Development of a level-specific test based on the ACTFL Reading Guidelines: an item response theory approach

Siwon Park

Abstract

The current study attempts to validate the text and skill hierarchy of the ACTFL Reading Guidelines by examining the developmental process of a level-specific second language (SL) reading test based on the Guidelines. With the data collected from 1,060 English learners in Korea, item parameters were calibrated using item response theory for carefully constructed test items, and a side-by-side comparison was performed of the items with their difficulties and their proposed levels. Also, an ANOVA and a post-hoc comparison between levels were performed in order to examine level distinctions. The findings were examined to see if the Reading Guidelines enabled the development of a reliable level-specific test and whether or not the level specifications of the Guidelines appropriately represent proficiency progression throughout levels from Novice-Mid to Superior. The study concludes: (a) constructing a reliable test for SL reading proficiency appears feasible, when the process is carefully pursued, (b) a general progression of SL reading proficiency was observed; yet inconsistency was noticeable between adjacent levels, and (c) the level-specific characterization of the ACTFL Reading Guidelines may not be entirely valid for the given text and skill hierarchy.

I. Introduction

Potential uses of proficiency guidelines for program and test development are uncontroversial as they can provide firm bases for such purposes – i.e., being a common yardstick to evaluate a program or being a reference for program and test developments (Bachman, 1990; Brown, 1996, North & Schneider, 1998; North, 2000). Also, in regards to the recent popularity of performance assessment, the guidelines' functional descriptors of developmental characteristics of second language (SL) may suggest useful frameworks in developing assessment criteria (Mislevy, 1995).

However, despite the seemingly useful potential, a number of researchers have voiced considerable criticism about the nature of the proficiency guidelines (e.g., Bachman & Savignon, 1986; Clark, 1985; Lantolf & Frawley, 1985; Pienemann & Johnston, 1987; Spolsky, 1986). Much of such criticism points to the fact that their descriptive nature and lack of empirical evidence has rendered no guarantee that the depiction of progressive language development and the number of proficiency levels are accurate, valid, or balanced. The central question in this regard is how closely level descriptors could approximate true developmental characteristics within each and across the domains that the guidelines depict of SL development. In addition, there is a practical concern with how closely test developers could materialize level specifications in their process of SL test development.

The ACTFL Guidelines are not exceptions to this concern. Especially, with the reading portion of the Guidelines, a few researchers have endeavored empirical investigations. However, they have observed inconsistent findings, especially in relation to the hierarchy construct put forth of SL reading development (Lange & Lowe, 1987; Lee & Musumeci, 1988; Mecartty, 1998; Park, 2004). With frustration, researchers have called for more systematic analyses with large samples – e.g., using item response theory (Kaya-Carton & Carton, 1986).

1. Taxonomy as a construct

When language testers prepare assessment tools, they face two main questions: what to test and how to test. The question of *what to test* is closely related to the definition of what language is (Chapelle & Douglas, 1993). *How to test* is often referred to as how to characterize stages in the development of communicative competence (Canale, 1988). What makes *how to test* problematic is the fact that there has been little consensus about the characterization(s) of global levels of language proficiency. The most important and fundamental characteristic of the proficiency guidelines can be found in their proficiency hierarchy. However, as Lantolf and Frawley (1988) maintain, there has been much disagreement on the number of levels to include in the proficiency hierarchy.

Ingram (1985) divides tests into two types: non-developmental versus developmental. Non-developmental tests may use an *ad hoc* manner in selecting the content of the test on either a linguistic or a behavioral basis. In contrast, developmental tests view proficiency as a developmental progression from zero to native-like. Considering Ingram's distinction, the ACTFL Guidelines are developmental in their characteristics. According to Ingram, "these developmental approaches to measuring language proficiency seek to relate statements of language proficiency to the learner's overall schedule of development in the language." Therefore, test development in direct testing is closely related to the evolving language behavior of SL learners. Consequently, what is required with the ACTFL Guidelines is defining this behavior. In addition to defining taxonomy, language proficiency needs to be clarified, in order to be used as a construct, since it is proficiency that constitutes the taxonomy. As Long (1985) argues, "[t]here is no reason to assume that acquisition proceeds in one step from zero proficiency to full target form or use." Furthermore, a linear description of language development cannot be justified by simply supposing its presumed systematization.

