



Institute for Test Development and Research

Professor Dr. Erwin Tschirner

Dr. Olaf Bärenfänger

 0341 9737 570

 0341 9737 547

Universität Leipzig, Herder-Institut, Beethovenstraße 15, 04107 Leipzig

Leipzig, August 15, 2013

Validating the ACTFL Listening and Reading Proficiency Computer Adaptive Test (ACTFL L&R CAT)

Technical Report 2013-US-PUB-4

Prepared for:

American Council on the Teaching of Foreign Languages
Washington, D.C.

Language Testing International
White Plains, NY

Prepared by:

Institute for Test Research and Development

Dr. Erwin Tschirner
Gerhard-Helbig-Professor of German as a Foreign Language

Dr. Olaf Bärenfänger
Director, Language Learning Centre



Validating the ACTFL Listening and Reading Proficiency Computer Adaptive Test (ACTFL L&R CAT)

Erwin Tschirner, Olaf Bärenfänger, Sabine Kutschera, Jupp Möhring

Executive Summary

Introduction

In 2013, *Language Testing International Inc.* (LTI) and ACTFL developed a computer-adaptive version of the *ACTFL Reading Proficiency Test* (ACTFL RPT) and the *Listening Proficiency Test* (ACTFL LPT). The rationale underlying a computer adaptive test (CAT) is that the person ability parameter, i.e. the estimate of a test taker's actual reading or listening ability, continues to be updated during test administration until the measurement is sufficiently precise. The difficulty of the test items is continuously adjusted to the then current estimate of the test taker's reading or listening ability. The CAT developed by LTI/ACTFL uses the single-faceted Rasch model for dichotomous items both for calculating the person ability estimate and for targeting test items according to the test taker's predicted level of ability.

A study with 201 participants at all levels of proficiency was conducted at four Credu Test Centers in Seoul, South Korea, from July 24 – 25, 2013. The study was intended to provide answers to the following research questions:

1. Does item targeting function as expected?
2. Is the final estimate of the CAT sufficiently precise?
3. How do test takers assess the difficulty and usability of the CAT?
4. What additional external indicators are there to validate the levels of the CAT?

Method

Subjects. From July 24-25, 2013, the listening part of the ACTFL L&R CAT was administered to 201 participants and the reading part to 200 participants. The great majority of these participants were university students or recent graduates.

Procedure. Participants first took the listening part of the ACTFL L&R CAT and then the reading part under proctored conditions in a computer lab. Each part of the test consisted of 27 items, the maximum time available for each part was 50 minutes. Before logging on to the ACTFL L&R CAT, participants were shown a short introductory video in order to familiarize them with the test. At the end of the test, participants completed a paper-and-pencil questionnaire with questions relating to biographical information, TOEIC scores, test usability, and self-evaluation of reading and listening competence.

Design. For each participant and each test item, the CAT provides the difficulty of the individual item chosen for the test taker, a preliminary person ability estimate, and the standard error of measurement (SEM) of the preliminary person ability estimate. The complete, item-by-item test-taking process was analyzed tracking all item difficulty values, the developing person ability estimate, and the associated decrease in the standard error of measurement. In addition, the distribution of final ratings was computed.

Biographical Information

The following is a summary of the information provided by Credu based on 178 completed questionnaires. 96% of the participants were 21-30 years of age, 4% were over 30. 61% were female, 39% were male. 40% were still attending university, 60% were university graduates. 93% were students or recent graduates looking for a job, 7% were employed.

Data Analysis

Three datasets were eliminated from the listening part of the L&R CAT and six datasets were eliminated from the reading part because participants had obviously completed them without paying attention to the test. Participants who spent six minutes or less on the reading part or 11 minutes or less on the listening part were eliminated. All of the six reading tests that were deleted had a rating of 0. One of the three listening tests had a rating of 0, the other two had ratings of NL and NM. Two additional reading test sets were excluded from analysis because of some kind of problem surfacing during test administration which resulted in two testlets (six items) not receiving any input. Thus, there were 198 listening and 192 reading datasets available for analysis.

Ratings

Table 1 shows the ratings test takers received in the two parts of the L&R CAT.

