OHIO TEST OF ENGLISH LANGUAGE ACQUISITION (OTELA)
MARCH 2015 ADMINISTRATION

STATISTICAL SUMMARY

Test Grade Cluster/ Subject	N-count	Max Raw Score	Raw Score Mean	Raw Score Standard Deviation	Raw Score SEM	Max Scaled Score	Scaled Score Mean	Scaled Score Standard Deviation	Scaled Score SEM	Reliability
Grade K Listening	6240	21	12.28	5.15	1.32	432.00	336.49	50.96	13.06	0.93
Grade K Speaking	6240	24	14.67	6.28	1.30	431.00	326.61	59.10	12.24	0.96
Grade K Reading	6240	42	21.77	10.31	2.36	391.00	303.17	32.61	7.47	0.95
Grade K Writing	6240	27	12.48	6.47	1.65	389.00	279.06	41.30	10.51	0.94
Grade 1-2 Listening	12571	21	15.25	4.75	1.18	421.00	351.99	52.20	13.01	0.94
Grade 1-2 Speaking	12571	24	17.87	5.55	1.20	426.00	353.79	56.83	12.28	0.95
Grade 1-2 Reading	12571	42	27.37	10.15	1.96	406.00	323.34	44.49	8.61	0.96
Grade 1-2 Writing	12571	27	18.10	6.23	1.50	397.00	311.63	45.74	11.02	0.94
Grade 3-5 Listening	14813	18	11.91	3.65	1.79	928.00	702.09	132.20	64.95	0.76
Grade 3-5 Speaking	14813	24	19.85	5.01	1.59	937.00	803.80	148.59	47.04	0.90
Grade 3-5 Reading	14813	20	12.52	4.45	1.85	925.00	653.13	151.97	63.28	0.83
Grade 3-5 Writing	14813	17	9.56	3.56	1.91	923.00	625.83	130.47	69.87	0.71
Grade 6-8 Listening	8009	18	12.48	3.76	1.68	937.00	765.49	134.07	60.08	0.80
Grade 6-8 Speaking	8009	24	19.93	6.00	1.43	948.00	831.20	163.69	39.12	0.94
Grade 6-8 Reading	8009	20	11.95	4.53	1.89	929.00	656.29	151.63	63.44	0.82
Grade 6-8 Writing	8009	17	9.45	3.27	1.56	999.00	698.62	167.80	80.22	0.77
Grade 9-12 Listening	7380	20	12.82	4.28	1.90	942.00	761.65	126.59	56.24	0.80
Grade 9-12 Speaking	7380	24	19.31	6.62	1.46	948.00	828.55	163.26	36.03	0.95
Grade 9-12 Reading	7380	20	11.20	4.78	1.95	933.00	668.79	146.22	59.75	0.83
Grade 9-12 Writing	7380	20	11.53	4.32	1.98	950.00	689.20	148.69	67.95	0.79

This table describes the population of Ohio limited English proficient (LEP) students completing all domains (i.e., not receiving DNA or INV for any of the four subjects) in the March 2015 OTELA administration.

OTELA Cut Score Points for All Performance Standards

		Performance Standard Cut Scores			
		Beginners	**Intermediate**	**Advanced**	**Full English Proficiency**
Grade K Listening	Raw Score	4	9	15	19
	Scaled Score	257	307	360	402
Grade K Speaking	Raw Score	7	12	18	22
	Scaled Score	262	300	351	397
Grade K Reading	Raw Score	10	21	35	39
	Scaled Score	272	301	341	359
Grade K Writing	Raw Score	8	16	21	26
	Scaled Score	255	300	329	375
Grade 1–2 Listening	Raw Score	6	11	16	19
	Scaled Score	257	305	357	391
Grade 1–2 Speaking	Raw Score	9	13	18	22
	Scaled Score	270	301	350	393
Grade 1–2 Reading	Raw Score	12	22	30	38
	Scaled Score	264	300	329	366
Grade 1–2 Writing	Raw Score	8	18	22	26
	Scaled Score	245	306	335	381
Grade 3–5 Listening	Raw Score	5	8	11	13
	Scaled Score	473	570	659	725
Grade 3–5 Speaking	Raw Score	6	10	16	21
	Scaled Score	456	554	685	818
Grade 3–5 Reading	Raw Score	6	11	13	17
	Scaled Score	450	595	653	805
Grade 3–5 Writing	Raw Score	5	9	12	14
	Scaled Score	472	602	703	786
Grade 6–8 Listening	Raw Score	7	9	12	14
	Scaled Score	581	643	739	815
Grade 6–8 Speaking	Raw Score	5	11	17	21
	Scaled Score	487	617	731	841