2. Studies on the ACTFL Reading Guidelines

L2 reading involves a variety of interactions – i.e., the interaction of reader and text. Specifically, it is the interaction of an array of processes and knowledge. In his explanation of the L2 reading process, Hudson (1998) argues that there are various factors involved in reading – e.g., basic decoding skills, higher level cognitive skills, interactional skills, and first and second language literacy interactions. Simply put, as reading involves more than a few complex applications, it cannot constitute a linear process (Hudson, 1991).

ACTFL understands reading as an interaction of reader and writer (or more specifically of reader and text). Although efforts were put to integrate other interaction factors into the Guidelines, the approach adopted for reading appears too simplistic, given its complex nature. Reading proficiency, proposed as the construct for the Reading Guidelines, is made up of abstract notions – e.g., cognitive skills, perceptual skills, cultural knowledge, linguistic knowledge, general knowledge, and L1 maturity, and *their interactions*. Such notions and their interactions make reading much more difficult to understand; yet, unless the construct is defined operationally, it is impossible to investigate reading proficiency (Kaya-Carton & Carton, 1986).

Lee and Musumeci (1988) validated the ACTFL Reading Guidelines with foci on three characteristics of reading proficiency – reading skill, text types, and level – and their interactions. Those characteristics were proposed as constructs of the Reading Guidelines, and they and their interactions were analyzed to investigate the proposed developmental stages of SL reading proficiency. However, the reader performance on different text types, on reading skills, and on their cross-sectional interactions was found not to be consistent with the proposed hierarchy for each construct. In conclusion, they suggested that the ACTFL Reading Guidelines instead be considered with respect to readers' cognitive perspectives, the linguistic elements of texts, and readers' and texts' cultural and social references.

Dandonoli and Henning (1990) also undertook a validation study of the Guidelines, using a multitrait-multimethod and a Rasch analysis with French and English. They successfully demonstrated the usefulness of the Guidelines for constructing relevant tests. The convergent validity was also revealed of each of the sub-skills between French and English. In addition, a generally acceptable discriminate validity was observed between each skill modality. The Rasch analysis revealed an adequate progression in the appropriate direction. Nonetheless, their study also failed to exhibit significance in some of the skill modalities such as French listening, English listening, and English writing. Also, the Rasch analysis revealed that the reading sections of the Guidelines were not sensitive enough to distinguish between intermediate and advanced levels. Furthermore, the fact that a notably high percentage of misfitting items for the reading part resulted from the Rasch analysis asks for further investigations into the Guidelines.

Park (2004) also reports a study with the same research objectives. With the data collected from a large pool of participants and based on carefully constructed test items, he investigated the Reading Guidelines. Using Guttman scaling and ANOVA, he also evidenced the usefulness of the Guidelines for constructing a reliable test, and demonstrated an adequate progression of SL reading proficiency in the appropriate direction within a statistically acceptable range. However, as he also recognizes, his findings are rather constrained, due to the deterministic characteristics of Guttman scaling (Nunnally & Bernstein, 1994). The results of Guttman scaling did not render much item-level information in examining the hierarchy construct of the Guidelines.

II. Purpose

The current study attempts to develop a level-specific test based on the ACTFL Reading Guidelines. Dandonoli and Henning (1990) argue that if the use of the Guidelines enables development of reliable as well as valid tests, the Guidelines themselves may be considered valid for the intended purposes of academic language proficiency assessment. Therefore, the successful development of a reliable and valid reading test based solely on the Guidelines must help validate their hierarchy construct at least indirectly.

In order to achieve the research goal of such, the following research questions are proposed:

1) Does the use of the ACTFL Reading Guidelines enable the development of a reliable level-specific SL reading test?
2) Do the level specifications of the ACTFL Reading Guidelines appropriately represent proficiency progression based on the scales throughout Levels from Novice-Mid to Superior?

Research Question 1 mainly concerns the reliability of the test developed based on the ACTFL Reading Guidelines. Research Question 2 seeks to answer the question regarding the validity of the overall hierarchical structure of the ACTFL Reading Guidelines. Initially it was intended to examine the proficiency progression of SL reading from Novice (0 ability in the ILR scales) to Distinguished (no specification in the ILR scales). However, operationalizing Novice-Low and obtaining enough sample data from the Distinguished level were considered unrealistic; hence, this study deals with the levels from Novice-Mid to Superior (accordingly, henceforth in this paper, the Reading Guidelines refer to all levels except Novice-Low and Distinguished Levels). If the level specifications are consistent with learners' developmental performance of reading proficiency, the difficulty continuum of participants' responses should show a smooth progression without too much noise among the items within and between the levels from Novice-Mid to Superior.

