

Idaho English Language Assessment

Technical Report

2006



Copyright ©2006, held by the Idaho State Board of Education. All rights reserved. Printed in the U.S.A. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without prior written permission. This assessment and its contents are the exclusive property of the State of Idaho.

Table of Contents

	Page
Structure of the IELA.....	3
Administration of the IELA	5
Scaling and Equating	5
Reliability of the IELA.....	5
Validity of the IELA.....	11
Content-related Validity	11
Criterion-related Validity	15
Standard Setting	17
Alignment Study	19
Glossary of Terms.....	20

Idaho English Language Assessment Technical Report

Structure of the IELA

The Idaho English Language Assessment (IELA) is a modified version of an assessment developed for the Mountain West Consortium. The test was designed to fulfill the requirements of ‘No Child Left Behind’ (NCLB) legislation. The IELA assesses English proficiency in Listening, Speaking, Reading, and Writing and reports scores in each of those language domains as well as in Comprehension (a combination of select items from the Listening and Reading test) and a total score. IELA test forms were designed for specific grade/grade clusters, K, 1-2, 3-5, 6-8, and 9-12, as shown in Table 1 (on page 4). For every grade cluster except Kindergarten, there are two forms differentiated by a number suffix (e.g., C1 and C2). The level 1 forms were designed to be administered to students on the lower end of the English proficiency scale (i.e., Beginner) and the level 2 forms designed for students on the upper end of the scale (i.e., Intermediate and Advanced). Within each grade cluster, the Listening and Speaking tests on level 1 and 2 forms are identical (i.e., feature the same items). The Reading and Writing tests on level 1 and 2 forms within a grade cluster are different, both in terms of the numbers of items and the content.

Prior to administration as the IELA, Mountain West test forms were reviewed and modified in several ways. The modifications fell into three areas:

- Directions for test administration. Some of the text intended to be read by the test administrator or by the test taker was modified to clarify directions.
- Rubrics for open-ended items. Some of the rubrics used to guide test administrators in scoring open-ended items were modified. The purpose of these modifications was to clarify rules for scoring and, in some cases, to add to the list of acceptable and unacceptable responses for each score point.
- Addition of linking items. In order to create a psychometric link between level 1 and 2 forms in each grade cluster, a sample of items (usually 5 each in reading and writing) was chosen from level 1 forms and these items were added to corresponding level 2 forms as common linking items.

Table 1 shows the test forms administered in each grade cluster and the composition of those forms, including the numbers of items by item type in each language domain as well as the number of points represented by those items. The items and points in the Comprehension column do not contribute to the Totals shown in the last two columns because all Comprehension items are part of the Listening or Reading tests.

Table 1. Structure and Content of IELA Test Forms

Grade Cluster	Form	Item Type	Listen		Speak		Read		Write		Comp		Total	
			Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts
K	A	MC	9	9	-	-	23	23	-	-	16	16	32	32
		SA	13	13	10	10	13	13	-	-	13	13	36	36
		ER	-	-	4	12	-	-	-	-	-	-	4	12
		Total	22	22	14	22	36	36	22*	22*	29	29	94	102
1-2	B1	MC	22	22	-	-	15	15	-	-	31	31	37	37
		SA	-	-	10	10	-	-	11	11	-	-	21	21
		ER	-	-	4	12	-	-	2	4	-	-	6	16
		Total	22	22	14	22	15	15	13	15	31	31	64	74
	B2	MC	22	22	-	-	20	20	-	-	39	39	42	42
		SA	-	-	10	10	-	-	10	10	-	-	20	20
		ER	-	-	4	12	-	-	3	10	-	-	7	22
		Total	22	22	14	22	20	20	13	20	39	39	69	84
3-5	C1	MC	22	22	-	-	15	15	4	4	31	31	41	41
		SA	-	-	10	10	-	-	5	5	-	-	15	15
		ER	-	-	4	12	-	-	2	6	-	-	6	18
		Total	22	22	14	22	15	15	11	15	31	31	62	74
	C2	MC	22	22	-	-	18	18	9	9	37	37	49	49
		SA	-	-	10	10	1	2	-	-	1	2	11	12
		ER	-	-	4	12	-	-	3	10	-	-	7	22
		Total	22	22	14	22	19	20	12	19	38	39	67	83
6-8	D1	MC	22	22	-	-	15	15	5	5	32	32	42	42
		SA	-	-	10	10	-	-	4	4	-	-	14	14
		ER	-	-	4	12	-	-	2	6	-	-	6	18
		Total	22	22	14	22	15	15	11	15	32	32	62	74
	D2	MC	22	22	-	-	18	18	10	10	38	38	50	50
		SA	-	-	10	10	-	-	-	-	-	-	10	10
		ER	-	-	4	12	2	6	3	10	2	6	9	28
		Total	22	22	14	22	20	24	13	20	40	44	69	88
9-12	E1	MC	22	22	-	-	15	15	7	7	32	32	44	44
		SA	-	-	10	10	-	-	2	2	-	-	12	12
		ER	-	-	4	12	-	-	2	6	-	-	6	18
		Total	22	22	14	22	15	15	11	15	32	32	62	74
	E2	MC	22	22	-	-	19	19	10	10	39	39	51	51
		SA	-	-	10	10	-	-	-	-	-	-	10	10
		ER	-	-	4	12	2	6	3	10	2	6	9	28
		Total	22	22	14	22	21	25	13	20	41	45	70	89

* Items on the Kindergarten Writing test are configured as a checklist completed by the examiner.

