{ijd} \geq c+1|\theta_d).$$

Finally, the boundary cases of $c = 0$ and $c = C-1$ can be defined as

$$\mathrm{P}(y_{ijd} = 0|\theta_d) = 1 - \mathrm{P}(y_{ijd} \geq 1|\theta_d)$$

and

$$\mathrm{P}(y_{ijd} = C-1|\theta_d) = \mathrm{P}(y_{ijd} \geq C-1|\theta_d).$$

Note that for items with $C = 2$, the model is equivalent to the two-parameter logistic (2PL) model. All latent domain proficiency variables were assumed to have a standard normal distribution, and item parameters (slopes and thresholds) were estimated freely fitting unidimensional models within domain.

In order to estimate the correlations between the domains, we fit a hierarchical model in which each test item loaded on the composite/general dimension (θ_g) and exactly one domain-specific dimension (θ_{d*}). In this model, the slope parameters for the two dimensions were constrained to be equal.

The probability of student i achieving item score $c = \{0,1,\dots,C-1\}$ for item $j = \{1,2,\dots,J\}$, given the student's standing on the general and domain-specific dimensions is represented by $\mathrm{P}(y_{ijd} = c|\theta_g, \theta_{d*})$. For all items, the probability $\mathrm{P}(y_{ijd} = c|\theta_g, \theta_{d*})$ was modeled using the multidimensional extension of the logistic graded response model (Samejima, 1969; Gibbons et al., 2007; Reckase, 2009).

If the item has *C* categories, then the possible item scores are $c = \{0,1,\dots,C-1\}$. The following logistic functions describe the cumulative probability of achieving an item score of c or greater (that is, $y_{ijd} \geq c$):

$$\mathrm{P}(y_{ijd} \geq 0|\theta_g, \theta_{d*}) = 1$$

$$\mathrm{P}(y_{ijd} \geq 1|\theta_g, \theta_{d*}) = \frac{1}{1 + exp(-[a_j\theta_g + a_j\theta_{d*} + d_{j1}])}$$

$$\vdots$$

$$\mathrm{P}(y_{ijd} \geq c|\theta_g, \theta_{d*}) = \frac{1}{1 + exp(-[a_j\theta_g + a_j\theta_{d*} + d_{jc}])}$$

$$\vdots$$

$$\mathrm{P}(y_{ijd} \geq C-1|\theta_g, \theta_{d*}) = \frac{1}{1 + exp(-[a_{jg}\theta_g + a_j\theta_{d*} + d_{j(C-1)}])},$$

where a_j is the common value for the slopes on the general and domain-specific dimensions, and $d_{j1}, \dots, d_{jc}, \dots, d_{j(C-1)}$ comprise a set of ordered intercept parameters. The probability of achieving score can then be defined as the difference of adjacent cumulative probabilities:

$$\mathrm{P}(y_{ijd} = c|\theta_g, \theta_{d*}) = \mathrm{P}(y_{ijd} \geq c|\theta_g, \theta_{d*}) - \mathrm{P}(y_{ijd} \geq c+1|\theta_g, \theta_{d*}).$$

Finally, the boundary cases of $c = 0$ and $c = C - 1$ can be defined as

$$\mathrm{P}(y_{ijd} = 0|\theta_g, \theta_{d*}) = 1 - \mathrm{P}(y_{ijd} \geq 1|\theta_g, \theta_{d*})$$

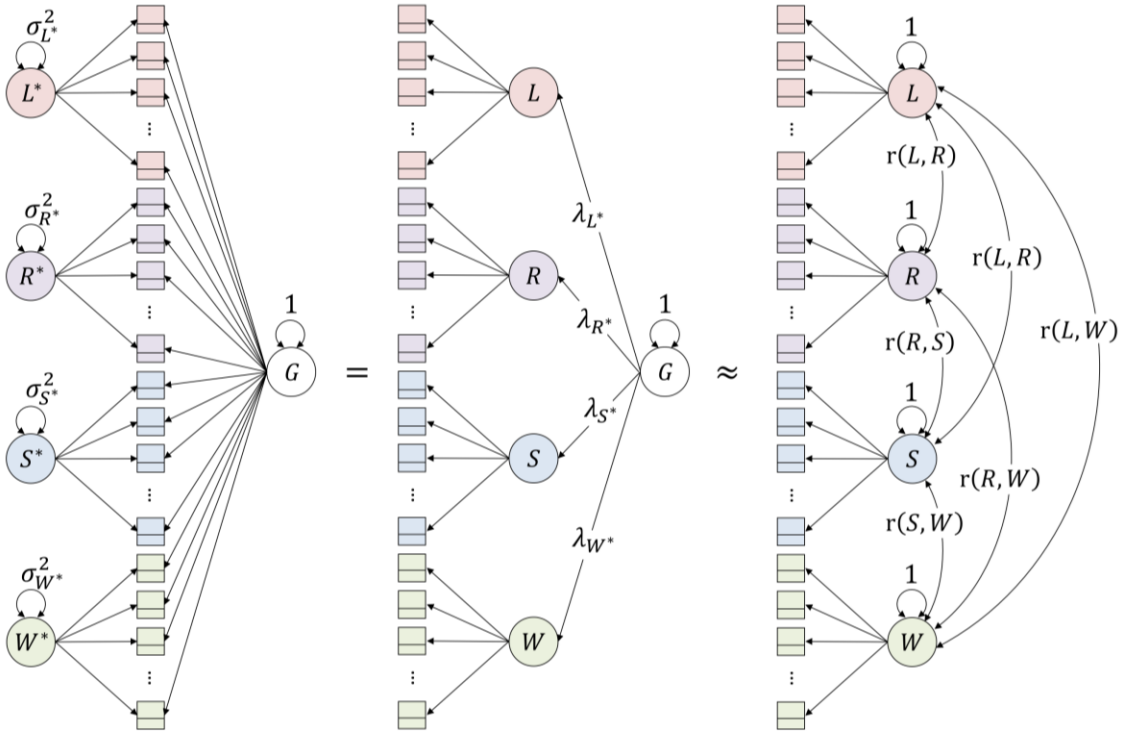
and

$$\mathrm{P}(y_{ijd} = C-1|\theta_g, \theta_{d*}) = \mathrm{P}(y_{ijd} \geq C-1|\theta_g, \theta_{d*}).$$

Note that for items with $C = 2$, the model is equivalent to the multidimensional extension of a two-parameter logistic model (Reckase, 2009).