III. Methods

1. Participants

1,964 students in Korea participated in the current study: 904 of them for the pilot study and the remaining for the main study. Their school standings varied from the second year in high school to seniors in college. Their formal learning experience of English all began in their junior high school, if not sooner. Hence, the participants must have spent a minimum of four years studying English and a maximum of more than ten years.

2. Test development

Successful completion of the current study lies on the valid construction of a research instrument. *Valid* here denotes that all passages and accompanying test items must reflect each level specification of the ACTFL Reading Guidelines. Three stages were implemented to develop a level-specific SL reading test with the following conditions in mind: a. text selection, b. test item construction, and c. rating.

a. Text selection: Following careful identification processes for the text characteristics of each level, candidate texts were selected from various sources. All of them were in use in the U.S., and were authentic and non-fictional. Most of the texts were used as they were, so as not to breach their authenticity. Out of 70 texts collected initially, 33 were chosen for the first draft of the test. In order to meet the assumptions in the Guidelines, Intermediate High (IH) and Advanced (A) shared the common reading texts, so did Advanced-High (AH) and Superior (S). Also, in order to heighten the precision of the test measure, more texts were allocated (therefore, items, too) around the mid levels – i.e., from Novice-High (NH) to Advanced.

b. Test item construction: Test items were constructed along with the text. They were to reflect the skill (i.e., functions) descriptions of the Reading Guidelines. The test was constructed with only multiple-choice items for the analyses using a three parameter IRT model. Following careful examination of each skill specification and considering its length, one to four items were generated along with each text. That procedure generated 91 initial items in total. Unlike the text selection, no items had to share the same levels to ensure that examinee's differential performance was due solely to skill difference.

c. Rating: The third stage was to ensure objectively that the selected reading texts and accompanying items were precisely reflective of each level specification of the Reading Guidelines. This rating procedure involved two raters, who were native speakers of English and whose specialty was in language testing, over two sessions. The first session was a familiarization and internalization of the Guidelines, together with a rough screening process for the texts and items. In the second session, they were asked to apply rigorous judgment in a dichotic manner about whether or not each item realized the functions specified and whether each text reflected well enough the proposed text types in the Guidelines. 21 passages with 59 accompanying items survived this first rating procedure. All of the items marked unacceptable by raters were disregarded. Three texts, one due to a mismatch with the proposed level and two due to authenticity violation were also unacceptable. Those texts together with their accompanying test items were also deleted. In addition, nine more texts were abandoned that appeared inappropriate in content as stimulus texts.

The second rating session took place a week after the first session. In the second session, the raters were asked to rate the text, items, and authenticity and legibility on a four-point Likert scale. For the second judgment task, the inter-rater reliability coefficient using modified Spearman-Brown Prophecy formula was calculated for the three ratings and the result is reported in Table 1.

Table 1 Correlation between Raters and Inter-rater Reliability Coefficient

	Correlation between raters	Reliability of the two ratings taken together
Text	1.00	1.00
Test Items	0.64	0.78
Authenticity & Legibility	0.88	0.93

The correlation and reliability were not calculated for passage, as both of the raters marked 4 (strongly agree) for all of the passages. They did not reach agreement for three test items and one passage of authenticity. The raters were highly consistent with their judgment of authenticity and legibility, and did so likewise with the test items.

Table 2 reports the total number of texts and items for each level that were subjected to the pilot test. Numbers within the parentheses indicate numbers of texts dedicated only for the particular level. Initially, the test was to include only common texts shared by IH and A and by AH and S. However, after completing the rating procedures, it was realized that it would be more informative for the study to designate one or two texts only dedicated for such levels as IH, A, AH, and S because that procedure will enable the observation of differential effects of text hierarchy as well as that of skill with the texts fixed. Therefore, although IH and A had four and three texts respectively, they shared only two texts. Consequently, two of the four IH texts and one of the three A texts were specifically designated for those particular levels. The same procedure was applied for AH and S.