		Performance Standard Cut Scores			
		Beginners	**Intermediate**	**Advanced**	**Full English Proficiency**
Grade 6–8 Reading	Raw Score	6	11	14	17
	Scaled Score	472	620	710	829
Grade 6–8 Writing	Raw Score	7	9	11	13
	Scaled Score	578	668	772	899
Grade 9–12 Listening	Raw Score	6	9	12	16
	Scaled Score	568	653	732	859
Grade 9–12 Speaking	Raw Score	7	12	18	21
	Scaled Score	572	663	772	854
Grade 9–12 Reading	Raw Score	7	10	14	17
	Scaled Score	552	631	741	854
Grade 9–12 Writing	Raw Score	6	10	13	17
	Scaled Score	512	637	735	893

Note: Scale score cuts may not be observable on all forms and may not correspond directly to the attainable raw score in each category. Observable scale scores are presented in the raw to scale score conversion tables below.

Percentage of Students at Each Performance Level

	Percentage of Students at Each Performance Level				
Test Grade Cluster/Subject	**Pre-functional**	**Beginners**	**Intermediate**	**Advanced**	**Full English Proficiency**
Grade K Listening	5.03	20.11	38.40	25.02	11.44
Grade K Speaking	12.50	17.13	34.01	21.81	14.55
Grade K Reading	14.26	30.16	43.30	8.11	4.17
Grade K Writing	24.92	40.72	22.15	10.64	1.57
Grade K Comprehension	14.02	30.21	43.53	8.16	4.09
Grade K Production	23.11	39.81	24.95	10.69	1.44
Grade K Composite	23.32	40.74	27.34	7.52	1.09
Grade 1-2 Listening	4.36	12.15	28.79	25.52	29.18
Grade 1-2 Speaking	7.92	7.56	24.17	29.31	31.04
Grade 1-2 Reading	8.07	19.04	25.29	30.25	17.35
Grade 1-2 Writing	7.08	31.33	27.87	25.13	8.59
Grade 1-2 Comprehension	7.92	18.80	25.95	30.36	16.98
Grade 1-2 Production	6.67	29.69	30.15	25.07	8.42
Grade 1-2 Composite	9.08	28.10	30.47	25.32	7.02
Grade 3-5 Listening	3.31	10.12	19.75	18.04	48.78
Grade 3-5 Speaking	3.34	2.19	8.65	25.75	60.06
Grade 3-5 Reading	8.40	23.61	12.85	33.01	22.13
Grade 3-5 Writing	9.09	28.63	29.87	17.94	14.47
Grade 3-5 Comprehension	6.92	21.32	17.59	32.85	21.32
Grade 3-5 Production	5.45	19.55	42.67	18.05	14.28
Grade 3-5 Composite	8.39	23.22	33.22	26.88	8.29
Grade 6-8 Listening	9.25	7.23	16.24	19.32	47.96
Grade 6-8 Speaking	5.13	4.58	8.10	14.92	67.26
Grade 6-8 Reading	10.38	26.57	20.43	24.81	17.82
Grade 6-8 Writing	19.62	13.96	22.49	26.67	17.27
Grade 6-8 Comprehension	9.10	23.42	25.07	25.06	17.34

	Percentage of Students at Each Performance Level				
Test Grade Cluster/Subject	**Pre-functional**	**Beginners**	**Intermediate**	**Advanced**	**Full English Proficiency**
Grade 6-8 Production	12.41	13.41	30.33	26.74	17.11
Grade 6-8 Composite	14.45	19.23	29.85	28.32	8.15
Grade 9-12 Listening	6.83	12.57	15.23	32.01	33.36
Grade 9-12 Speaking	8.69	5.57	10.12	11.48	64.15
Grade 9-12 Reading	21.15	17.60	24.00	20.93	16.31
Grade 9-12 Writing	11.75	16.99	23.09	38.05	10.12
Grade 9-12 Comprehension	17.48	20.01	25.53	20.88	16.10
Grade 9-12 Production	10.19	12.53	29.51	37.80	9.96
Grade 9-12 Composite	15.62	20.28	30.23	28.60	5.26

This table describes the population of Ohio limited English proficient (LEP) students completing all domains (i.e., not receiving DNA or INV for any of the four subjects) in the March 2015 OTELA administration.