The primary dimension was assumed to have a standard normal distribution. The domain-specific dimensions were assumed to have normal distribution with mean of zero. The four variances of the domain-specific factors were freely estimated, along with the item parameters (slopes and thresholds). This constrained hierarchical model—with equal slopes on the general and domain-specific factors and orthogonal latent variables—is a version of the test response model (Wainer, Bradlow, & Wang, 2007) and is isomorphic to a second-order item factor analysis model (see, e.g., Yung, McLeod, & Thissen, 1999; Rijmen, 2010). The hierarchical and higher-order models, in turn, approximate an independent clusters model. Figure 3 shows the relationships among these models.

Figure 3. Relationships Among the Hierarchical (Left), Second-Order (Middle), and Independent Clusters (Right) Item Factor Analysis Models.



Note: Circles represent latent variables (factors), squares represent manifest variables (test items), single-headed arrows represent regression paths (of items onto the latent variables), and double-headed arrows indicate latent variable variances and covariances. Adapted from Figure 1 in Thissen (2013).

Thissen (2013) proposed that the hierarchical model (left panel of Figure 3) could be used to approximate the parameters of the independent clusters model (right panel). This is desirable because the parameters of the full independent clusters model could not be estimated using fixed quadrature without greatly reducing the number of quadrature points. In contrast, the hierarchical model could be estimated by applying analytical dimension reduction (Gibbons et al, 2007; Rijmen, 2009; Cai, Yang, and Hansen, 2011). The approximation would also tend to be accurate enough for operational purposes because the fitted approximating model is very nearly the same (only 2 degrees of freedom different) as the desired independent clusters model.

By fitting the hierarchical model (left panel of Figure 3), we obtained estimates of the domain-specific variances ($\sigma^2_{L^*}$, $\sigma^2_{R^*}$, $\sigma^2_{S^*}$, and $\sigma^2_{W^*}$). From these variances, we computed (applying Thissen's [2013] Equation 1) loadings of the domain factors onto a second-order dimension (λ_{L^*}, λ_{R^*}, λ_{L^*}, and λ_{W^*}) in the equivalent higher-order model (middle panel):

$$\boldsymbol{\lambda}_g = \begin{bmatrix} \lambda_{L^*} \\ \lambda_{R^*} \\ \lambda_{S^*} \\ \lambda_{W^*} \end{bmatrix} = \begin{bmatrix} \left(1 + \sigma^2_{L^*}\right)^{-1/2} \\ \left(1 + \sigma^2_{R^*}\right)^{-1/2} \\ \left(1 + \sigma^2_{S^*}\right)^{-1/2} \\ \left(1 + \sigma^2_{W^*}\right)^{-1/2} \end{bmatrix}$$

These loadings, in turn, imply a correlation structure for the latent variable in the independent clusters model, shown in the right panel of Figure 3 (see Thissen's [2013] Equation 2):

$$\boldsymbol{R} = \boldsymbol{\lambda}_g \boldsymbol{\lambda}'_g + \left[\mathbf{I} - \mathrm{diag}\left(\boldsymbol{\lambda}_g \boldsymbol{\lambda}'_g\right)\right].$$

Table 10 shows the domain-specific variance estimates and the corresponding (model-implied) correlation matrices.

Table 10. Hierarchical (Testlet Response) Model Factor Variances and Implied Correlation Matrix for the Independent Clusters Model.

Domain	Hierarchical (Testlet Response) Model					Independent Clusters Model			
	G	L*	R*	S*	W*	L	R	S	W
Grade K									
General (G)	1.000								
Listening (L)	.000	.118				1.000			
Reading (R)	.000	.000	.214			.858	1.000		
Speaking (S)	.000	.000	.000	1.339		.618	.593	1.000	
Writing (W)	.000	.000	.000	.000	1.461	.603	.578	.417	1.000
Grade 1									
General (G)	1.000								
Listening (L)	.000	1.036				1.000			
Reading (R)	.000	.000	.191			.642	1.000		
Speaking (S)	.000	.000	.000	1.264		.466	.609	1.000	
Writing (W)	.000	.000	.000	.000	.163	.650	.850	.616	1.000
Grades 2-3									
General (G)	1.000								
Listening (L)	.000	.559				1.000			
Reading (R)	.000	.000	.090			.767	1.000		
Speaking (S)	.000	.000	.000	.802		.597	.714	1.000	
Writing (W)	.000	.000	.000	.000	.128	.754	.902	.702	1.000
Grades 4-5									
General (G)	1.000								
Listening (L)	.000	.275				1.000			
Reading (R)	.000	.000	.160			.822	1.000		
Speaking (S)	.000	.000	.000	.700		.679	.712	1.000	
Writing (W)	.000	.000	.000	.000	.108	.841	.882	.729	1.000
Grades 6-8									
General (G)	1.000								
Listening (L)	.000	.159				1.000			
Reading (R)	.000	.000	.214			.843	1.000		
Speaking (S)	.000	.000	.000	.694		.714	.697	1.000	
Writing (W)	.000	.000	.000	.000	.157	.864	.844	.714	1.000
Grades 9-12									
General (G)	1.000								
Listening (L)	.000	.083				1.000			
Reading (R)	.000	.000	.160			.892	1.000		
Speaking (S)	.000	.000	.000	.861		.704	.681	1.000	
Writing (W)	.000	.000	.000	.000	.198	.878	.848	.670	1.000

Notes: In these hierarchical models, slopes on the general and domain-specific factors were constrained to be equal, general factor variances were fixed to 1, all factor covariances were fixed to 0, and domain-specific factor variances were freely estimated. Correlations were computed using the estimated domain-specific factor variances, following Thissen (2013).

The estimated correlations between domains were strong but variable, ranging from .417 (between Kindergarten Speaking and Writing) to .902 (Grade 2-3 Reading and Writing). The correlation matrices shown on the right side of Table 10 were used in specifying the multivariate normal priors used in computing the Domain scale scores and error covariances (ELPA21, 2016b). Specifically, the population distribution of $\boldsymbol{\theta} = (\theta_L, \theta_R, \theta_S, \theta_W)^t$ for examinees within a gradeband is assumed to be multivariate normal with means of zero and covariance $\boldsymbol{R}$. Item parameters (slopes and intercepts) obtained from Model 1 can be provided upon request.