Table 2 No. of Texts and Items per Level of the Pilot Version of the Instrument

Level	No. of Texts	No. of Items
NL	2	2
NM	2	3
NH	3	7
IL	3	6
IM	3	6
IH	4 (2 for IH only)	10
A	3 (1 for A only)	6
AH	4 (2 for AH only)	7
S	3 (1 for S only)	12
Total	27 (23 – total # of the texts not shared)	59

3. The pilot test

In the pilot study, IRT analysis was performed to identify misfitting items from the pilot version (Henning, 1984; Perkins & Miller, 1984). Misfitting items were checked for the sources of such misfit, and were given further consideration as to whether or not to disregard or revise and include them. All of the responses were entered into spreadsheets using EXCEL and coded for scoring and data analyses. They were tallied as ones or zeros so they could be used for analyses using SPSS and PC-BILOG 3.11 (Mislevy & Bock, 1990). Three texts and seven items were further eliminated from the pilot version of the test since it was considered that responding to 59 items with 23 texts may be too demanding a task for the test takers. That procedure left 52 items and 20 texts with the final version, as shown in Table 3.

Table 3 Number of texts & items per level of the final version of the test

	No. of Texts	No. of Items	Deleted text	Deleted item
NL	0	0	P17, P20	I43, I53
NM	2	3		
NH	3	7		
IL	3	6		
IM	3	6		
IH	4 (2 for IH only)	10		
A	3 (1 for A only)	6		
AH	3 (2 for AH only)	6	P12	I 27
S	2 (1 for S only)	8	P12	I 7, I 28, I 29, I 30
Total	27 (20 texts not shared)	52	4	7

IV. Analyses

1. Preliminary analyses

Table 4 reports the descriptive statistics for the scores on the test.

Table 4 Descriptive statistics for examinees' performance on the test

N	*k*	Mean	Mode	Median	*SD*	Hi	Lo	Range
1060	52	34.39	35	35	7.77	51	9	43

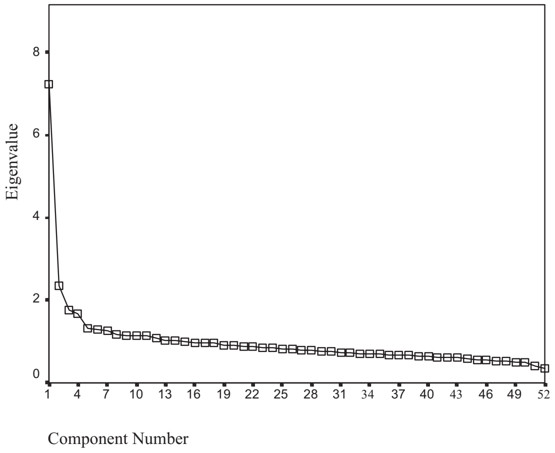
Even though the highest possible score was 52, 51 was observed as being the actual highest score, while 9 was the lowest. Also, the mean score of 34 and the mode and the median of 35, imply that the test was negatively skewed with an *SD* of 7.77. The reliability coefficients using Cronbach alpha resulted in .86 with a standardized item alpha of .87. Considering the improvement of the Cronbach alpha results of .81 from the pilot version of the test to .87 of the final version, the quality of the test was improved through the piloting process. In addition, based on its high reliability coefficients the test appeared to be a reliable measure for the research purpose.

2. IRT analyses

First, two preliminary IRT assumptions were checked: unidimensionality and local independence. For the unidimensionality assumption, a factor analysis was performed to look for a dominant component as demonstrated in Figure 1 (Scree plot). Local independence among items was double checked qualitatively by the item raters.

As the scree plot in Figure 1 demonstrates, there is a noticeable break between the first and the second components resulting in the eigenvalue difference of approximately five. In reality, a test can hardly be psychologically unidimensional. Hambleton and Swaminathan (1985), as an alternative investigation of unidimensionality, suggest finding a dominant factor that influences test performance. The factor analysis procedure in this study also accomplished this goal by confirming the existence of a dominant factor in terms of the eigenvalue. Also, to the argument against using a unidimensional statistical model to validate possibly multidimensional constructs such as reading proficiency, Henning (1987) successfully demonstrated that psychometric dimensionality does not have to be compatible with psychological dimensionality.

Figure 1 A scree plot of components by Eigenvalues



PC-BILOG 3 was used to calibrate item parameters. The information that BILOG generated was examined based on each phase especially for the item fits and model fit. The EM cycles stopped after 14 cycles with the largest change at 0.00707. Table 5 provides information about item parameters and chi-square item fit statistics from Phase 2 of BILOG.