Equating and Scaling: How Raw Scores Are Converted into Scaled Scores

Test Form Construction

The Ohio Test of English Language Acquisition (OTELA) is based on the English Language Development Assessment (ELDA) developed under the direction of a consortium of 18 member states of the LEP State Collaborative on Assessment and Student Standards (LEP-SCASS) and the Council of Chief State School Officers. The ELDA was designed to allow states to meet federal requirements under NCLB concerning the annual assessment of LEP students regarding their acquisition of and progress toward developing English language proficiency in listening, speaking, reading, and writing.

The OTELA is a battery of tests designed to allow schools to measure progress in the acquisition of English language proficiency skills among non-native English-speaking students. The battery consists of separate tests for listening, speaking, reading, and writing, for each of five grade clusters: K, 1–2, 3–5, 6–8 and 9–12. The tests are aligned with Ohio's English language proficiency standards and were constructed to provide content coverage across four academic topic areas (English Language Arts; Mathematics, Science and Technology; and Social Studies), and one non-academic topic area, School-Environmental, which is related to aspects of the school environment such as extracurricular activities, student health, homework, classroom management, and lunchtime. Although the OTELA tests measure language skills with content drawn from age-appropriate curricular and non-curricular sources, they are not tests of academic content. Students do not need any external or prior content-related knowledge to respond to the test questions.

To measure a wide range of English language proficiency, the full-length ELDA includes many items and requires substantial test administration time. Although administration of the ELDA test battery is not officially timed, general guidelines indicate approximately four hours of test administration time. In addition, most students to whom the ELDA was administered scored in the upper ranges of the raw score distribution. These performance results indicated that the ELDA operational forms could be shortened substantially by eliminating the easiest items in the operational item bank while maintaining a proportional representation of items across content standards within each domain.

OTELA items were selected on the basis of their psychometric properties, contribution to measurement at key points on the scale (such as the intermediate cut score), and content coverage. When, for example, the easiest items within a domain proved to be concentrated within specific content standards, the Ohio Department of Education (ODE) opted to maintain breadth of content coverage, rather than to simply increase form difficulty. In addition, although a primary goal was to reduce test length as much as possible, estimated form reliabilities were used to determine the appropriate number of items to include in each test form.

Common Item Equating

Grade clusters 3–5, 6–8, and 9–12. Following the first operational administration of grades 3–12 ELDA forms in 2005, items included in the first operational test forms were recalibrated, with the resulting item parameter estimates serving as the reference scales for ELDA. All subsequent grades 3–12 ELDA test forms are linked to these scales.

Because the first set of operational forms were constructed to include a set of common items between adjacent grade clusters, the grades 3–5, 6–8 and 9–12 forms were jointly calibrated in a single Winsteps run for each domain, resulting in a common, vertically linked scale across grade clusters for each domain. For each Winsteps run, the mean of the item difficulty parameters was fixed to zero so that the average difficulty for all items across grade clusters was equal to zero within each domain for the first operational form.

For the 2005 field test, a common item design was used to allow common item equating across field-test forms and the first operational form. Following the common item design of the field test, items were jointly calibrated in a single Winsteps run for each domain and grade-cluster combination. Because all of the 2005 ELDA field-test forms shared items in common with operational Form 1, a common item equating method was used to link the field-test items to the ELDA operational Form 1 scale. For each field-test form within each grade cluster, shared items were fixed to their operational Form 1 parameter estimates, while the remaining items were freely estimated. This placed all the field-test items on the operational Form 1 scale.