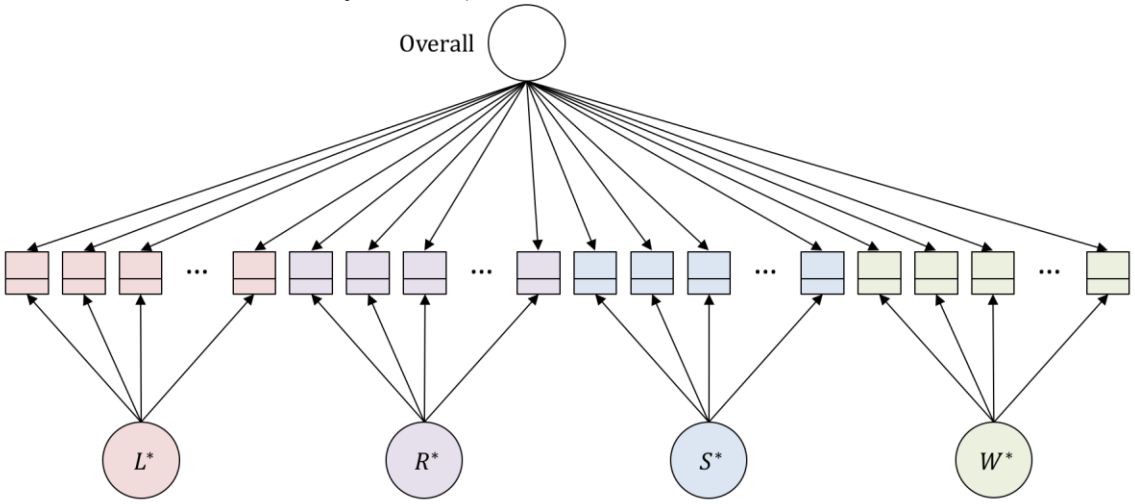
Calibration of Model 2: A Hierarchical Item Factor Analysis Model for Overall English Language Proficiency

Although the results obtained from Model 1 are the primary basis for student-level reporting and decision-making (providing characterization of performance level by domain and an overall proficiency determination), there are contexts (e.g., in program evaluation and accountability reporting) in which it may be useful to have a single-number summary of performance across the four domains. Figure 3 illustrates the structure of the five-dimensional hierarchical model used to obtain parameters relating item performance to standing on an Overall dimension (based on performance across all domains), with additional factors to explain domain-specific variations in performance.

This model closely resembles the hierarchical model shown in Figure 3 that was used to obtain estimates of the correlation structure for the independent clusters model (Model 1). Like that model, the slopes on the general and domain-specific factors were constrained equal, the means of all the latent variables were fixed to zero, the variance of the general factor was fixed to one, and the domain-specific variances were estimated.

Unlike the hierarchical model in the previous section, Model 2 is further constrained to require the domain-specific variances be equal. This constraint ensures that the general factor is not dominated by one or more domains but instead represents an average giving equal weight to the domains.

Figure 4. Path Diagram for Calibrating Items with Respect to Overall English Language Proficiency (Hierarchical Item Factor Analysis Model).



Note: Circles represent latent variables (factors), squares represent manifest variables (test items), single-headed arrows represent regression paths (of items onto the latent variables).

Table 11 provides the estimated factor variances obtained from this model. These variances were used in specifying the multivariate normal priors used in computing the Overall scale score and standard error (ELPA21, 2016b). Specifically, the population distribution of $\boldsymbol{\theta} = (\theta_{Overall}, \theta_{L^*}, \theta_{R^*}, \theta_{S^*}, \theta_{W^*})^t$ for examinees within a gradeband is assumed to be multivariate normal with means of zero and covariance matrix with all off-diagonal elements equal to zero and diagonal elements of $\left(1, \sigma^2_{d^*}, \sigma^2_{d^*}, \sigma^2_{d^*}, \sigma^2_{d^*}\right)$, where $\sigma^2_{d^*}$ is the

common variance estimate for the four domain-specific factors. Item parameter estimates (slope and intercepts) for Model 2 can be provided upon request.

Table 11. Overall English Language Proficiency: Estimated Common Variances for the Domain-specific Factors

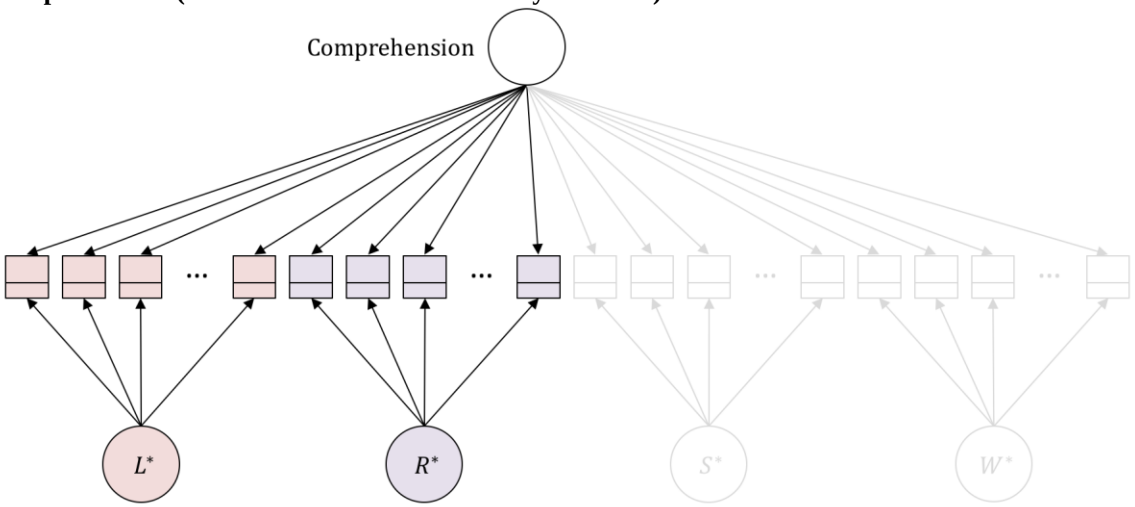
	O	L*	R*	S*	W*
		Grade K			
Overall (O)	1.000				
Listening (L*)	.000	.633			
Reading (R*)	.000	.000	.633		
Speaking (S*)	.000	.000	.000	.633	
Writing (W*)	.000	.000	.000	.000	.633
		Grade 1			
Overall (O)	1.000				
Listening (L*)	.000	.514			
Reading (R*)	.000	.000	.514		
Speaking (S*)	.000	.000	.000	.514	
Writing (W*)	.000	.000	.000	.000	.514
		Grades 2-3			
Overall (O)	1.000				
Listening (L*)	.000	.309			
Reading (R*)	.000	.000	.309		
Speaking (S*)	.000	.000	.000	.309	
Writing (W*)	.000	.000	.000	.000	.309
		Grades 4-5			
Overall (O)	1.000				
Listening (L*)	.000	.283			
Reading (R*)	.000	.000	.283		
Speaking (S*)	.000	.000	.000	.283	
Writing (W*)	.000	.000	.000	.000	.283
		Grades 6-8			
Overall (O)	1.000				
Listening (L*)	.000	.290			
Reading (R*)	.000	.000	.290		
Speaking (S*)	.000	.000	.000	.290	
Writing (W*)	.000	.000	.000	.000	.290
		Grades 9-12			
Overall (O)	1.000				
Listening (L*)	.000	.296			
Reading (R*)	.000	.000	.296		
Speaking (S*)	.000	.000	.000	.296	
Writing (W*)	.000	.000	.000	.000	.296