Table 5 Item parameter and chi-square item fit statistics

Item #	*a*	*b*	*c*	Chi-square (Prob.)	Item #	*a*	*b*	*c*	Chi-square (Prob.)
1*	0.602	0.365	0.188	18.0 (0.0352)	27*	0.670	-2.274	0.204	20.0 (0.0029)
2	0.519	-4.501	0.251	6.9	28	1.093	1.889	0.323	13.5

				(0.1417)					(0.1392)
3	0.747	-1.702	0.236	8.1 (0.2275)	29	0.741	1.037	0.259	15.7 (0.0729)
4	0.616	-2.483	0.251	10.3 (0.1132)	30	1.660	1.579	0.242	14.6 (0.1018)
5	0.612	-1.341	0.198	7.3 (0.3959)	31	2.445	1.458	0.381	7.1 (0.5274)
6	0.756	-1.482	0.223	11.0 (0.1356)	32*	2.469	1.323	0.307	23.7 (0.0027)
7	1.029	1.633	0331	10.8 (0.2881)	33	1.870	2.178	0.208	13.9 (0.1235)
8	0.930	1.878	0.352	10.7 (0.2957)	34	1.482	1.656	0.366	11.3 (0.2519)
9	1.150	1.277	0.310	5.7 (0.7700)	35	0.649	-1.355	0.250	7.9 (0.3409)
10	0.913	-0.733	0.235	6.5 (0.4789)	36	0.925	-1.599	0.195	3.6 (0.7346)
11	0.926	-0.616	0.242	5.2 (0.6376)	37	0.763	-2.385	0.199	8.5 (0.1283)
12	0.967	-0.892	0.245	7.2 (0.4069)	38	1.239	0.816	0.176	14.3 (0.0746)
13	0.470	-0.170	0.269	13.1 (0.1557)	39*	0.935	0.741	0.170	26.9 (0.0016)
14	0.498	-0.693	0.287	10.9 (0.2064)	40*	0.776	0.417	0.110	69.2 (0.0000)
15	1.040	-1.442	0.230	7.7 (0.1699)	41*	0.452	1.467	0.201	17.8 (0.0373)
16	0.706	-1.762	0.221	7.9 (0.2447)	42	0.649	0.705	0.117	10.5 (0.3117)
17	1.351	0.199	0.372	8.6 (0.2791)	43	0.890	-0.236	0.352	6.4 (0.4925)
18	0.697	-0.123	0.194	11.5 (0.1722)	44	1.028	0.333	0.262	15.1 (0.0560)
19*	0.795	-0.748	0.174	22.6 (0.0021)	45	1.149	0.318	0.255	10.9 (0.2064)
20*	0.851	0.428	0.138	24.6 (0.0036)	46	0.833	0.760	0.184	15.39 (0.0688)
21	0.844	-0.682	0.248	5.6 (0.5831)	47	1.099	-0.060	0.247	11.7 (0.1084)
22*	1.162	0.061	0.315	17.0 (0.0175)	48	0.935	-0.834	0.233	11.0 (0.1357)
23	1.399	0.052	0.164	11.3 (0.1244)	49	1.303	-0.114	0.365	5.0 (0.5450)
24	1.091	-2.377	0.201	3.0 (0.3897)	50	1.567	-0.234	0.388	3.0 (0.7070)
25	0.990	-2.696	0.200	4.6 (0.975)	51	0.917	-2.003	0.234	1.5 (0.8294)

26	0.596	-2.223	0.213	10.0 (0.1896)	52	0.893	-2.207	0.232	1.6 (0.8164)

a: Slope; *b:* Threshold; *c*: Asymptote
* misfitting items with probability lower than 0.05

Nineteen items in total demonstrate high discrimination power with slopes over 1.000. Among them, Items 31 and 32 are particularly discriminating in ability dispersion with slopes of over 2.000. As for the threshold (difficulty), 29 items recorded threshold values lower than 0.00 – i.e., with minus values, implying that they were relatively easy for the examinees. In particular, Item 2 turned out to be the easiest with a threshold value of – 4.501. As shown in the fourth column of Table 5, the asymptotes of each item ranged from 0.110 to 0.388. Since four options were provided for each item in the test, the probability of getting an item by blind guessing would be 25%, i.e., 0.250. However, the actual guessing values found in the results varied depending on items. That is, examinees seemed to apply different degrees of guessing in their responses in search for the correct answer. Such a finding justifies the use of the three-parameter IRT model for this study.