In addition, a small subset of items were field tested in 2004 but were not included in the 2005 operational forms. These items were also placed on the 2005 operational ELDA scale. Because all items in the 2005 operational test came from the 2004 field-test item pool, the 2005 operational test items were used as linking items. The mean-mean procedure was used to find the linking constant. To ensure that the final set of anchor items (i.e., common items) was free of item parameter drift, a stepwise deletion procedure was used to select anchor items and calculate the linking constant needed to bring the field test items onto the reference scale defined by the first operational administration. Following this procedure, a linking constant was calculated, using all anchor items, and then applied the linking constant to bring the items back to the reference scale. Anchor item parameter estimates were then examined to determine whether the difference between any adjusted or linked parameter estimates and the reference scale parameter estimates was greater than .3 logits. At each step, the item with the greatest difference between its linked and reference item parameter estimates was eliminated from the anchor set, provided the difference was greater than .3. A new linking constant was then computed and applied to the test items and the parameter estimates for the remaining anchor items were again examined to determine whether any exceeded the .3 tolerance level. This process was repeated until all remaining anchor items met the tolerance-level specifications. The linking constant was computed on the basis of this final anchor item set, and then applied to the 2004 ELDA field-test item parameters.

The result of these analyses was to place all items in each of the grade 3–12 ELDA domain item banks on the common scale defined by the first operational administration.

Additional items were subsequently developed for the OTELA assessment program and these items were embedded in the operational test forms for the 2009 and 2010 administrations of OTELA. Operational and embedded field test items were concurrently calibrated. The operational test items were used to link items from the 2009 and 2010 operational administration to the original ELDA scale. The average item difficulty for the operational test items were then computed based on both the spring 2009 and 2010 operational administration and the bank item parameter estimates from the original ELDA operational administration to identify the linking constant necessary to bring the 2009 and 2010 operational item parameters back to the ELDA reference scale. The resulting linking constant was then applied to the field test items to place the field test item parameter estimates on the original ELDA scale.

Grade clusters K and 1–2. Items in the grades K and 1–2 OTELA forms were calibrated independently of the items in the grades 3–12 scales and are not reported on the vertical scale used to report scores on the grades 3–12 OTELA tests. A large proportion of items in the listening and speaking tests are common across the grades K and 1–2 test forms, while item overlap between the grades K and 1–2 reading and writing test forms is minimal. Consistent with this perspective, item difficulties for the kindergarten and grades 1–2 OTELA test forms were calibrated following two distinct strategies. Parameters for all OTELA kindergarten and grade 1–2 items were estimated using Masters' partial credit model, an extension of the Rasch model for polytomous items. Student item scores were obtained from the Spring 2006 operational administration of the OTELA. For the reading and writing assessments, items in each of the grades K and 1–2 operational test forms were calibrated in separate Winsteps runs. For the listening and speaking items, parameters for items in both the grades K and 1–2 forms were estimated simultaneously in a joint calibration. Once the listening and speaking items were calibrated, the resulting cross-grade item parameter estimates were used to generate form-specific raw score to theta scale conversion tables.

Reporting scales for the grades K and 1–2 OTELA forms were established by setting the "intermediate" or level 3, performance standard for each of the assessments to 300. Therefore, for both the grades K and 1–2 assessments, and across the four English language domains assessed, a score of 300 indicates attainment of an intermediate level of English language proficiency. The standard deviation of the scale was set to 15.

Refreshing Bank Item Parameters

Because item parameter estimates may change over time, it is desirable to update the bank item parameter estimates. To accomplish this, in early 2014 student records from the previous operational administration of the OTELA forms were used to recalibrate and equate bank item parameters. Mean-mean equating was used to link recalibrated item parameters back to the reference OTELA scale. Items showing evidence of drift were examined to ensure that there were no changes to item content or presentation that might be expected to change the performance of the item and warrant dropping the item from the linking set. Based on the results of this review, all operational test items were used to compute the final linking constants.

Performance Standards

The OTELA is designed to provide student performance-level assessment results that are fully comparable with those from the ELDA. To achieve this goal, the OTELA uses the same performance standards adopted by the LEP-SCASS for the ELDA. Performance levels range from Full English Proficiency, a level at which an LEP student is deemed to be able to function effectively and consistently through the medium of academic English in the school system (and thus ceases to be defined as LEP), to Pre-functional, a level at which an LEP student is consistently unable to communicate with any success in the English of the school environment, although the student may have some limited knowledge of English. Student performance levels are reported for each of the four language domain scores, as well as for English language comprehension (derived from student performance on the listening and reading tests), production (derived from student performance on the speaking and writing assessments), and a composite performance level that reflects student performance in both English language comprehension and production.