MC - Multiple Choice; SA - Short Answer; ER - Extended Response

Administration of the IELA

The IELA was administered in Spring (March 1 - April 14) 2006.

Scaling and Equating of the IELA

In order to accommodate short-answer and constructed response items on the Speaking and Writing subtests, as well as all multiple-choice items administered across the language domains, the Rasch Partial Credit Model (PCM), as implemented in WINSTEPS, version 3.57.1, was used. Within each grade cluster, all items on both forms (e.g., C1 and C2) were concurrently calibrated. This procedure placed all items from both forms on the same Rasch item difficulty scale, effectively equating level 1 and 2 forms.

By using the Rasch item parameter estimates from the concurrent calibration for just those items that are in each form, separate raw score to Rasch ability (θ) conversion tables were produced for each form. Cut scores were established via the August, 2006 IELA standard setting and these cut scores were then used to transform the Rasch ability estimates to scale scores. Specifically, for the total test, scale scores were determined by setting the Early Fluent and Fluent proficiency level cut scores, at the lowest grade in each grade cluster, to 400 and 425, respectively. Since separate cut scores were established by grade for the 1-2 and the 3-5 grade clusters, the scale scores of 400 and 425 for Early Fluent and Fluent apply only to the lowest grade of each grade cluster. However, the same cut scores were established for each grade for the 6-8 and the 9-12 grade clusters and, thus, the same set of scale score cuts applies to each grade for these two grade clusters. For each subtest, scale scores were determined by setting the Advanced Beginning and Early Fluent level cut scores, at the lowest grade in each grade cluster, to 80 and 100, respectively. Scale scores corresponding to each proficiency level by grade are shown in the Standard Setting section of this report.

Reliability of the IELA

Data bearing on the reliability of IELA 2006 Test Forms are shown in the multiple panels of Table 2 (pages 6-10). The panels show, by grade, test form, and language domain (and comprehension and the total test), the number of students (N) who were administered the form, coefficient Alpha, a measure of internal-consistency reliability, the maximum raw score attainable, and the mean, standard deviation, and standard error of measurement (SEM) in both raw score and scale score units. This table includes scores for students identified as LEP (limited English proficient) and LEP1¹ but not those identified as LEPX.² Number of students represents the number for whom there was a valid test score and may vary across language domains in a grade to the extent that there were students who did not attempt one or more of the language domain tests. There is a total score for each student regardless of whether or not all language domain tests were attempted.

1 New to U.S. school within last 10 months. (This rule has been changed to 12 months for the 2006-07 school year.)

2 Exited out of an LEP program within the last 2 years.

Table 2. Reliability, Raw Score and Scale Score Descriptive Statistics for IELA Test Forms by Grade

Grade K				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
A	Listening	2,054	0.85	22	12.5	4.8	1.83	101.1	21.2	8.11
	Speaking	2,046	0.84	22	12.7	5.2	2.05	100.9	24.0	9.54
	Reading	2,031	0.96	36	19.8	10.1	2.10	100.6	28.9	5.99
	Writing	1,983	0.94	22	12.8	5.9	1.45	104.1	32.4	8.00
	Comprehen	2,059	0.89	29	14.3	6.3	2.15	100.5	20.9	7.09
	Total	2,071	0.96	102	56.7	21.2	4.19	399.7	36.8	7.28

Grade 1

Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
B1	Listening	658	0.76	22	16.0	3.2	1.58	103.0	14.3	7.08
	Speaking	657	0.87	22	13.4	5.5	2.00	101.3	23.9	8.64
	Reading	651	0.82	15	11.2	2.9	1.24	100.2	19.4	8.29
	Writing	650	0.88	15	9.8	4.0	1.39	101.7	23.8	8.26
	Comprehen	666	0.81	31	22.2	5.0	2.17	100.3	14.7	6.41
	Total	667	0.93	74	49.5	13.6	3.66	399.8	42.0	11.28

B2	Listening	1,188	0.83	22	17.4	2.6	1.09	109.2	13.5	5.64
	Speaking	1,183	0.81	22	16.2	3.8	1.67	112.9	17.6	7.65
	Reading	1,225	0.72	20	15.0	3.1	1.63	107.8	16.2	8.53
	Writing	1,223	0.80	20	11.0	3.5	1.59	105.4	17.3	7.83
	Comprehen	1,227	0.81	39	29.3	5.4	2.35	106.7	13.3	5.77
	Total	1,227	0.90	84	58.4	11.5	3.72	418.1	34.2	11.04