Considering the chi-square fit statistics of each item, Items 1, 19, 20, 22, 27, 32, 39, 40, and 41 – nine items in total – appeared to be misfitting with probabilities lower than 0.05. The source of the misfitting was examined further with their item characteristic curves (ICCs). As far as their ICCs are concerned, the fits of Items 1, 22, 27, and probably 41 did not appear critical, despite their low probabilities. Most of the residuals in these items were within the bounds of the significance level for the goodness-of-fit of the item-response function. Misfits may still have occurred with certain ability groups as in the rest of five items (e.g., Appendix for the ICC of Item 40).

V. Results and discussion

1. IRT results

In order to examine if and to what degree the items are referenced to the levels that they were supposed to represent, they were sorted by the difficulty parameter and placed on the latent difficulty continuum. Table 6 provides a side-by-side comparison of items on such a continuum and their proposed levels from the Reading Guidelines.

Table 6 Difficulty order of item with proposed levels

Item #	Threshold (difficulty)	Proposed Level	Item #	Threshold (difficulty)	Proposed Level
I2	-4.501	NM	I18	-0.123	A
I25	-2.696	NH	I49	-0.114	IH
I4	-2.483	IL	I47	-0.060	IL
I37	-2.385	NH	I29	1.037	AH
I24	-2.377	NH	I23	0.052	IH
I27	-2.274	NH	I22	0.061	IH
I26	-2.223	NH	I17	0.199	IH
I52	-2.207	NM	I45	0.318	A
I51	-2.003	NM	I44	0.333	IH
I36	-1.599	NH	I1	0.365	A
I16	-1.762	IH	I40	0.417	S
I3	-1.702	IL	I20	0.428	A
I6	-1.482	IM	I42	0.705	IH
I15	-1.442	IM	I39	0.741	AH
I35	-1.355	NH	I46	0.760	A
I5	-1.341	IM	I38	0.816	AH
I12	-0.892	IL	I9	1.277	S
I48	-0.834	IL	I32	1.323	S
I19	-0.748	A	I31	1.458	S
I10	-0.733	IM	I41	1.467	S
I14	-0.693	IM	I30	1.579	S
I21	-0.682	IH	I7	1.633	S
I11	-0.616	IM	I34	1.656	AH
I43	-0.236	IH	I8	1.878	S
I50	-0.234	IH	I28	1.889	AH
I13	-0.170	IL	I33	2.178	AH

Also, Figures 2a, 2b, and 3 (overleaf) visually present the items on the continuum: Figures 2a and 2b based on the “ideal” item difficulties and Figure 3 based on the actual difficulty parameters from this study. Since achieving the same difficulties with the items within a single level is considered impossible and some amount of difficulty variances is expected, what one may expect to observe would be Figure 2b – i.e., a progressive development of item difficulties even within a single level and its smooth connection with the previous as well as the next levels for the item difficulty progression in the right direction, low to high levels. Even a brief examination of Figures 2a and 2b and their comparison with Figure 3 provide information about the overall progression of the item difficulties across levels and about the difficulty consistency among items within individual levels. It appears that three levels – IL, IH, and AH are especially problematic, although a general progression across levels has been achieved.