Calibration of Model 3: A Hierarchical Item Factor Analysis Model for Proficiency in English Language Comprehension

Test users may also wish to have a means of quantifying students' ability with respect to English language comprehension (or reception). Figure 5 illustrates the structure of the three-dimensional hierarchical model used to obtain parameters relating item performance to standing on a Comprehension dimension, with

additional factors to explain domain-specific variations in performance. Note that unlike Models 1 and 2, the model for Comprehension uses items from the Listening and Reading domains only).

Figure 5. Path Diagram for Calibrating Items with Respect to Proficiency in English Language Comprehension (Hierarchical Item Factor Analysis Model).



Note: Circles represent latent variables (factors), squares represent manifest variables (test items), single-headed arrows represent regression paths (of items onto the latent variables). Speaking and Writing items are excluded from the calibration.

The model for Comprehension was specified in the same manner as the model for Overall English Language Proficiency. Specifically, slopes on the general and domain-specific factors were constrained equal, the means of the latent variables were fixed to zero, the variance of the general factor was fixed to one, and a common domain-specific variance was estimated.

Table 12 provides the estimated factor variances obtained from this model. These variances were used in specifying the multivariate normal priors used in computing the Comprehension scale score and standard error (ELPA21, 2016b). Specifically, the population distribution of $\boldsymbol{\theta} = \left(\theta_{Comprehension}, \theta_{L^*}, \theta_{R^*}\right)^t$ for examinees within a gradeband is assumed to be multivariate normal with means of zero and covariance matrix with all off-diagonal elements equal to zero and diagonal elements of $\left(1, \sigma^2_{d^*}, \sigma^2_{d^*}\right)$, where $\sigma^2_{d^*}$ is the common variance estimate for the two domain-specific factors. Item parameter estimates (slope and intercepts) for Model 3 can be provided upon request.

Calibration of Paper-only Items

A total of 2,113 items were calibrated using data obtained from the online administration of the 2015-16 summative assessments. The parameters obtained from these calibrations were used in scoring both online and paper versions of the assessments. However, there was one item (VH197897) appearing on the paper form for Kindergarten Writing that was not administered in any online form. In order to calibrate this item, we utilized a sample of 566 students who completed the paper form for this gradeband.

The calibrations were performed by fixing the item parameters to their prior estimates for all but the previously uncalibrated item. Because the population of students administered paper forms would be expected to differ from the general population of examinees, the population parameters (means and variances) were freely estimated, along with the parameters of the paper-only item. Separate calibration runs were performed in order to obtain item parameters for Models 1 (Domains) and 2 (Overall English Language Proficiency). This brought the total number of items calibrated from the 2015-16 summative assessment administration to 2,114.

Table 12. Proficiency in English Language Comprehension: Estimated Common Variances for the Domain-specific Factors

	C	L*	R*
Grade K			
Overall (O)	1.000		
Listening (L*)	.000	.137	
Reading (R*)	.000	.000	.137
Grade 1			
Overall (O)	1.000		
Listening (L*)	.000	.521	
Reading (R*)	.000	.000	.521
Grades 2-3			
Overall (O)	1.000		
Listening (L*)	.000	.274	
Reading (R*)	.000	.000	.274
Grades 4-5			
Overall (O)	1.000		
Listening (L*)	.000	.172	
Reading (R*)	.000	.000	.172
Grades 6-8			
Overall (O)	1.000		
Listening (L*)	.000	.137	
Reading (R*)	.000	.000	.137
Grades 9-12			
Overall (O)	1.000		
Listening (L*)	.000	.090	
Reading (R*)	.000	.000	.090

Part 4: Cross-gradeband Analyses

Additional IRT calibrations were performed in order to examine the relationships of domain performance across gradebands. These analyses utilized a small number of non-operational items administered in the gradeband above their intended level (e.g., Kindergarten items embedded in the Grade 1 test forms).

Table 13 shows the number of off-gradeband items that were embedded, by grade level and domain subtest. These items were distributed across the five or six alternative test forms, so the actual number administered per student was smaller.

Table 13. Number of Off-gradeband Items Embedded Within 2015-16 Summative Assessment Forms.

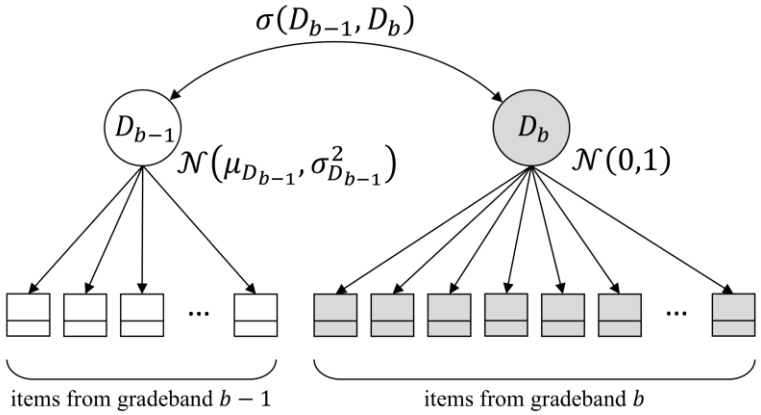
Gradeband		Domain			
of Test Forms	of Embedded Items	L	R	S	W
Grade 1	Kindergarten	33	38	11	28
Grades 2-3	Grade 1	39	34	13	14
Grades 4-5	Grades 2-3	36	30	17	20
Grades 6-8	Grades 4-5	27	30	18	15
Grades 9-12	Grades 6-8	39	26	13	11

Items were selected to represent the range of tasks administered in the gradeband. As in the operational portions of the test forms, the number of items was generally higher for the Listening and Reading Domains. This is largely due to the fact that Speaking and Writing items generally require substantially more time to administer. In addition, many Speaking items are scored as clusters: responses to multiple items in a set are scored using a rubric that is applied to the collected responses. The counts in Table 13 are based on the number of distinct scored items (with cluster-scored sets counted as one item).