Table 1
Ratings: All Levels

	Test	
	Listening	Reading
0	0	2
Novice Low	0	6

Novice Mid	0	8
Novice High	4	5
Intermediate Low	27	51
Intermediate Mid	59	40
Intermediate High	30	19
Advanced Low	71	43
Advanced Mid	7	17
Advanced High	0	1
Superior	0	0
Total	198	192

The median ACTFL listening level was Intermediate High and the median ACTFL reading level was Intermediate Mid.

Table 2 shows the numbers and percentages of ratings test takers received at the four major levels of the ACTFL scale. The ratings of 0 and NL were collapsed in one rating, i.e., NL.

Table 2
Ratings: Major Levels

	Listening		Reading	
	<i>N</i>	Percentage	<i>N</i>	Percentage
Novice	4	2.0	21	10.9
Intermediate	116	58.6	110	57.3
Advanced	78	39.4	61	31.8
Superior	0	0	0	0
Total	198		192	

Table 2 shows that most test takers received a final rating of Intermediate in both listening and reading (close to 60 per cent) followed by a final rating of Advanced (close to 40 per cent in listening and a little over 30% in reading). Very few test takers received a rating of Novice and no one received a rating of Superior.

Research Questions

Question 1: Does item targeting function as expected?

Item targeting functions as expected if the standard error of measurement (SEM) continually decreases from the first item administered to the participant to the last item. An analysis of all 198 listening and 192 reading test sets showed that the SEM continually decreases from the first to the last item.

Item targeting, accordingly, functions as expected.

Question 2: Is the final estimate of the CAT sufficiently precise?

The final estimate of the CAT is sufficiently precise, if the standard error of measurement (SEM) reaches the required logit level after the last item administered.

Table 3 shows the proportion of participants who reached the required SEM. The CAT provides a very precise estimation of the person ability parameter, if it reaches that logit level.

*Table 3
Final Value of the Standard Error of Measurement (SEM)*

SEM	Listening		Reading	
	N	Percentage	N	Percentage
Required Logit Level	190	96%	188	98%
Within 20% of the required Level	8	4%	4	2%
Within 33% of the required level	0	0%	0	0%
Total	198	100%	192	100%

As Table 3 shows, the final estimate of the CAT reached the required SEM in all but a very few number of cases (96% for listening and 98% for reading).

The final estimate of the CAT, therefore, is clearly sufficiently precise for high stakes testing.

Question3: How do test takers assess the difficulty and usability of the CAT?

The following is a summary of the questionnaire analysis completed by Credu. The analysis included feedback from 178 test takers.

Table 4 shows the responses test takers gave with respect to the difficulty and usability of the listening part of the L&R CAT.

*Table 4
Difficulty and Usability of Listening Part*

	Agree	Average	Disagree
User-friendly	52%	33%	15%
Good comprehensibility of passages	57%	19%	25%
Enough time to answer questions	85%	13%	3%
Good assessment tool	46%	42%	13%
	Difficult	Appropriate	Easy
Difficulty of questions	45%	46%	10%

52% of the participants thought that the listening part of the L&R CAT was “user-friendly”, 33% thought it was average, 15% did not think that it was user-friendly. 57% felt that they understood most of the passages they listened to reasonably well, 19% felt that their understanding was average, 25% did not understand most of them well enough. 85% thought they had enough time to respond to questions in the listening part of the test, 13% felt the time they had was average, and only 3% did not feel they had enough time. 46% thought that the listening part of the test is a good tool for assessing listening competence, 42% thought it is an average tool, and 13% did not think it was a good tool. 45% felt that the level of difficulty of most questions was high, 46% thought that their level was appropriate for their level of proficiency, and 10% felt the questions were easy.

In general, a large majority of the test takers thought that the listening part was user-friendly, a good assessment tool and that they had enough time for the test. Equal proportions of test takers thought the questions were difficult or appropriate, very few thought that they were easy or too easy.

Table 5 shows the responses test takers gave with respect to the difficulty and usability of the reading part of the L&R CAT.