Figure 2 Visual presentations of "ideal" item difficulty by level

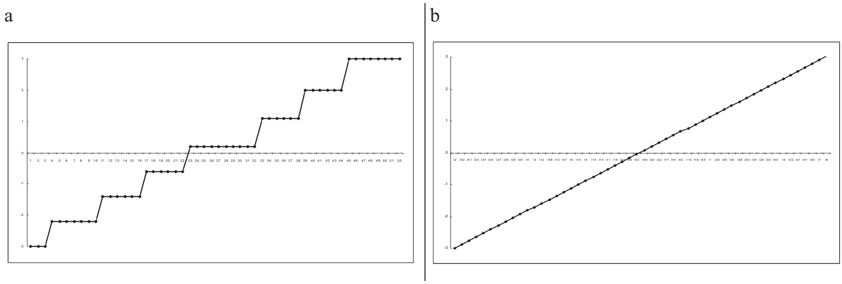
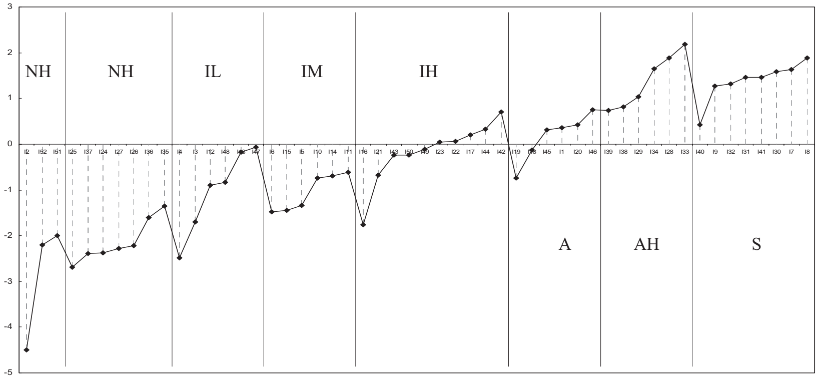


Figure 3 Visual presentation of "actual" item difficulty by level



2. ANOVA results

For more generic analyses of the level distinctions between the four levels – Novice (N), Intermediate (I), Advanced (A), and Superior (S) – an ANOVA was performed, and the results are reported in Table 7. In addition, in order to locate the significant differences among groups, multiple mean comparisons using Scheffé were calculated, and the results are reported in Table 8.

Table 7 ANOVA results for mean differences among N, I, A, and S

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	1985294.087	3	661764.696	58.012*	.000
Within Groups	501923.393	3177	11407.350		
Total	2487217.479	3180			

Table 8 A multiple comparison among levels

(I) Level	(J) Level	Mean Difference (I-J)	*S.E.*	*Sig.*
Novice	Intermediate	173.65*	41.727	.002
	Advanced	479.28*	45.731	.000
	Superior	551.41*	52.634	.000
Intermediate	Novice	-173.65*	41.727	.002
	Advanced	305.64*	39.383	.000
	Superior	377.77*	47.223	.000
Advanced	Novice	-479.28*	45.731	.000
	Intermediate	-305.64*	39.383	.000
	Superior	72.13	50.796	.574
Superior	Novice	-551.41*	52.634	.000
	Intermediate	-377.77*	47.223	.000
	Advanced	-72.13	50.796	.574

* significant at the .05 level.

Table 7 confirms that the mean difference between levels of N, I, A, and S was significant at *F* (3, 3177) = 58.012, *p* < .05. Also, as Table 8 suggests, significant mean differences resulted mainly between each combination of the three levels, N, I, and A. S was significantly different from N and I; yet, no difference was observed between A and S. Based on the linear development of each level from N to S, there were differences between N and I and I and A except between A and S. Dandonoli and Henning (1990) found that the middle levels (i.e., Intermediate and Advanced) were not sensitive enough for the distinction between them; however, the current study revealed that it was the distinction between the Advanced and Superior levels which was not clear. Such an unclear distinction between A and S appears to be due especially to the AH items of I29, I34, and I28. As noticeable in Figure 3, they were rising up steeply going over the difficulty levels of most of the S items. They were more difficult than they were supposed to be.

3. Item examinations by level

a. Novice level (mid and high)

For the Novice Level, ten items and five texts were used in order to investigate the skill and text characteristics of the Guidelines. Items for Novice Level demonstrate relatively good consistency in their positions in a sequence except for Item 2 (for NM) and Item 35 (for NH). Item 2 turned out to be too easy, whereas Item 35 was found more difficult than it should have been. Item 2 was for simple fact-finding based on a CD ad. An explicit graphic feature of the stimulus of Item 2 may have enhanced examinee's comprehension of the question. Item 35 required a simple interpretation skill based on a standardized message: questioning the amount of money to be paid based on a course registration form. It was assumed that this item would be less cognitively demanding, since a simple word matching skill between the stem of the questions and the text was expected good enough to respond to the item. However, this item was more difficult than even several Intermediate items. In addition, Items 36 and 37 were for the same skill to be applied. Yet, Item 35 was found much more difficult than either of the other two. The unusual pattern of Item 35 may be due to the examinees' lack of cultural familiarity with the text. The registration forms used as the stimuli may not be easily accessible by English learners in Korea – i.e., contextualized in the target culture. The unexpectedly high difficulty of Item 35 may have been due to its culture-specific demand rather than the cognitive skill demand.