Figure 6 illustrates the model used to estimate the relationships between adjacent gradebands. In each model, there were three parameters that were estimated:

- $\mu_{D_{b-1}}$, the mean ability of students in gradeband b on the dimension underlying items from gradeband $b-1$ (i.e., one gradeband below the students' actual gradeband) for domain D;
- $\sigma^2_{D_{b-1}}$, the variance of ability of students in gradeband b on the dimension underlying items from gradeband $b-1$ for domain D (i.e., one gradeband below the students' actual gradeband; and
- $\sigma(D_{b-1}, D_b)$, the covariance (among students in gradeband b between the dimensions underlying (a) items from gradeband $b-1$ for domain D and (b) items from gradeband b for domain D.

Figure 6. Model for Cross-gradeband Calibrations.



All item parameters were fixed to the estimates obtained from Model 1 (Independent Clusters), thus imposing an assumption of measurement invariance across the adjacent gradebands. The mean and variance of the ability students in gradeband b on the dimension underlying items from gradeband b were fixed to zero and one respectively—exactly as was done in the Model 1 calibrations.

Note that the assumption of measurement invariance is not the same as assuming a vertical scale, only that the relationship between items in gradeband $b-1$ and the construct of gradeband $b-1$ domain D (e.g., Grade 1 Listening) is the same for students in gradeband $b-1$ as for gradeband b. We assume that domain as measured in one gradeband is distinct from (though related to) the same domain (in name) as measured in the adjacent gradeband.

Estimated parameters from the series of cross-gradeband calibrations are provided in Table 14. The domain scores were generally found to correlate strongly across gradebands (above 0.93, on average). In addition, performance of students in the higher gradeband was generally found to be higher on average than the performance of students in the lower gradeband.[3] These differences were largest in the lowest gradebands, however, perhaps due to the fact that higher ability students exit the population of examinees. Note that students in Grades 9-12 were actually found to perform slightly worse than Students in Grades 6-8 on the Grade 6-8 Listening.

[3] Note that, per the assumption imposed in the calibrations of Model 1, on-gradeband performance has a mean of zero and standard deviation of one. Thus, the values in the Mean column of Table 14 may be interpreted as group differences in units based on the standard deviation for students in the lower gradeband.

Table 14. Parameter Estimates from Cross-gradeband Analyses.

Domain	Mean $\mu_{D_{b-1}}$	Variance $\sigma^2_{D_{b-1}}$	Covariance $\sigma(D_{b-1}, D_b)$	Correlation $r(D_{b-1}, D_b)$
Students in Grade 1 – Performance on Items from Kindergarten				
Listening	1.031	1.950	1.342	.961
Reading	1.111	2.168	1.084	.737
Speaking	.673	.997	.933	.934
Writing	1.305	1.067	.990	.959
Students in Grades 2-3 – Performance on Items from Grade 1				
Listening	.718	1.494	1.114	.912
Reading	1.460	2.253	1.459	.972
Speaking	.810	1.592	1.193	.946
Writing	.797	1.137	.965	.905
Students in Grades 4-5 – Performance on Items from Grades 2-3				
Listening	.935	2.072	1.192	.828
Reading	.892	1.694	1.197	.920
Speaking	.387	1.099	.997	.951
Writing	.861	1.232	1.059	.954
Students in Grades 6-8 – Performance on Items from Grades 4-5				
Listening	.373	1.672	1.222	.945
Reading	.620	1.890	1.309	.952
Speaking	.051	1.406	1.164	.982
Writing	.560	1.472	1.144	.943
Students in Grades 9-12 – Performance on Items from Grades 6-8				
Listening	.031	1.460	1.177	.974
Reading	.288	1.526	1.207	.977
Speaking	-.065	1.538	1.217	.981
Writing	.185	1.440	1.176	.980

Aside from providing insights into the relationships between the skills measured by the items in each gradeband, the results of these analyses can facilitate the computation of score projections across gradebands. That is, one could use the parameter estimates to translate performance within one gradeband to an expected level of performance on the adjacent gradeband (Thissen, Liu, Magnus, & Quinn, 2015). This method was used to evaluate the ordering of cut scores during the ELPA21 standard setting (Pacific Metrics, 2017).

Summary

This report describes analyses of the ELPA21 item pools. The vast majority of items were included in the final operational pools, with a total of 2,114 items calibrated and only two items excluded (two due to poor statistics and one due to a factual error in the stimulus text).

Calibrations were performed by applying a series of item response theory models to the scored item response data within each gradeband, all using the logistic graded response model and its multidimensional extensions.

The primary scoring model for ELPA21 is an independent clusters model with four underlying dimensions. Parameters for this model were estimated in two steps. First, the item parameters (slopes and intercepts) were obtained by fitting unidimensional models within domain. Second, the correlations between domains were estimated by fitting a hierarchical model. Implied correlations can be computed from the-specific variance estimates from this model (thus providing much more computationally feasible approach for obtaining estimates of these correlations than directly fitting the independent clusters model).

Secondary models were also estimated to obtain parameters for generating scores for Overall English Language Proficiency and Comprehension. These calibrations used highly restricted hierarchical models such that the general factors can be interpreted as cross-domain composite in which the underlying domains are equally weighted.

Item parameters for the 2,114 items can be provided upon request. Parameters were produced for each of the three scoring models: Domains (independent clusters item factor analysis model), Overall English Language Proficiency (hierarchical item factor analysis model), and Proficiency in English Language Comprehension (hierarchical item factor analysis model, for Listening and Reading items only). Structural parameters used to specify the priors used in scoring were provided in Tables 10-12.