Grade 2

Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
B1	Listening	406	0.82	22	17.6	3.3	1.41	111.8	17.5	7.51
	Speaking	402	0.89	22	15.4	5.9	1.97	110.4	27.0	9.08
	Reading	403	0.87	15	12.5	2.8	1.04	111.4	21.3	7.84
	Writing	396	0.91	15	11.4	3.9	1.16	112.5	25.0	7.50
	Comprehen	411	0.86	31	24.8	5.2	1.93	110.3	18.2	6.72
	Total	411	0.95	74	55.7	15.1	3.44	423.5	50.4	11.49

B2	Listening	1,283	0.77	22	19.2	2.0	0.94	120.0	14.7	7.02
	Speaking	1,281	0.76	22	18.1	3.2	1.57	121.8	18.3	8.87
	Reading	1,300	0.58	20	17.2	2.1	1.37	121.3	15.6	10.14
	Writing	1,298	0.71	20	14.3	2.6	1.37	123.6	16.4	8.81
	Comprehen	1,300	0.75	39	33.4	3.9	1.93	119.2	13.6	6.81
	Total	1,300	0.86	84	68.3	8.4	3.17	452.2	33.4	12.62

Grade 3				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
C1	Listening	398	0.83	22	14.6	4.5	1.85	99.2	14.8	6.04
	Speaking	398	0.86	22	15.4	5.4	2.02	98.5	19.4	7.34
	Reading	396	0.80	15	9.9	3.4	1.50	99.8	15.9	7.08
	Writing	397	0.80	15	10.0	3.1	1.39	99.5	16.7	7.51
	Comprehen	399	0.87	31	19.7	6.2	2.29	98.7	13.7	5.01
	Total	399	0.94	74	49.8	14.2	3.61	395.9	25.5	6.51
C2	Listening	1,230	0.75	22	17.0	3.3	1.66	106.6	12.6	6.30
	Speaking	1,228	0.71	22	18.4	3.0	1.64	109.6	14.5	7.83
	Reading	1,234	0.76	20	13.3	3.7	1.85	105.5	12.7	6.29
	Writing	1,231	0.73	19	12.0	2.9	1.52	105.5	13.2	6.87
	Comprehen	1,234	0.83	39	27.5	6.0	2.50	105.4	10.7	4.44
	Total	1,234	0.88	83	60.5	10.4	3.55	411.0	19.1	6.49

Grade 4				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
C1	Listening	350	0.85	22	15.5	4.6	1.78	102.0	15.0	5.84
	Speaking	347	0.88	22	16.4	5.4	1.91	103.0	20.7	7.31
	Reading	350	0.79	15	10.6	3.2	1.48	102.4	15.2	7.05
	Writing	348	0.82	15	10.5	3.2	1.35	102.7	17.8	7.49
	Comprehen	351	0.88	30	21.1	6.4	2.19	101.6	14.0	4.79
	Total	351	0.95	74	52.6	15.1	3.50	402.0	28.7	6.67
C2	Listening	1,169	0.74	22	18.2	2.8	1.43	111.6	12.5	6.35
	Speaking	1,171	0.71	22	19.4	2.7	1.44	114.4	14.7	7.97
	Reading	1,168	0.76	20	14.9	3.4	1.63	111.2	13.0	6.33
	Writing	1,167	0.74	19	13.2	2.7	1.38	111.6	14.0	7.08
	Comprehen	1,175	0.82	39	30.2	5.4	2.32	110.5	11.1	4.77
	Total	1,175	0.88	83	65.3	9.3	3.27	421.1	19.8	7.01

Grade 5				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
C1	Listening	328	0.84	22	16.2	4.3	1.75	104.8	15.1	6.15
	Speaking	309	0.91	22	16.4	5.3	1.64	103.0	20.3	6.26
	Reading	308	0.88	15	11.0	3.1	1.09	105.1	15.9	5.54
	Writing	308	0.86	15	10.6	3.0	1.10	103.5	17.1	6.32
	Comprehen	328	0.87	31	21.6	6.1	2.18	102.9	13.5	4.80
	Total	328	0.95	74	51.9	15.7	3.55	401.9	30.0	6.77
C2	Listening	1,066	0.68	22	18.9	2.5	1.42	114.9	12.6	7.17
	Speaking	1,064	0.70	22	19.8	2.4	1.34	117.2	14.8	8.10
	Reading	1,067	0.73	20	16.0	3.1	1.60	116.0	13.3	6.92
	Writing	1,067	0.66	19	14.1	2.4	1.43	116.7	14.4	8.44
	Comprehen	1,067	0.80	39	32.0	4.8	2.15	114.6	11.2	5.05
	Total	1,067	0.86	83	68.7	8.2	3.04	429.6	20.4	7.52

Grade 6				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
D1	Listening	247	0.85	22	13.6	5.1	1.98	92.9	12.0	4.71
	Speaking	239	0.91	22	13.9	6.7	1.97	91.8	17.9	5.30
	Reading	245	0.81	15	9.4	3.5	1.51	91.3	12.5	5.42
	Writing	245	0.84	15	9.2	3.5	1.39	91.5	14.6	5.80
	Comprehen	247	0.89	32	18.9	7.2	2.42	91.6	10.7	3.58
	Total	247	0.95	74	45.5	16.9	3.75	380.8	23.7	5.24

