OHIO TEST OF ENGLISH LANGUAGE ACQUISITION (OTELA)
MARCH 2014 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	6058	21	12.12	5.30	1.32	432	335.08	52.78	13.13	0.94
Grade K Speaking	6058	24	14.38	6.46	1.34	431	324.03	60.74	12.57	0.96
Grade K Reading	6058	42	21.63	10.46	2.37	391	303.04	33.36	7.54	0.95
Grade K Writing	6058	27	12.42	6.56	1.66	389	278.46	42.40	10.74	0.94
Grade 1-2 Listening	11888	21	15.27	4.70	1.21	421	352.26	51.66	13.29	0.93
Grade 1-2 Speaking	11888	24	17.91	5.41	1.20	426	354.01	55.62	12.31	0.95
Grade 1-2 Reading	11888	42	27.73	9.99	1.99	406	324.99	43.79	8.73	0.96
Grade 1-2 Writing	11888	27	18.28	6.11	1.51	397	312.92	45.05	11.13	0.94
Grade 3-5 Listening	13585	18	12.18	3.91	1.77	924	693.33	142.51	64.56	0.79
Grade 3-5 Speaking	13585	24	20.46	4.63	1.47	933	800.06	140.99	44.85	0.90
Grade 3-5 Reading	13585	20	12.96	4.28	1.84	924	654.42	149.09	64.06	0.82
Grade 3-5 Writing	13585	17	9.18	3.34	1.65	946	609.38	140.51	69.31	0.76
Grade 6-8 Listening	7397	18	12.81	3.47	1.63	940	771.67	135.14	63.70	0.78
Grade 6-8 Speaking	7397	24	20.65	5.22	1.32	947	849.60	144.85	36.64	0.94
Grade 6-8 Reading	7397	20	12.82	4.24	1.86	924	656.02	144.97	63.64	0.81
Grade 6-8 Writing	7397	17	10.02	3.11	1.60	947	685.52	117.47	60.40	0.74
Grade 9-12 Listening	6740	20	13.09	4.14	1.86	943	765.93	125.94	56.68	0.80
Grade 9-12 Speaking	6740	24	19.24	6.34	1.52	955	824.58	164.91	39.48	0.94
Grade 9-12 Reading	6740	20	10.96	4.09	1.96	936	663.78	131.28	62.93	0.77
Grade 9-12 Writing	6740	21	11.42	4.18	1.90	969	666.90	133.36	60.72	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 2014 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	248	300	355	399
Grade K Speaking	Raw Score	7	12	18	22
	Scaled Score	255	300	349	394
Grade K Reading	Raw Score	10	21	35	39
	Scaled Score	270	300	338	359
Grade K Writing	Raw Score	8	16	21	26
	Scaled Score	251	300	328	375
Grade 1–2 Listening	Raw Score	6	11	16	19
	Scaled Score	254	300	348	382
Grade 1–2 Speaking	Raw Score	9	13	18	22
	Scaled Score	266	300	344	388
Grade 1–2 Reading	Raw Score	12	22	30	38
	Scaled Score	262	300	328	364
Grade 1–2 Writing	Raw Score	8	18	22	26
	Scaled Score	245	300	329	369
Grade 3–5 Listening	Raw Score	5	8	12	14
	Scaled Score	450	544	645	725
Grade 3–5 Speaking	Raw Score	6	11	18	22
	Scaled Score	450	547	668	809
Grade 3–5 Reading	Raw Score	7	11	14	17
	Scaled Score	450	580	648	770
Grade 3–5 Writing	Raw Score	6	9	11	14
	Scaled Score	450	577	669	785
Grade 6–8 Listening	Raw Score	7	9	12	14
	Scaled Score	554	626	718	806
Grade 6–8 Speaking	Raw Score	5	11	17	21
	Scaled Score	458	611	719	825

		Performance Standard Cut Scores			
		Beginners	**Intermediate**	**Advanced**	**Full English Proficiency**
Grade 6–8 Reading	Raw Score	7	12	15	18
	Scaled Score	460	612	690	829
Grade 6–8 Writing	Raw Score	7	10	12	15
	Scaled Score	553	653	722	894
Grade 9–12 Listening	Raw Score	6	9	13	16
	Scaled Score	556	632	729	850
Grade 9–12 Speaking	Raw Score	8	13	19	21
	Scaled Score	570	650	765	850
Grade 9–12 Reading	Raw Score	7	10	13	17
	Scaled Score	545	630	718	850
Grade 9–12 Writing	Raw Score	6	11	14	17
	Scaled Score	509	631	719	850