The parameters obtained in the calibrations described here were used to produce individual score results for both the 2015-16 and 2016-17 summative assessments. The same item pools will continue to be used in 2017-18 (both for the summative assessments and for the operational screener, beginning in August 2017). Newly developed items will be embedded in field test slots in the 2017-18 summative assessment. The procedures and models applied here will again be used to evaluate and calibrate these new items.

References

Cai, L. (2016). FlexMIRT®: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

Cai, L., Yang, J., & Hansen, M. (2011). Generalized full information item bifactor analysis. *Psychological Methods, 16*, 221–248. doi:10.1037/a0023350

English Language Proficiency for the 21st Century (2016a). *Score reporting specifications: School year 2015–2016 summative assessment: Grades K-12.* Washington, DC: Council of Chief State School Officers.

English Language Proficiency for the 21st Century (2016b). *ELPA21 Scoring specifications: School year 2015–2016*. Washington, DC: Council of Chief State School Officers.

English Language Proficiency for the 21st Century (2017). *ELPA21 AAA TMT review of items showing differential item functioning for English Language Learners with disabilities.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*(1), 4–19. doi:10.1177/0146621606289485

Han, K. C. T., & Paek, I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement, 38*(6), 486–498. doi:10.1177/0146621614536770

Houts, C. R., & Cai, L. (2015). FlexMIRT®*: Flexible multilevel multidimensional item analysis and test scoring. User's manual version 3.0RC*. Seattle, WA: Vector Psychometric Group, LLC.

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

Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society: Series B, 62*, 257–270. doi:10.1111/1467-9868.00230

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*(2), 99–114. doi:10.1177/01466210022031552

Michaelides, M. P. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment Research & Evaluation, 13*(7).

Pacific Metrics (2017). *ELPA21 standard setting technical report.* Monterey, CA: Author.

Questar Assessment (2016). *ELPA21 technical report: Spring 2015 field test*. Apple Valley, MN: Author.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Rijmen, F. (2009). *Efficient full information maximum likelihood estimation for multidimensional IRT models* (Tech. Rep. No. RR-09-03). Princeton, NJ: Educational Testing Service.

Rijmen, F. (2010). Formal relations and an empirical comparison between the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*, 361–372. doi:361–372. doi:10.1111/j.1745-3984.2010.00118.x

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monographs No. 17). Richmond, VA: Psychometric Society.

Thissen, D. (2013). Using the testlet response model as a shortcut to multidimensional item response theory subscore computation. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, C. M. Woods (Eds.), *New Developments in Quantitative Psychology* (pp. 29–40). New York: Springer.

Thissen, D., Liu, Y., Magnus, B., & Quinn, H. (2015). Extending the use of multidimensional IRT calibration as projection: Many-to-one linking and linear computation of projected scores. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, and S.-M. Chow (Eds.), *Quantitative Psychology Research* (pp. 1–16). New York: Springer.

Thissen, D., & Wainer, H. (2001). *Test scoring.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., Bradlow, E., & Wang, X. (2007). *Testlet response theory and its applications.* New York: Cambridge University Press.

Yung, Y. F., McLeod, L. D., & Thissen, D. (1999). The development of hierarchical factor solutions. *Psychometrika, 64*, 113–128.

Yung, Y. F., McLeod, L. D., & Thissen, D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika 64*(2), pp 113–128. doi: 10.1007/BF02294531

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233–251. doi:10.1111/j.1745-3984.1993.tb00425.x

Appendix A. Distribution of Operational Tasks in 2015-16 Summative Assessment

This appendix describes the number of each task type to be included in each test form. Separate tables are provided for each gradeband. Note that only operational tasks (those contributing to students' scores) are listed; non-operational tasks (embedded field test items and off-grade items used in cross-gradeband analyses) are excluded.

Table A1. Summative Assessment Blueprint, Grade K

Slot	Task Type	# Tasks/Form	Response	# Items/Task
	Listening			
1	Listen and Match: Word	3	TE	1
2	Listen and Match: Phrase	3	TE	1
3	Listen and Match: Sentence	3	TE	1
4	Follow Instructions	2	TE	3, 4, or 5
5	Short Conversation	1	TE	1
6	Long Conversation	1	SR,TE	2 or 3
7	Read-Aloud Story	1	SR,TE	3
8	Teacher Presentation	1	SR,TE	3
	Reading			
1	Read and Match: Word	3	TE	1
2	Read and Match: Phrase	3	TE	1
3	Read and Match: Sentence	3	TE	1
4	Word Wall	1	SR,TE	5
5	Read-Along Story	1	SR,TE	3
6	Short Correspondence	1	SR,TE	3
7	Informational Set	1	SR,TE	3
	Speaking			
1	Classroom Tableau	1	CR	6
2	Show and Share Questions	1	CR	2
3	Show and Share Presentation	2	CR	4
4	Picture Description	1	CR	5
5	Observe and Report	1	CR	4
	Writing			
1	Word Builder: Word	2	TE	1
2	Word Builder: Phrase	2	TE	1
3	Word Builder: Sentence	1	TE	1
4	Sentence Builder	2	TE	1
5	Complete the Story	1	TE	2
	Writing Supplement (Paper)			
1	Copy a Word	1	CR	1
2	Complete a Word	1	CR	1
3	Write a Word	1	CR	1
4	Write a Sentence	1	CR	1
5	Opinion	1	CR	1

Note: CR = Constructed Response (audio response, text response), SR = Selected Response (multiple-choice single selection, multiple-choice multiple selection), and TE = Technology-Enhanced (match single selection, match multiple selection, zone single selection, zone multiple selection)