D2	Listening	1,059	0.77	22	17.8	3.2	1.55	103.1	10.2	4.89
	Speaking	1,058	0.77	22	19.2	3.0	1.44	107.3	11.9	5.78
	Reading	1,061	0.76	24	14.5	3.9	1.91	102.0	8.9	4.33
	Writing	1,062	0.69	20	12.7	3.0	1.64	102.4	9.4	5.21
	Comprehen	1,065	0.84	43	30.4	6.5	2.56	101.8	8.3	3.28
	Total	1,065	0.88	88	63.9	10.5	3.57	404.4	15.3	5.22

Grade 7

Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
D1	Listening	228	0.86	22	14.1	5.1	1.92	94.1	11.9	4.50
	Speaking	220	0.91	22	14.1	6.5	1.98	93.2	18.0	5.49
	Reading	229	0.81	15	9.9	3.6	1.54	93.0	13.9	5.99
	Writing	228	0.83	15	9.8	3.3	1.40	94.2	14.7	6.14
	Comprehen	229	0.89	32	19.9	7.2	2.39	93.4	11.1	3.68
	Total	229	0.95	74	47.2	16.9	3.74	383.9	24.7	5.46

D2	Listening	945	0.72	22	18.7	2.8	1.51	106.2	10.4	5.55
	Speaking	940	0.74	22	19.3	2.6	1.33	107.2	11.1	5.63
	Reading	945	0.74	24	15.8	3.6	1.87	104.9	8.9	4.56
	Writing	945	0.62	20	13.4	2.7	1.66	104.6	9.3	5.74
	Comprehen	946	0.82	44	32.7	5.7	2.41	104.9	8.2	3.50
	Total	946	0.85	88	67.0	9.0	3.43	409.2	14.9	5.70

Grade 8

Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
D1	Listening	232	0.87	22	13.8	5.3	1.94	93.8	12.7	4.61
	Speaking	225	0.90	22	13.8	6.7	2.09	92.6	19.2	6.02
	Reading	230	0.81	15	10.0	3.5	1.50	93.3	13.1	5.68
	Writing	229	0.81	15	9.9	3.3	1.41	94.8	14.0	6.05
	Comprehen	232	0.89	32	19.8	7.2	2.39	93.3	11.2	3.71
	Total	232	0.95	74	46.8	16.8	3.83	383.6	24.5	5.59

D2	Listening	851	0.79	22	19.0	2.9	1.31	107.9	10.9	5.00
	Speaking	850	0.77	22	19.3	2.9	1.37	107.6	11.5	5.51
	Reading	852	0.78	24	16.3	3.8	1.78	106.4	9.5	4.45
	Writing	855	0.69	20	13.9	2.9	1.63	106.4	10.3	5.76
	Comprehen	855	0.85	44	33.4	6.0	2.32	106.2	8.9	3.40
	Total	856	0.89	88	68.2	10.4	3.38	411.8	16.7	5.44

Grade 9				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E1	Listening	222	0.86	22	12.7	5.2	1.94	87.8	13.0	4.84
	Speaking	219	0.91	22	12.0	6.8	2.05	86.0	17.7	5.31
	Reading	218	0.83	15	8.4	3.7	1.52	87.8	13.3	5.48
	Writing	216	0.83	15	6.5	3.5	1.43	87.1	14.2	5.84
	Comprehen	223	0.89	32	17.3	7.3	2.40	87.0	12.2	4.00
	Total	224	0.95	74	38.7	17.3	3.82	376.0	21.3	4.71
E2	Listening	807	0.78	22	18.6	3.1	1.46	103.5	11.2	5.30
	Speaking	798	0.80	22	19.2	3.0	1.35	105.6	11.6	5.17
	Reading	807	0.78	25	16.3	4.4	2.07	102.0	9.8	4.57
	Writing	798	0.73	20	11.7	3.2	1.66	102.0	9.9	5.16
	Comprehen	807	0.86	44	33.1	6.7	2.52	102.1	9.3	3.49
	Total	808	0.91	89	65.4	11.9	3.62	403.3	14.8	4.52

Grade 10

Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E1	Listening	205	0.87	22	13.5	5.1	1.88	89.4	12.1	4.46
	Speaking	204	0.89	22	14.0	6.1	2.06	91.9	15.2	5.16
	Reading	199	0.80	15	8.7	3.3	1.46	88.5	11.6	5.17
	Writing	199	0.76	15	7.1	3.0	1.46	89.5	11.5	5.61
	Comprehen	206	0.88	32	18.3	6.8	2.38	88.2	10.2	3.61
	Total	207	0.93	74	42.4	15.0	3.87	380.3	17.0	4.39
E2	Listening	716	0.79	22	19.2	2.6	1.21	105.5	10.9	5.01
	Speaking	713	0.81	22	19.4	3.0	1.31	106.6	11.6	5.12
	Reading	716	0.78	25	17.3	4.1	1.91	104.0	9.6	4.45
	Writing	719	0.68	20	12.3	3.0	1.71	103.7	9.6	5.40
	Comprehen	721	0.85	45	34.4	6.3	2.48	103.7	8.7	3.42
	Total	721	0.89	89	67.7	10.8	3.52	406.1	13.9	4.52