Table 5
Difficulty and Usability of Reading Part

	Agree	Average	Disagree
User-friendly	35%	26%	40%
Good comprehensibility of texts	47%	25%	27%
Enough time to answer questions	66%	28%	8%
Text window big enough	29%	28%	45%
Good assessment tool	36%	51%	13%
	Difficult	Appropriate	Easy
Difficulty of questions	50%	42%	8%

35% of the participants thought that the reading part of the L&R CAT was “user-friendly”, 26% thought it was average, 40% did not think that it was user-friendly. 47% felt that they understood most of the texts they read reasonably well, 25% felt their understanding was average, 27% did not understand most of them well enough. 66% thought they had enough or more than enough time to respond to questions in the reading part of the test, 28% felt it was average, and 8% did not feel they had enough time. 29% felt that the text window was big enough for them to read comfortably, 28% thought it was average and 45% did not think it was big enough. 36% thought that the reading part of the test is a good tool for assessing reading competence, 51% thought it was average, and 13% did not think it was a good tool. 50% thought that the level of difficulty of most questions was too high for their level of proficiency, 42% felt the level was appropriate and 8% felt the questions were easy.

As with the listening part, a large majority thought the reading part was a good assessment tool and that they had enough time to answer the questions. The proportions of test takers saying the questions were appropriate, difficult, or easy was also similar as was the number of test takers agreeing with the statement that most texts were comprehensible enough. Test takers differed in their evaluation of the listening and reading part, however, with respect to its user-friendliness: while 60% found it good or average, a large minority of 40% did not find the reading part user-friendly. The most likely reason for this may be found in their answers to the question if they thought the text window was big enough. 45% disagreed with this statement.

Question 4: What additional external indicators are there to validate the CAT?

All L&R CAT texts and passages as well as test items were thoroughly validated in pilot studies as were the ACTFL levels associated with particular item logit ranges (Bärenfänger & Tschirner 2013a; Bärenfänger & Tschirner 2013b; Swender, Tschirner & Bärenfänger 2012). The goal of the present study was not to provide validity evidence for logit values. However, the study yielded a few additional external indicators of the general validity of the L&R CAT which will be discussed in this section.

To provide additional external validity evidence, the following relationships were investigated in the present study:

- The relationship between ACTFL levels and the period of time test takers visited or lived in an English-speaking country;
- The relationship between ACTFL levels and participants' self-evaluation of their listening and reading proficiency (can-do statements).

The relationship between ACTFL levels and the period of time test takers visited or lived in an English-speaking country.

Test takers were asked to indicate how much time they spent in an English-speaking country prior to taking the L&R CAT on a scale ranging from “never”, “up to three months”, “three months to one year”, “one to two years” to “more than two years. Table 6 shows the results for the listening and Table 7 the results for the reading part.

Table 6
Time spent in an English-Speaking Country: Listening

	NH	IL	IM	IH	AL	AM	Total
None	2	12	22	10	15	1	62
Up to 3 months		7	8	4	9		28
3 months to 1 year	1	2	17	3	16	1	40
1 to 2 years		2	1	3	4	2	12
More than 2 years			1		5	2	8
Total	3	23	49	20	49	6	150

Table 7
Time spent in an English-Speaking Country: Reading

	0-NM	NH	IL	IM	IH	AL	AM	AH	Total
None	8	2	17	15	4	10	5		61
Up to 3 months	2	1	8	8	2	3	3		27
3 months to 1 year	3		13	3	6	12	2		39
1 to 2 years			1	3		4	2	1	11
More than 2 years			2	1		5			8
Total	13	3	41	30	12	34	12	1	146

A one-way between subjects ANOVA was conducted to compare the effect of time spent in an English-speaking country on the results of the listening and the reading parts of the L&R CAT. There was a significant effect of time on the results of the listening part at the $p < .05$ level [$F(1,148)=9.425, p=0.003$]. Similarly, there was a significant effect of time on the results of the reading part at the $p < .05$ level [$F(1,144)=6,661, p=0.019$].

Figures 1 and 2 show the results of the ANOVA for listening and reading.