Appendix A. Distribution of Operational Tasks in 2015-16 Summative Assessment

Table A2. Summative Assessment Blueprint, Grade 1

Slot	Task Type	# Tasks/Form	Response	# Items/Task
	Listening			
1	Listen and Match: Word	4	TE	1
2	Listen and Match: Sentence	4	TE	1
3	Follow Instructions	1	TE	4 or 5
4	Short Conversation	2	TE	1
5	Long Conversation	1	TE	3
6	Read-Aloud Story	1	SR,TE	4
7	Teacher Presentation	1	SR,TE	2 or 3
	Reading			
1	Read-Along Sentence	4	TE	1
2	Read and Match: Word	4	SR	1
3	Read and Match: Sentence	6	SR	1
4	Read for Details	2	TE	1
5	Short Correspondence	2	SR,TE	2
6	Procedural Text	1	SR,TE	3 or 4
7	Literary Set	1	SR,TE	3 or 4
8	Informational Set	1	SR,TE	3 or 4
	Speaking			
1	Classroom Tableau	1	CR	5
2	Conversation	1	CR	3
3	Picture Description	1	CR	1
4	Opinion	1	CR	2
5	Observe and Report	1	CR	1
	Writing			
1	Word Builder: Word	2	TE	1
2	Word Builder: Sentence	3	TE	1
3	Sentence Builder	5	TE	1
	Writing Supplement (Paper)			
1	Copy a Word	1	CR	1
2	Write a Word	1	CR	1
3	Write a Sentence	1	CR	1
4	Storyboard	1	CR	1

Note: CR = Constructed Response (audio response, text response), SR = Selected Response (multiple-choice single selection, multiple-choice multiple selection), and TE = Technology-Enhanced (match single selection, match multiple selection, zone single selection, zone multiple selection)

Table A3. Summative Assessment Blueprint, Grades 2-3

Slot	Task Type	# Tasks/Form	Response	# Items/Task
	Listening			
1	Listen and Match	4	TE	1
2	Listen and Match	4	TE	1
3	Follow Instructions	1	TE	4
4	Short Conversation	2	SR,TE	1
5	Long Conversation	1	SR,TE	3
6	Read-Aloud Story	1	SR,TE	3 or 4
7	Teacher Presentation	1	SR,TE	4
	Reading			
1	Read-Along Sentence	4	TE	1
2	Read and Match: Word	3	SR	1
3	Read and Match: Sentence	3	SR	1
4	Read for Details	1	TE	1
5	Short Correspondence	2	SR,TE	2 or 3
6	Procedural Text	1	SR,TE	3 or 4
7	Literary Set	1	SR,TE	3 or 4
8	Informational Set	1	SR,TE	3 or 4
	Speaking			
1	Classroom Tableau	1	CR	5
2	Conversation	1	CR	3
3	Compare Pictures	1	CR	1
4	Opinion	1	CR	1
5	Observe and Report	1	CR	1
	Writing			
1	Word Builder: Word	5	TE	1
2	Sentence Builder	5	TE	1
3	Picture Caption	2	CR	1
4	Opinion	1	CR	1
5	Storyboard	1	CR	1

Note: CR = Constructed Response (audio response, text response), SR = Selected Response (multiple-choice single selection, multiple-choice multiple selection), and TE = Technology-Enhanced (match single selection, match multiple selection, zone single selection, zone multiple selection)

Appendix A. Distribution of Operational Tasks in 2015-16 Summative Assessment

Table A4. Summative Assessment Blueprint, Grades 4-5

Slot	Task Type	# Tasks/Form	Response	# Items/Task
	Listening			
1	Listen and Match: Word	4	TE	1
2	Listen and Match: Sentence	4	TE	1
3	Follow Instructions	1	TE	3 or 4
4	Listen for Information	3	TE	1
5	Short Conversation	1	SR	3
6	Teacher Presentation: Read Aloud	1	SR	3 or 4
7	Interactive Student Presentation	1	SR,TE	3 or 4
8	Student Discussion	1	SR,TE	3 or 4
	Reading			
1	Match Picture to Word and Sentence	6	SR	1
2	Short Correspondence Set	1	SR,TE	4
3	Short Literary Set	1	SR,TE	3 or 4
4	Short Informational Set	1	SR,TE	4
5	Extended Literary Set	1	SR,TE	3, 4, or 5
6	Extended Informational Set	1	SR,TE	4 or 5
	Speaking			
1	Oral Vocabulary	0	CR	5
2	Conversation	1	CR	4
3	Compare Pictures	1	CR	1
4	Language Arts Presentation	1	CR	3
5	Observe and Report	1	CR	1
6	Analyze a Visual	1	CR	2
	Writing			
1	Discrete editing tasks	2	TE	1
2	Word Builder: Word	3	TE	1
3	Sentence Builder	3	TE	1
4	Writing questions task	1	CR	3
5	Write an Opinion	1	CR	1
6	Storyboard	1	CR	1

Note: CR = Constructed Response (audio response, text response), SR = Selected Response (multiple-choice single selection, multiple-choice multiple selection), and TE = Technology-Enhanced (match single selection, match multiple selection, zone single selection, zone multiple selection)

Appendix A. Distribution of Operational Tasks in 2015-16 Summative Assessment

Table A5. Summative Assessment Blueprint, Grades 6-8

Slot	Task Type	# Tasks/Form	Response	# Items/Task
	Listening			
1	Listen and Match: Word	5	TE	1
2	Listen and Match: Sentence	5	TE	1
3	Follow Instructions	1	TE	2 or 3
4	Listen for Information	3	TE	1
5	Short Conversation	2	SR	3 or 4
6	Academic Lecture or Discussion	1	SR	4
7	Interactive Student Presentation	1	SR,TE	3 or 4
8	Academic Debate	1	SR,TE	3 or 4
	Reading			
1	Short Paragraph	3	SR	2
2	Short Literature Set	1	SR,TE	4 or 5
3	Short Informational Set	1	SR,TE	4
4	Extended Literature Set	1	SR	3 or 4
5	Extended Informational Set	1	SR,TE	5 or 7
6	Argument and Support Essay Set	1	SR	4 or 5
	Speaking			
1	Oral Vocabulary	0	CR	5
2	Compare Pictures	1	CR	1
3	Language Arts Presentation	1	CR	3
4	Analyze a Visual and a Claim	1	CR	2
5	Observe and Report	1	CR	1
	Writing			
1	Discrete editing tasks	2	TE	1
2	Writing questions task	1	CR	3
3	Respond to a Peer E-mail	1	CR	1
4	Storyboard	1	CR	1
5	Construct a claim	1	CR	1

Note: CR = Constructed Response (audio response, text response), SR = Selected Response (multiple-choice single selection, multiple-choice multiple selection), and TE = Technology-Enhanced (match single selection, match multiple selection, zone single selection, zone multiple selection)

Table A6. Summative Assessment Blueprint, Comparison of Alternate Test Forms, Grades 9-12