Grade 11

Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E1	Listening	140	0.87	22	14.8	4.9	1.78	92.5	12.3	4.51
	Speaking	141	0.86	22	15.1	5.4	2.00	94.2	13.9	5.18
	Reading	137	0.77	15	9.7	2.9	1.37	91.7	10.4	5.00
	Writing	136	0.81	15	7.8	3.0	1.34	92.4	11.9	5.27
	Comprehen	141	0.87	32	20.2	6.5	2.30	91.5	10.7	3.83
	Total	142	0.93	74	46.4	13.9	3.74	385.0	16.8	4.53
E2	Listening	517	0.78	22	19.1	2.9	1.36	105.7	11.6	5.50
	Speaking	506	0.85	22	19.7	2.9	1.10	108.0	11.8	4.52
	Reading	517	0.80	25	17.7	4.3	1.96	105.2	10.2	4.62
	Writing	516	0.71	20	12.6	3.2	1.72	104.7	10.3	5.54
	Comprehen	518	0.87	45	34.9	6.6	2.43	104.9	9.8	3.60
	Total	518	0.91	89	68.4	11.8	3.51	407.9	15.8	4.70

Grade 12				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E1	Listening	82	0.86	22	14.4	5.1	1.92	91.7	13.0	4.93
	Speaking	82	0.86	22	14.6	5.7	2.12	92.9	13.6	5.06
	Reading	80	0.79	15	9.8	3.1	1.42	92.2	11.0	5.03
	Writing	80	0.74	15	7.9	2.7	1.37	92.7	10.4	5.30
	Comprehen	82	0.88	32	20.3	6.9	2.36	91.7	11.4	3.91
	Total	82	0.94	74	46.2	14.8	3.72	385.0	18.1	4.53
E2	Listening	431	0.81	22	19.5	2.5	1.09	107.3	11.4	5.00
	Speaking	427	0.84	22	19.5	3.0	1.22	107.5	12.0	4.81
	Reading	432	0.81	25	17.6	4.4	1.90	105.3	10.7	4.66
	Writing	434	0.70	20	12.7	3.1	1.71	105.0	10.1	5.56
	Comprehen	435	0.85	45	35.0	6.4	2.45	105.2	10.1	3.87
	Total	435	0.90	89	68.6	11.0	3.55	408.0	15.3	4.97

Validity of the IELA

Content-related Validity. Validity of the IELA begins with test content. The following excerpt from the *Mountain West Consortium Foundation Document* provides background information on the design of the assessment.

Mountain West Assessment Consortium Foundation Document *Introduction*

The *Mountain West Assessment Consortium Foundation Document* is part of a response to the No Child Left Behind Act (NCLB) of 2001 that mandates assessment of English language learners' progress in attaining proficiency of academic English. Since regular state assessments may not accurately reflect the gains English language learners have made in attaining English proficiency, the Mountain West Assessment Consortium has developed an English language proficiency assessment to serve a dual purpose: to measure students' language proficiency and to measure students' progress toward meeting state standards. Through the development and administration of this assessment, Mountain West Consortium states will satisfy the NCLB requirements for monitoring the development of English proficiency of the English language learners in their public schools.

The *Mountain West Assessment Consortium Foundation Document* describes the elements of language proficiency that are the basis for the Mountain West Assessment Consortium's English Language Proficiency Assessment. The purpose of the assessment is to gauge English language learners' progress in learning to listen to, speak, read, and write in the English language. The assessment follows a developmental progression across and within distinct grade spans. It is based on five communication standards recognized as the linguistic underpinnings of language: phonology, morphology, vocabulary, syntax, and function. The standards have been further detailed in benchmark performance descriptors.

Standards and benchmark descriptors are common elements of any framework that describes what students should know and be able to do. Standards are like umbrellas; they are broad-based, encompassing a set of related skills and/or knowledge bases. Benchmarks are more specific statements that describe discrete tasks students will perform in order to demonstrate knowledge or skills within a standard. For example, under the vocabulary standard in reading, one benchmark descriptor is, "*Reads and understands common idioms.*"

The Mountain West Assessment Consortium English Language Proficiency Assessment includes separate modules for children at these grade spans: kindergarten through early first grade; mid-first grade through second grade; third

grade through fifth grade; sixth grade through eighth grade; ninth grade through twelfth grade. Within each of these designated grade spans, assessment items have been developed to evaluate growth in English language acquisition across three broad developmental levels: early acquisition, intermediate, and transitional. The assessment battery modules include test items at each of the three developmental levels across the four modalities of listening, speaking, reading, and writing.

It is important to emphasize the breadth of these developmental levels and to recognize that they are not proficiency levels. The developmental levels of the standards are intentionally broad; they are used simply to make general classifications of test items within the assessment. Proficiency or performance levels specify what a student has achieved or demonstrated *relative to a set of standards*. There may, in fact, be as many as five distinct proficiency levels within these three broad developmental levels. Proficiency or performance levels are determined through standard-setting activities that yield cut-scores within the total range of test scores. There are several ways to determine proficiency levels, and each state that elects to use the Mountain West Assessment Consortium English Language Proficiency Assessment will apply its own process to determine proficiency levels.