Figure 1
Time in English-Speaking Country: Listening

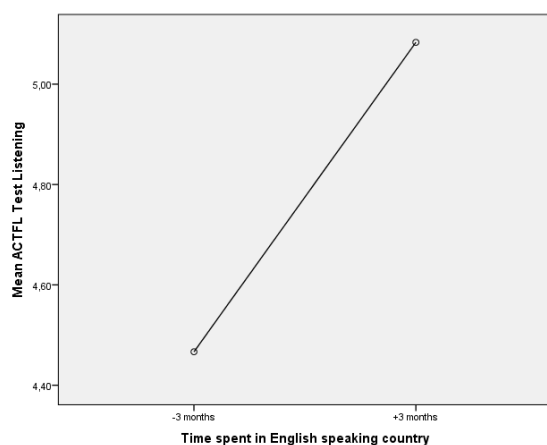
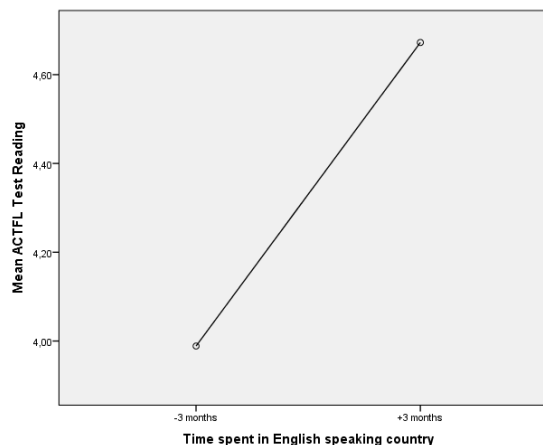


Figure 2
Time in English-Speaking Country: Reading



The relationship between ACTFL levels and participants' self-evaluation (can-do statements)

In the questionnaire, test takers were asked to evaluate their listening and reading proficiency in the form of can-do statements. Test takers were asked if they agreed or disagreed with five can-do statements relating to their listening proficiency and five statements relating to their reading proficiency. The can-do statements were taken from the DIALANG self assessment statements (Council of Europe 2001: Appendix C). The levels A2 (IM), B1 (AL) and C1 (S) were represented by one statement for each skill, the level B2 (AM) was represented by two statements. The ten can-do statements can be seen below. They are ordered by the item number on the questionnaire. Their level is indicated at the end of the statement in parenthesis.

20: I can understand films which contain a considerable degree of slang and idiomatic usage. (C1)

21: I can understand most TV news and current affairs programs such as documentaries, live interviews, talk shows, plays and the majority of films in standard language. (B2)

22: I can understand simple directions relating how to get from X to Y, by foot or public transport. (A2)

23: I can understand simple technical information, such as operation instructions for everyday equipment. (B1)

24: I can understand the main ideas of complex speech on both concrete and abstract topics delivered in a standard language including technical discussions in my field of specialization. (B2)

25: I can read quickly enough to cope with the demands of an academic course. (C1)

26: I can understand articles and reports concerned with contemporary problems in which the writers adopt particular stances or viewpoints. (B2)

27: I can understand clearly written straightforward instructions for a piece of equipment. (B1)

28: I can understand standard routine letters and emails. (A2)

29: I have a broad reading vocabulary, but I sometimes experience difficulty with less common words and phrases. (B2)

Table 8 shows the number of test takers who agreed and strongly agreed (= yes) or disagreed and strongly disagreed (= no) according to their final ACTFL major level on the listening part of the test.

Table 8
Can-do Statements by ACTFL Major Levels: Listening

Item	ACTFL	Agree	N	I		A		Total
20	S	yes	2	16	28%	11	39%	29
20	S	no	1	42	72%	17	61%	60
21	AM	yes	1	21	47%	22	60%	44
21	AM	no	1	24	53%	15	40%	40
22	IM	yes	3	74	96%	52	100%	129
22	IM	no	0	3	4%	0	0%	3
23	AL	yes	2	55	77%	33	87%	90
23	AL	no	0	16	23%	5	13%	21
24	AM	yes	1	19	35%	17	49%	37
24	AM	no	1	36	65%	18	51%	55

Table 8 shows how many test takers (strongly) agreed or (strongly) disagreed with the five listening can-do statements. Note that the results of test takers who neither agreed nor disagreed have been deleted. Therefore, the number of test takers listed per can-do statement varies.

There were too few test takers at the Novice level to warrant any analysis. For both Intermediate (I) and Advanced (A), the number of test takers agreeing or disagreeing are given in the first column. The second column lists the percentage of test takers agreeing or disagreeing. A one-way between groups ANOVA was conducted for each can-do statement comparing the effect of agreeing and disagreeing on the mean final ACTFL level. There was a

significant effect of the response to item 23 (AL) on the results of the listening part of the test at the $p < .05$ level [$F(1,109)=4.375$, $p=.039$].