OTELA Performance Levels

Level	Label
5	Full English Proficiency
4	Advanced
3	Intermediate
2	Beginners
1	Pre-functional

In the process of adopting ELDA performance standards for the OTELA, ODE, in consultation with the Ohio LEP Advisory Committee, elected to revise one ELDA performance level cut score. In the ELDA performance standards for writing, students in the grade 3–5 cluster must substantially outperform students in both the 6–8 and 9–12 grade clusters to achieve Full English Proficiency. To address this issue, a linear regression approach was used to identify a cut score for Full English Proficiency at the grades 3–5 cluster from the cut scores identified for Beginning, Intermediate, and Advanced performance levels on the grade 3–5 writing assessment. This analysis identified a cut score of 2.08 (in the theta metric; 867 on the ELDA reporting scale) for the Full English Proficiency cut score at the 3–5 grade cluster. AIR submitted the cut score and estimated impact data for the revised performance standard to the Ohio LEP Advisory Committee for their consideration. The Ohio LEP Advisory Committee recommended that ODE adopt the revised performance standard, which ODE has done.

While performance levels for the four domain tests (Listening, Speaking, Reading, and Writing) are based on scaled scores, performance levels for the three derived scores (Comprehension, Production and Composite) are based on the performance levels of the underlying domain tests. The Comprehension performance level is based on the set of rules relating student performance levels on the Listening and Reading domain tests shown in the table below. Following these rules, if a student performed at level 3 on the Reading test and at level 2 on the Listening test, then the student would receive a level 3 for English language Comprehension. If the levels were reversed, so that a student performed at level 2 on Reading and level 3 on Listening, then the assigned Comprehension performance level would be 2.

Rules for Combining Listening and Reading Levels to Yield Student Comprehension Level

Rules for Combining Listening and Reading Levels to Yield Student Comprehension Level		
If *Reading* Level is:	And *Listening* Level is:	Then *Comprehension* Level is:
1	1	1
	2	1
	3	1
	4	2
	5	2
2	1	2
	2	2
	3	2
	4	2
	5	3
3	1	2
	2	3
	3	3
	4	3
	5	3
4	1	3
	2	3
	3	4
	4	4
	5	4
5	1	3
	2	3
	3	4
	4	5
	5	5

Similarly, performance levels for Production are based on the set of rules shown below describing the relationship between Speaking and Writing performance levels. For example, a student performing at level 5 on the Writing test and at level 4 on the Speaking test would receive

a 5 for English language Production. If the levels were reversed, however, so that the student performed at level 4 in Writing and level 5 on the Speaking test, then the Production performance level would be set to 4.

Rules for Combining Writing and Speaking Levels to Yield Student Production Level

Rules for Combining Writing and Speaking Levels to Yield Student Production Level		
Writing Level is:	And *Speaking* Level is:	Then *Production* Level is:
1	1	1
	2	1
	3	1
	4	2
	5	2
2	1	2
	2	2
	3	2
	4	2
	5	3
3	1	2
	2	3
	3	3
	4	3
	5	3
4	1	3
	2	3
	3	4
	4	4
	5	4
5	1	3
	2	3
	3	4
	4	5
	5	5

Performance levels for Comprehension and Production are in turn evaluated to create an overall Composite level using the rules shown below. When the Comprehension and Production performance levels are not the same, the rule is to average the two levels and round down. For example, if the performance level for Production were 3 and the performance level for Comprehension were 4, the average would be 3.5, and the final Composite performance level would be reported as 3.