Benchmarks have been grouped within five standards to reflect the dimensions of communicative competency:

- Phonology/Orthography standards are used to evaluate students' progress in understanding and correctly manipulating the sound system of English.
- Morphology standards are used to evaluate students' progress in understanding and using the rules of English word formation.
- Vocabulary standards are used to evaluate students' understanding and appropriate use of English words and phrases (semantic knowledge).
- Syntax standards are used to evaluate students' progress in understanding and using the rules of English sentence formation.
- Function/Discourse standards are used to evaluate students' ability to use and comprehend English in various oral and written contexts.

Since elements of some standards must be in place before others develop, the application of these five language standards varies across both grade spans and developmental levels. For example, phonology benchmarks are generally addressed more extensively at the early acquisition level than at intermediate or transitional levels. In addition, the requirements for competency in the four modalities (listening, speaking, reading, and writing) vary so that one modality may emphasize some standards over others. For example, expectations for syntax use are more pronounced in the language production modalities of speaking and writing. Similarly, assessment of function/discourse skills is addressed in greatest depth at the transitional level.

All of the standards and benchmarks included in this document are addressed in the assessment. The majority of the benchmarks are addressed in specific assessment tasks. Other benchmarks are addressed indirectly through holistic acts of listening, speaking, reading, or writing. In the receptive processes of listening or reading,

acquisition of some benchmarks is inherent in demonstrations of comprehension of the language presented. Holistic scoring rubrics have been developed to encompass such benchmarks in the language production modalities of speaking and writing.

The order in which progress across the four language modalities is assessed also reflects a developmental perspective. The modalities generally considered informal –listening and speaking– precede assessment of the more formal language modalities of reading and writing. Moreover, since a degree of language comprehension generally precedes language production, receptive language skills are addressed before production skills in both informal and formal order in the assessment. Thus, listening skills are assessed first, followed by speaking, reading, and writing skills in that order.

The developmental continuum is also reflected in this assessment in the degree to which language is decontextualized. At the early acquisition level, care has been taken to provide directions that are simple and concrete. Demonstration and practice items are also provided to help students understand what is expected of them. In addition, language in the test directions for intermediate and transitional level items begins to approximate the language found in mainstream assessments.

More detailed information about the content of the assessment is included in the full text of the *Mountain West Consortium Foundation Document*.

In addition to test design considerations, test results also bear on the content validity of the assessment. In very general terms, the distribution and range of scores within each grade cluster and grade level (Table 2, pages 6-10) provide evidence that the IELA can capture a range of abilities. And, Table 3 (on page 14) provides information on the validity of the assessment showing intercorrelations among components of the test. This table shows, by grade cluster and by test form, Pearson product moment correlations among scale scores on each subtest (Listening, Speaking, Reading, Writing, Comprehension). Correlations are not reported for subtests that share common items (e.g., Reading and Comprehension) nor are they reported for subtests and Total IELA. The number below the correlation coefficient in each cell represents the number of students on which the correlation is based.

Table 3. Correlations Among Scale Scores on Individual Language Domain Tests

Grade	K	1-2		3-5		6-8		9-12		
r	A	B1	B2	C1	C2	D1	D2	E1	E2	Avg.
L x S	0.72 2,040	0.57 1,048	0.36 2,642	0.58 1,053	0.33 3,454	0.60 683	0.27 2,840	0.62 641	0.39 2,433	0.49
L x R	0.53 2,026	0.61 1,041	0.48 2,469	0.66 1,052	0.54 3,458	0.67 703	0.56 2,847	0.70 631	0.58 2,462	0.59
L x W	0.31 1,967	0.59 1,035	0.51 2,465	0.65 1,052	0.49 3,458	0.62 701	0.46 2,852	0.69 628	0.52 2,458	0.54
S x R	0.52 2,021	0.54 1,037	0.34 2,462	0.57 1,048	0.34 3,456	0.58 681	0.27 2,840	0.62 627	0.44 2,434	0.47
S x W	0.31 1,960	0.62 1,031	0.44 2,461	0.63 1,049	0.36 3,454	0.65 680	0.22 2,845	0.62 624	0.39 2,433	0.47
S x C	0.71 2,044	0.55 1,058	0.40 2,464	0.62 1,054	0.38 3,463	0.62 684	0.29 2,847	0.62 643	0.46 2,443	0.52
R x W	0.42 1,946	0.70 1,045	0.60 2,519	0.67 1,050	0.61 3,458	0.71 702	0.63 2,853	0.68 629	0.68 2,458	0.63
W x C	0.40 1,971	0.65 1,046	0.60 2,541	0.71 1,053	0.62 3,465	0.70 702	0.63 2,861	0.40 1,971	0.66 2,467	0.60
Avg.	0.49	0.60	0.47	0.64	0.46	0.64	0.42	0.62	0.52	0.54