A closer look at Table 8, however, shows that there are clear additional trends even if not at all ACTFL levels. Item 20 (S), e.g., was answered negatively by 72% of Intermediate test takers and 61% of Advanced test takers who had either agreed or disagreed. Note that respondents who had neither agreed nor disagreed were excluded from the tally. Item 21 (AM) was answered positively by 60% of the Advanced test takers. Item 22 (IM) was answered positively by 74% of the Intermediate test takers and by 100% of the Advanced ones. Item 23 (AL), of course, was the item that distinguished significantly between all levels and item 24 (AM) was answered negatively by 65% of the Intermediate test takers.

Table 9 shows the number of test takers who agreed and strongly agreed (= yes) or disagreed and strongly disagreed (= no) according to their final ACTFL major level on the reading part of the test.

Table 9
Can-do Statements by ACTFL Major Levels: Reading

Item	ACTFL	Agree	N		I		A		Total
25	S	yes	3	38%	17	38%	26	72%	46
25	S	no	5	62%	28	62%	10	28%	43
26	AM	yes	2	20%	25	58%	21	70%	48
26	AM	no	8	80%	18	42%	9	30%	35
27	AL	yes	9	69%	55	92%	27	84%	91
27	AL	no	4	31%	5	8%	5	16%	14
28	IM	yes	16	100%	73	99%	45	100%	134
28	IM	no	0	0%	1	1%	0	0%	1
29	AM	yes	5	100%	45	94%	32	97%	82
29	AM	no	0	0%	3	6%	1	3%	4

Table 9 shows how many test takers (strongly) agreed or (strongly) disagreed with the five reading can-do statements. For Novice (N), Intermediate (I) and Advanced (A), the number of test takers agreeing or disagreeing are given in the first column. The second column lists the percentage of test takers agreeing or disagreeing. A one-way between groups ANOVA was conducted for each can-do statement comparing the effect of agreeing and disagreeing on the mean final ACTFL level. There was a significant effect of the response to item 25 (AL) on the results of the reading part of the test at the $p < .01$ level [$F(1,87)=7.292$, $p=.008$]. In addition, there was a significant effect of the response to item 26 on the mean ACTFL level at $p < .05$ [$F(1,81)=5.848$, $p=.018$].

Again, there are clear trends for additional items at least at some additional ACTFL levels. As stated, items 25 (S) and 26 (AM) distinguish significantly between all ACTFL levels. In addition, 69% of the Novice respondents, 92% of the Intermediate respondents and 84% of the Advanced respondents who either agreed or disagreed responded positively to item 27 (IM). Item 28 (IM) was positively responded to by 100% of Novice, 99% of Intermediate, and 100% of Advanced respondents. Similarly, item 29 was responded to positively by 100% of Novice, 94% of Intermediate and 97% of Advanced test takers who either agreed or disagree.

In summary, the analysis of test takers' responses to the ten can-do statements yields a significant relationship between statements and mean ACTFL levels for three statements and clear trends for all other statements for at least some ACTFL levels.

TOEIC Scores

Participants also provided TOEIC scores. These scores will be analyzed in this section to provide a fuller picture of the population of test takers in the present study. Care must be taken not to over-interpret the results because the age of the TOEIC data varies considerably.

Table 10 shows the year in which the TOEIC results were achieved, the number and the percentage of TOEIC test results for all participants who provided them.

Table 10
Year of Administration, Number and Percentage of TOEIC Results

Year	N	Percentage
2013	73	50.3
2012	45	31.0
2011	19	13.1
2010	3	2.1
2008	1	0.7
2005	1	0.7
No Information	3	2.1
Total	145	100

Table 10 shows that about 50% of the TOEIC results were recent, while the other 50% had been received in previous years.

Table 11 shows the distribution of TOEIC scores separately for listening and reading. In addition to minimum, maximum, mean, median and standard deviation, the number and percentage of

TOEIC scores are listed with respect to TOEIC score quartiles, i.e., the number of scores that fall into the lowest, the second lowest, the second highest and the highest 25 per cent of TOEIC scores. This provides an indication of how evenly distributed TOEIC scores are. Note that no separate listening and reading scores were provided for three test takers. The total number of scores analyzed, therefore, was 142.