Rules for Combining Comprehension and Production Levels to Yield Student Composite Level

Rules for Combining Comprehension and Production Levels to Yield Student Composite Level		
If *Production* Level is:	And *Comprehension* Level is:	Then *Composite* Level is:
1	1	1
	2	1
	3	2
	4	2
	5	3
2	1	1
	2	2
	3	2
	4	3
	5	3
3	1	2
	2	2
	3	3
	4	3
	5	4
4	1	2
	2	3
	3	3
	4	4
	5	4
5	1	3
	2	3
	3	4
	4	4
	5	5

Spring 2015 Raw Score to Scaled Score Conversion Table—Grades K–2

	Scaled Scores Corresponding to Raw Score Points							
Raw Score	**Grade K Listening**	**Grade K Speaking**	**Grade K Reading**	**Grade K Writing**	**Grade 1–2 Listening**	**Grade 1–2 Speaking**	**Grade 1–2 Reading**	**Grade 1–2 Writing**
0	197	183	206	178	179	176	184	175
1	214	197	220	193	195	191	198	189
2	230	212	233	207	211	205	212	202
3	244	223	242	218	223	216	221	212
4	257	234	248	227	235	226	228	219
5	268	244	253	235	246	235	234	226
6	279	253	258	242	257	245	239	232
7	289	262	262	249	267	254	243	239
8	298	271	265	255	278	262	248	245
9	307	278	269	261	287	270	252	251
10	316	286	272	267	297	278	256	258
11	325	293	275	273	305	286	260	264
12	333	300	278	278	315	293	264	270
13	342	308	281	284	324	301	268	276
14	351	316	283	289	335	310	272	281
15	360	324	286	294	346	319	276	287
16	369	333	289	300	357	329	279	293
17	379	342	291	305	368	340	283	299
18	390	351	294	311	379	350	287	306
19	402	360	296	317	391	360	290	313
20	417	371	299	323	406	370	294	320
21	432	383	301	329	421	381	297	327
22		397	304	336		393	300	335
23		414	306	343		410	304	344
24		431	309	351		426	307	353
25			311	361			311	365
26			314	375			314	381
27			317	389			318	397
28			319				322	
29			322				325	

	Scaled Scores Corresponding to Raw Score Points							
Raw Score	**Grade K Listening**	**Grade K Speaking**	**Grade K Reading**	**Grade K Writing**	**Grade 1–2 Listening**	**Grade 1–2 Speaking**	**Grade 1–2 Reading**	**Grade 1–2 Writing**
30			325				329	
31			328				334	
32			331				338	
33			334				342	
34			337				346	
35			341				351	
36			344				356	
37			349				361	
38			354				366	
39			359				373	
40			367				381	
41			379				394	
42			391				406	

Spring 2015 Raw Score to Scaled Score Conversion Table—Grades 3–12

	Scaled Scores Corresponding to Raw Score Points											
Raw Score	**Grade 3–5 Listening**	**Grade 3–5 Speaking**	**Grade 3–5 Reading**	**Grade 3–5 Writing**	**Grade 6–8 Listening**	**Grade 6–8 Speaking**	**Grade 6–8 Reading**	**Grade 6–8 Writing**	**Grade 9–12 Listening**	**Grade 9–12 Speaking**	**Grade 9–12 Reading**	**Grade 9–12 Writing**
0	161	135	147	204	178	226	150	74	215	284	169	197
1	237	219	176	277	273	300	194	186	306	353	261	276
2	330	302	271	350	367	374	290	299	398	423	353	354
3	389	355	331	398	427	421	351	376	455	466	411	405
4	435	395	377	437	473	457	398	436	499	499	454	445
5	473	427	416	472	513	487	437	487	536	527	491	480
6	508	456	450	505	548	513	472	534	568	550	523	512
7	540	482	481	538	581	537	504	578	598	572	552	544
8	570	507	511	570	613	559	534	623	626	592	579	575
9	599	531	539	602	643	579	563	668	653	610	605	606
10	629	554	567	634	674	599	592	717	679	628	631	637
11	659	576	595	668	706	617	620	772	705	646	657	669
12	691	599	624	703	739	636	649	836	732	663	684	701
13	725	620	653	742	775	654	678	899	760	680	711	735
14	764	642	685	786	815	672	710	924	790	697	741	769
15	809	664	720	843	862	691	744	954	822	714	773	806
16	868	685	759	902	909	710	783	991	859	732	810	847
17	914	708	805	923	923	731	829	999	903	751	854	893
18	928	731	866		937	753	889		914	772	897	907
19		757	907			778	911		928	795	915	928
20		785	925			806	929		942	822	933	950
21		818				841				854		
22		863				888				898		
23		912				922				924		
24		937				948				948		