All of the correlation coefficients in Table 3 are significantly different from zero. In addition, all are high enough to suggest that the individual subtests are assessing related abilities but low enough to suggest the abilities are not identical. There is one relatively systematic result in this table that deserves mention. The average correlations for each test form, shown in the bottom row, reveal that in each grade cluster where two forms were administered, the average correlation for the level 1 form is higher than the average for the level 2 forms. Looking within each grade cluster and comparing the correlations for pairs of subtests on form 1 and form 2 shows that, for any pair of subtests including listening or speaking (the first 6 rows of the table), the correlation is lower on form 2 than on form 1. There are differences between forms 1 and 2 for the other two pairs (RxW and WxC) but those differences are smaller and less systematic. The reason for the disparity in correlations between forms 1 and 2 is likely due to the very high level of performance in Listening and Speaking on level 2 forms. An examination of Table 2 shows that, in each grade cluster, the average raw scores on listening and speaking are higher on form 2 than on form 1. This makes sense insofar as the Listening and Speaking tests on level 1 and 2 forms were identical and students taking level 2 forms had a higher level of English proficiency. The scores on level 2 forms were so high though (averaging 18.5 and 19.0 correct in Listening and Speaking, respectively, out of a possible 22 on each test) that the correlations of those subtests with other subtests were attenuated.

Criterion-related Validity. Table 4 (on page 16) shows, for each grade cluster and LEP group, the number of students to whom the test was administered (N) and mean and standard deviation of the scale scores for each language domain plus comprehension and the total test. These data are collapsed over grades and test forms within a grade cluster. Several points can be made from reviewing this table. First, for each grade cluster, a large majority of students who were administered the IELA were in the LEP rather than LEP1 or LEPX group. The proportion of LEP1 students was somewhat higher in Kindergarten than in other grade clusters. Second, in each grade cluster and for each language domain test and the total test, scores for LEPX students were higher on average than either LEP or LEP1. This difference was smaller in the higher grades, i.e., middle and high school, than in the lower grades. Third, for all grade clusters except K, scores for LEP1 students were lower on average than those of LEP students. Because LEP condition was determined independently of scores on this test and is based on criteria related to English proficiency, the differences in scores by LEP condition can be used as a source of criterion-related validity.

Table 4. LEP Groups Scale Scores by Grade Cluster

LEP1				LEP			LEPX		
IELA-A	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Listening	734	99.4	21.2	1,320	102.1	21.1	23	121.3	17.2
Speaking	727	99.4	24.1	1,319	101.7	23.9	23	121.0	19.0
Reading	726	101.0	26.8	1,305	100.4	30.0	23	124.1	25.7
Writing	725	100.7	31.6	1,258	106.1	32.7	23	134.9	27.6
Comprehen	736	99.0	20.3	1,323	101.4	21.2	23	121.0	15.7
Total	740	398.2	36.3	1,331	400.5	37.1	23	441.8	28.5

IELA-B	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Listening	251	101.3	16.1	3,284	113.1	15.6	253	122.7	17.4
Speaking	245	92.2	32.0	3,278	115.3	19.8	252	124.2	18.2
Reading	254	96.4	22.8	3,325	112.9	18.2	254	123.4	17.6
Writing	250	96.3	29.7	3,317	113.4	20.1	253	124.1	17.2
Comprehen	255	99.0	17.1	3,349	111.3	15.5	254	122.1	16.3
Total	255	389.1	53.1	3,350	430.6	40.1	254	457.7	35.8

IELA-C	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Listening	218	95.8	17.5	4,327	109.4	13.5	499	117.3	13.6
Speaking	218	92.3	25.6	4,303	111.6	16.1	497	119.0	14.2
Reading	219	96.8	18.7	4,308	109.3	14.1	498	118.9	13.0
Writing	217	95.0	23.1	4,305	109.5	14.9	498	118.9	14.2
Comprehen	220	95.7	17.4	4,338	108.5	12.2	499	117.4	12.0
Total	220	388.1	36.7	4,338	416.7	22.9	499	434.2	21.2

IELA-D	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Listening	230	90.6	14.0	3,335	104.0	11.3	494	109.0	10.2
Speaking	214	86.1	22.7	3,321	105.6	12.8	494	111.0	11.4
Reading	230	89.3	13.7	3,335	102.8	10.4	494	107.1	9.4
Writing	229	91.1	18.7	3,338	103.0	10.7	495	107.8	10.3
Comprehen	230	90.0	12.3	3,347	102.7	9.5	495	107.2	8.5
Total	230	375.3	30.3	3,348	405.1	18.2	495	415.1	15.8

IELA-E	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Listening	200	86.9	13.7	2,922	103.0	12.5	372	108.0	11.7
Speaking	199	83.7	16.7	2,893	104.7	13.1	356	109.9	11.3
Reading	198	85.9	12.9	2,910	101.9	11.2	372	105.7	9.7
Writing	196	86.7	14.6	2,904	101.7	11.2	372	106.2	10.5
Comprehen	201	85.9	12.3	2,934	101.6	10.8	373	105.6	9.0
Total	202	374.2	21.7	2,937	402.4	17.4	375	408.9	15.6

Standard Setting

A formal IELA Standard Setting was conducted in August 2006. That process involved 25 Idaho educators divided into two panels: One panel focused on the lower grades, K, 1-2, 3-5, and the second panel focused on middle and high school grades, 6-8, 9-12.