Table 11
Distribution of TOEIC Listening and Reading Scores

	Listening		Reading	
	<i>N</i> = 142	Percentage	<i>N</i> = 142	Percentage
Quartile 1	0	0	0	0
Quartile 2	3	2.1	14	9.9
Quartile 3	20	14.1	49	34.5
Quartile 4	119	83.8	79	55.6
Minimum	200		170	
Maximum	495		495	
Mean	418.8		368.6	
Median	435		380	
SD	58.03		74.79	

Table 11 shows the skewed distribution that seems to be typical of Korean TOEIC scores. 83.8% of listening scores are in the top 25% range of TOEIC scores, i.e. scores ranging from 375 to 495. This is supported by the high mean and median scores of 418.8 and 435, respectively. Similarly, 55.6% of reading scores are in the top 25% range and 90% are in the top 50%. Again, this is corroborated by a high mean and median of 386.6 and 380, respectively. If one compares these results with the results obtained by the L&R CAT (Tables 1 and 2), one notices that ACTFL levels are much more evenly distributed across the same population of test takers with a median at IM and IH, respectively.

Table 11 also shows that listening scores are significantly higher with a difference of approximately 50 points between means and medians and a considerably higher percentage of test takers scoring in the top 25% in listening when compared to reading. This apparent superiority of the listening skill is also noticeable in the L&R CAT results.

Conclusion

There were two main goals of the Credu Korea Study from July 24-25, 2013: The first goal was to determine if the L&R CAT functions as intended to determine a precise person ability value and the second one was to get feedback on the test's usability and other features by the target

population of the test, i.e., college students and recent graduates looking for employment. In addition, the study yielded further evidence of the validity of the test which has been established in several previous studies. A total of 201 test takers participated in the study. A very small percentage of tests needed to be eliminated from analysis because of a number of reasons yielding a final tally of 198 tests of listening comprehension and 192 tests of reading comprehension.

Both, the listening and reading tests, yielded a fairly even distribution of proficiency levels of the target population. 2% of the test takers were rated Novice in listening, about 59% were rated Intermediate, close to 40% were rated Advanced and 0% were rated Superior. The median rating was Intermediate High. For reading, 11% were rated Novice, 57% were rated Intermediate, and 32% were rated Advanced. Again, there were no Superior ratings. The median rating was Intermediate Mid.

The two most important functions of the L&R CAT that were investigated were the item targeting function and the precision of the final person ability estimate. Both functions worked as expected for a high stakes English proficiency test. This convincingly demonstrates that the refinement of the algorithm initiated after the small-scale June 2013 Leipzig study had its desired effect.

Test takers were given an opportunity to comment on several features of the test, including its perceived user-friendliness, the difficulty level of reading texts, listening passages and items, the appropriate length of time given to take the test, and if they considered the L&R CAT a good assessment tool. Results were very positive and indicated a high level of agreeability to most features of both listening and reading. About half of the questions a test taker answered were considered appropriate to their level and the other half as being difficult.

To provide further external evidence of the L&R CAT's validity, two factors were evaluated in detail: the relationship between the amount of time a test taker had spent in an English-speaking country and their proficiency level as well as their own perceived proficiency estimate and their final rating. Both the descriptive statistics and a number of ANOVAs comparing time of residency and self-evaluative can-do statements with final ratings supported the validity argument established in previous studies as well as the appropriateness of the logit values established for all ACTFL levels from Novice Low to Superior.

Bärenfänger, O., & Tschirner, E. (2013a). Assessing Evidence of Validity of the ACTFL Reading Proficiency Test (RPT) (Technical Report 2013-US-PUB-1). Leipzig: ITT.

Bärenfänger, O., & Tschirner, E. (2013b). Assessing Evidence of Validity of the ACTFL Listening Proficiency Test (LPT) (Technical Report 2013-US-PUB-2). Leipzig: ITT.

Swender, E., Tschirner, E. & Bärenfänger, O. (2012). Comparing ACTFL/ILR and CEFR Based Reading Tests. In E. Tschirner, ed., *Aligning frameworks of reference in language testing: The ACTFL Proficiency Guidelines and the Common European Framework of Reference*, Tübingen: Stauffenburg, 123-138.