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	6.29	20.04	37.50	23.61	12.56
Grade K Speaking	14.38	16.72	33.05	21.76	14.10
Grade K Reading	15.22	30.06	42.32	7.31	5.08
Grade K Writing	26.00	39.50	22.78	10.22	1.50
Grade K Comprehension	14.94	30.06	42.64	7.38	4.99
Grade K Production	24.38	38.46	25.70	10.04	1.42
Grade K Composite	24.74	39.63	26.92	7.54	1.16
Grade 1-2 Listening	3.72	12.89	29.54	24.12	29.74
Grade 1-2 Speaking	7.02	7.74	25.78	28.95	30.51
Grade 1-2 Reading	7.39	18.67	25.26	29.96	18.71
Grade 1-2 Writing	6.59	30.76	27.89	25.76	9.01
Grade 1-2 Comprehension	7.31	18.49	25.83	30.15	18.22
Grade 1-2 Production	6.35	28.73	30.37	25.77	8.78
Grade 1-2 Composite	8.46	27.25	30.55	26.06	7.67
Grade 3-5 Listening	3.97	10.61	24.24	17.52	43.67
Grade 3-5 Speaking	2.84	2.11	10.18	29.41	55.47
Grade 3-5 Reading	9.91	17.42	20.54	28.78	23.35
Grade 3-5 Writing	16.22	24.14	20.07	31.12	8.44
Grade 3-5 Comprehension	9.02	16.86	22.81	29.33	21.98
Grade 3-5 Production	7.53	22.33	30.62	31.22	8.30
Grade 3-5 Composite	10.95	21.12	30.65	31.92	5.37
Grade 6-8 Listening	7.41	4.87	15.26	21.47	50.99
Grade 6-8 Speaking	3.38	3.43	6.16	14.70	72.33
Grade 6-8 Reading	10.72	22.66	24.60	30.16	11.86
Grade 6-8 Writing	14.45	23.37	26.44	31.28	4.45
Grade 6-8 Comprehension	9.21	19.40	29.59	30.21	11.59

	Percentage of Students at Each Performance Level				
Test Grade Cluster/Subject	**Pre-functional**	**Beginners**	**Intermediate**	**Advanced**	**Full English Proficiency**
Grade 6-8 Production	8.35	15.03	40.94	31.26	4.42
Grade 6-8 Composite	11.21	20.18	39.35	27.40	1.85
Grade 9-12 Listening	5.01	11.82	22.60	26.93	33.64
Grade 9-12 Speaking	8.55	6.35	13.53	9.58	61.99
Grade 9-12 Reading	16.57	19.48	25.15	30.01	8.78
Grade 9-12 Writing	10.95	25.73	27.60	27.09	8.64
Grade 9-12 Comprehension	14.55	19.84	27.18	29.97	8.46
Grade 9-12 Production	9.33	16.87	38.43	26.99	8.38
Grade 9-12 Composite	14.58	22.61	34.91	25.04	2.85

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 2014 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, 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 2014 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 2014 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	160	205	144	127	157	217	146	251	187	267	159	220
1	231	273	162	213	224	290	171	321	282	337	237	295
2	322	341	256	299	322	364	265	390	378	407	333	369
3	379	383	316	358	386	410	324	436	438	450	394	416
4	423	414	362	405	437	447	370	473	485	483	441	452
5	460	440	400	447	481	477	409	508	523	510	481	482
6	493	462	435	486	520	504	443	542	558	533	516	510
7	523	483	466	523	557	528	474	575	589	554	548	536
8	552	501	495	559	592	551	504	609	618	574	578	561
9	580	519	524	595	627	572	532	643	646	592	607	587
10	608	536	552	633	661	592	560	678	673	610	635	614
11	637	552	580	674	696	612	589	714	701	627	663	642
12	668	569	609	718	733	632	618	754	728	644	692	672
13	701	586	639	769	772	651	648	798	757	662	721	705
14	738	603	671	829	815	670	679	851	788	680	753	740
15	782	621	707	895	867	689	714	899	822	699	787	780
16	840	640	747	921	910	709	753	923	860	718	825	824
17	910	662	794	946	925	731	800	947	905	740	871	875
18	924	685	857		940	753	861		915	763	901	903
19		713	905			779	906		929	789	919	920
20		745	924			808	924		943	820	936	945
21		784				843				857		969
22		835				889				902		
23		905				922				928		
24		933				947				955		