Panel members received books containing test items for a particular grade span, with each page corresponding to a test item and pages ordered in terms of increasing item difficulty. Using the Bookmark or item mapping procedure, panelists made “cuts” by placing markers in the books to indicate the item on which 50% of the students at a particular proficiency level and in a particular grade would answer correctly. Three rounds of cuts were planned for each grade span. Following each of the first two rounds, panelists were shown frequency distributions and medians of recommended cuts and were given the opportunity to discuss the process. The second round was followed by impact data, i.e., the percent of students in each grade who would be placed in each proficiency level based on the median cuts assigned by the group. The third round of cuts was accepted as the panelists’ final recommendations.

Final recommendations were adjusted to eliminate minor variations within grade clusters. For example, minor adjustments to recommended cut scores in 8th grade resulted in one set of cut scores for the 6-8 grade cluster. A similarly minor adjustment in grades 9 and 10 produced one set of cut scores for the grade 9-12 cluster. A second reason for adjustments was to create a more consistent pattern of proficiency levels across the grades. The panelists’ final recommendations resulted in disparities over grades in the percent of students at different proficiency levels. This outcome is not uncommon when there are different panels working on different grade clusters. The second set of adjustments was made to reduce these disparities.

Table 5a (on page 18) shows, for each form and grade, the range of IELA scale scores corresponding to each proficiency level.

Table 5a. Total IELA Scale Scores Corresponding to Proficiency Levels

Total IELA Proficiency Levels						
Form	Grade	Beginning	Advanced Beginning	Intermediate	Early Fluent	Fluent
A	K	Below 362	362-380	381-399	400-424	425 and above
B1 or B2	1	Below 345	345-371	372-399	400-424	425 and above
	2	Below 354	354-384	385-424	425-465	466 and above
C1 or C2	3	Below 359	359-379	380-399	400-424	425 and above
	4	Below 362	362-382	383-414	415-433	434 and above
	5	Below 370	370-389	390-416	417-437	438 and above
D1 or D2	6-8	Below 357	357-373	374-399	400-424	425 and above
E1 or E2	9-12	Below 364	364-375	376-399	400-424	425 and above

Table 5b shows scale score ranges corresponding to proficiency levels in each of the language domains (Listening, Speaking, Reading, Writing) and Comprehension. In the case of language domain tests, three proficiency levels are reported. Individual language domain tests do not include enough items to reliably report more than three levels of proficiency.

Table 5b. Language Domain IELA Scale Scores Corresponding to Proficiency Levels

Language Domain Proficiency Levels				
Form	Grade	Beginning	Advanced Beginning to Intermediate	Early Fluent and above
A	K	Below 80	80-99	100 and above
B1 or B2	1	Below 80	80-99	100 and above
	2	Below 83	83-108	109 and above
C1 or C2	3	Below 80	80-99	100 and above
	4	Below 81	81-106	107 and above
	5	Below 85	85-107	108 and above
D1 or D2	6-8	Below 80	80-99	100 and above
E1 or E2	9-12	Below 80	80-99	100 and above

Alignment Study

An alignment study was conducted in September 2006 for the purpose of determining the extent to which IELA 2006 and 2007 test forms (adaptations of Mountain West Forms 1 and 2, respectively) were aligned with Idaho English Language Proficiency Standards. The results of that study will be presented in a separate report.

Glossary of Terms

Alpha - Coefficient alpha is a measure of internal consistency reliability based on the average inter-item correlation. Coefficient alpha can vary from 0 to 1.0, with a higher value indicating a more reliable test.

Raw Score - The raw score is the total number of points earned on the test, determined by summing the number of correct answers on multiple-choice items and the number of points earned on open-ended items.

Reliability - The reliability of a test refers to the extent to which it produces consistent, stable results. Test reliability is typically expressed as a reliability coefficient (see Alpha) or by the standard error of measurement (see SEM).

Scale Score - A type of derived score, which is a transformation of the raw score, developed through a process called scaling. Scale scores provide a basis for comparing scores on different forms of a test (e.g., C1 and C2). Scale scores on the IELA cannot be compared across test levels (e.g., B and C) or across different subtests (e.g., Listening and Writing).

SEM - The standard error of measurement is a statistic used to indicate the amount by which a score might vary due to errors of measurement. It can be thought of as the standard deviation of an individual's observed scores from repeated administrations of a test (or parallel forms of a test). Standard error of measurement is estimated using information about the reliability and standard deviation of a set of test scores.

Validity - The validity of an assessment is the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by the purposes for which the test is used. Validity is commonly defined as the extent to which a test measures what it is intended to measure. Content-related validity refers to the extent to which a test represents a balanced and adequate sampling of the content domain in terms of the knowledge, skills, and objectives assessed. Criterion-related validity refers to the extent to which a test is a measure of a particular criterion. It can be assessed in terms of how well test results predict performance on some future criterion measure or are in agreement with the results of some current criterion measure.



IDAHO STATE BOARD OF EDUCATION

650 W. State Street • P.O. Box 83720 • Boise, ID 83720-0037

208/334-2270 • FAX: 208/334-2632

e-mail: board@osbe.idaho.gov