Slot	Task Type	# Tasks/Form	Response	# Items/Task
	Listening			
1	Listen and Match: Word	2	TE	1
2	Listen and Match: Sentence	3	TE	1
3	Listen for Information	3	TE	1
4	Short Conversation	1	SR,TE	4 or 5
5	Academic Lecture and Discussion	1	SR	4 or 5
6	Interactive Student Presentation	1	SR	3, 4, or 5
7	Academic Debate	1	SR,TE	3 or 4
	Reading			
1	Discrete Items	4	SR	2 or 3
2	Short Literary Set	1	SR,TE	4 or 6
3	Short Informational Set	1	SR,TE	4 or 5
4	Extended Literary Set	1	SR,TE	6 or 7
5	Extended Informational Set	1	SR,TE	5, 6, or 7
6	Argument and Support Essay Set	1	SR	4, 5, or 6
	Speaking			
1	Oral Vocabulary	0	CR	5
2	Compare Pictures	1	CR	1
3	Language Arts Presentation	1	CR	3
4	Observe and Report	1	CR	3
5	Analyze a Visual and a Claim Argument	1	CR	2
	Writing			
1	Discrete Editing Tasks	2	TE	1
2	Writing Questions Task	1	CR	3
3	Respond to a Peer E-mail	1	CR	1
4	Storyboard	1	CR	1
5	Construct a Claim	1	CR	1

Note: CR = Constructed Response (audio response, text response), SR = Selected Response (multiple-choice single selection, multiple-choice multiple selection), and TE = Technology-Enhanced (match single selection, match multiple selection, zone single selection, zone multiple selection)

Appendix B. Composition of Alternate Test Forms, 2015-16 Summative Assessment

This appendix describes the number of items, scores, and score points for each alternate test form. Separate tables are provided for each gradeband. Note that only operational items (those contributing to students' scores) are listed.

Table B1. Comparison of Alternate Test Forms, Grade K

Form	# Items	# Scores	# Score Points
	Listening		
L1	29	29	29
L2	27	27	27
L3	29	29	29
L4	27	27	27
L5	27	27	27
L6	27	27	27
	Reading		
R1	23	23	23
R2	23	23	23
R3	23	23	23
R4	23	23	23
R5	23	23	23
R6	23	23	23
	Speaking		
S1	25	11	27
S2	25	11	27
S3	25	11	27
S4	25	11	27
S5	25	11	27
S6	25	11	27
	Writing		
W1	10	10	10
W2	10	10	10
W3	10	10	10
W4	10	10	10
W5	10	10	10
W6	10	10	10
	Writing Supplement (Paper)		
WS1	5	5	12
WS2	5	5	12
WS3	5	5	12
WS4	5	5	12

Table B2. Comparison of Alternate Test Forms, Grade 1

Form	# Items	# Scores	# Score Points
	Listening		
L1	25	25	25
L2	24	24	24
L3	25	25	25
L4	24	24	24
L5	25	25	25
L6	24	24	24
	Reading		
R1	30	30	30
R2	30	30	30
R3	30	30	30
R4	31	31	31
R5	31	31	32
R6	30	30	31
	Speaking		
S1	12	9	25
S2	12	9	25
S3	12	9	25
S4	12	9	25
S5	12	9	25
S6	12	9	25
	Writing		
W1	10	10	10
W2	10	10	10
W3	10	10	10
W4	10	10	10
W5	10	10	10
W6	10	10	10
	Writing Supplement (Paper)		
WS1	4	4	11
WS2	4	4	11
WS3	4	4	11
WS4	4	4	11
WS5	4	4	11
WS6	4	4	11

Appendix B. Composition of Alternate Test Forms, 2015-16 Summative Assessment

Table B3. Comparison of Alternate Test Forms, Grades 2-3

Form	# Items	# Scores	# Score Points
	Listening		
L1	24	24	26
L2	25	25	27
L3	25	25	26
L4	24	24	26
L5	24	24	26
L6	24	24	26
	Reading		
R1	28	28	35
R2	29	29	37
R3	27	27	35
R4	29	29	37
R5	27	27	35
R6	29	29	35
	Speaking		
S1	11	9	25
S2	11	9	25
S3	11	9	25
S4	11	9	25
S5	11	9	25
S6	11	9	25
	Writing		
W1	14	14	24
W2	14	14	24
W3	14	14	24
W4	14	14	24
W5	14	14	24
W6	14	14	24

Table B4. Comparison of Alternate Test Forms, Grades 4-5

Form	# Items	# Scores	# Score Points
	Listening		
L1	28	28	32
L2	28	28	32
L3	27	27	31
L4	26	26	30
L5	28	28	32
L6	29	29	33
	Reading		
R1	27	27	29
R2	26	26	28
R3	27	27	30
R4	28	28	30
R5	28	28	31
R6	24	24	26
	Speaking		
S1	11	8	30
S2	11	8	30
S3	11	8	30
S4	11	8	30
S5	11	8	30
S6	11	8	30
	Writing		
W1	13	13	30
W2	13	13	30
W3	13	13	30
W4	13	13	30
W5	13	13	30
W6	13	13	30

Appendix B. Composition of Alternate Test Forms, 2015-16 Summative Assessment

Table B5. Comparison of Alternate Test Forms, Grades 6-8

Form	# Items	# Scores	# Score Points
	Listening		
L1	33	33	36
L2	33	33	36
L3	32	32	34
L4	34	34	36
L5	33	33	35
	Reading		
R1	28	28	32
R2	29	29	33
R3	28	28	33
R4	29	29	33
R5	28	28	32
	Speaking		
S1	7	7	27
S2	7	7	27
S3	7	7	27
S4	7	7	27
S5	7	7	27
	Writing		
W1	8	8	28
W2	8	8	28
W3	8	8	28
W4	8	8	28
W5	8	8	28

Appendix B. Composition of Alternate Test Forms, 2015-16 Summative Assessment

Table B6. Comparison of Alternate Test Forms, Grades 9-12

Form	# Items	# Scores	# Score Points
	Listening		
L1	25	25	27
L2	26	26	28
L3	24	24	27
L4	24	24	28
L5	24	24	27
	Reading		
R1	35	35	36
R2	36	36	38
R3	36	36	38
R4	37	37	38
R5	35	35	38
	Speaking		
S1	7	7	27
S2	7	7	27
S3	7	7	27
S4	7	7	27
S5	7	7	27
	Writing		
W1	8	8	28
W2	8	8	28
W3	8	8	28
W4	8	8	28
W5	8	8	28