The two NM items – Items 51 and 52 – also were found to be relatively more difficult than the other NH items. The texts for Item 51 and 52 were four types of business cards, which usually aim to provide explicit information about the card givers. Items 51 and 52 were unusually difficult, especially compared to Item 2, which turned out to be the easiest. The target words for the two Items were "bank" and "email", with which most examinees must already be familiar. However, considering that the examinees for the study were students and they usually do not carry business cards, they may have found the cards unfamiliar, especially when they were written in English. Again, although texts are simple and standardized, reading may not happen effectively if the contents of the texts are heavily target culture-specific (or genre-specific in this case).

b. Intermediate level (low, mid, and high)

Overall, the items for the Intermediate Level except for the ones for IL appear consistent with regard to their difficulty positions. Item 4 (for IL) was placed in a fairly low position on the difficulty continuum, possibly due to its low skill demand: information extracting. However, Item 3, which was based on the same text and of the main idea, was placed in a relatively difficult position. Such difference in difficulty of Items 3 and 4 may suggest a reevaluation of the assumptions regarding cognitive difficulties with the two skill areas of the Guidelines: skimming and scanning. The two were posited as the representative skills of the Intermediate Levels. However, such supposition about the degree of cognitive difficulty in processing the two skills may not necessarily be true in actual reading performance.

Another example is Items 16 and 47, both to find the main idea. Their positions in the difficulty continuum are far apart, producing a large difference in their difficulty parameter values. Interestingly, although Item 16 targeted a higher proficiency level (IH) than Item 47 (IL), Item 47 was more difficult than Item 16. Such a finding may be due to their interaction effect between text and skill – i.e., the interaction may reverse the predicted difficulty order with the actual difficulties of items.

Skimming and scanning may be more dependent on the text characteristics, demanding both linguistic and extra-linguistic knowledge for any text by the reader. In particular, Item 28 was of interest, since it also required the skimming skill to respond correctly. Interestingly, the item was placed in the second most difficult position on the overall difficulty continuum. Skimming and scanning are skills that often need to be activated in many reading tasks. Placing these skills on the lower end of the cognitive difficulty continuum may have been the source of unpredictability for the difficulties of items.

c. Advanced level (advanced and advanced high)

The most problematic pattern with regard to the item order by difficulty was found in the Advanced Level. Items for the Advanced Level were scattered over a large difficulty range. Some items for AH were found to be more difficult than the items for the S. Especially, Items 33 and 34 for AH were positioned the highest on the difficulty continuum. Those items were based on a newspaper editorial. Neither the syntactic structure nor the lexical characteristics of the editorial text appeared demanding. However, newspaper editorials require much schematic background knowledge about the target culture and current social events. The difficulty of such texts may have been due to such knowledge requirements of the target culture or society for examinees to interact efficiently with the texts. The level descriptor of AH describes the ability for that level as "Able to understand … texts which involve aspects of target-language culture." In addition, Child's (1987) model assumes that as an FL reader becomes more proficient, the texts that the reader can interact with become more individual- or culture-specific. The high difficulty of Items 34 and 35 illustrates that some of the proficiency suppositions concerning A or AH Levels may need to be reevaluated as opposed to such suppositions and also in relation to those of the Superior Level.

d. Superior level

Although within the Superior Level, items seem equivalent with their difficulty, Table 6 and Figure 3 reveals that distinction between A and S Levels is not clear. Item 40 was especially too easy for S Level. Item 40 was based on a text that was lexically as well as syntactically demanding. The required skill was inference, which is often considered possessing a cognitively higher skill demand for processing. However, considering the contents of Item 40, examinees may have simply applied the common word matching strategy in responding to the item, since the wording of the stem was almost the same as the cue sentences in the text and the wording of the options was relatively short.

e. Shared texts

As another area of research interest, some of the levels shared the same texts for certain items. This research procedure was aimed mainly at observing the order effect on the reading performance only by skill. The Advanced Level shared some of its texts with the Intermediate, and the Superior Level did likewise with the Advanced High. In fact, Lee and Musumeci (1988) tested the opposite phenomena, i.e., how text affects the reading performance when the same reading skills are applied. They reported that different text characteristics differentiated the reading performance by the same skill.

In this study, ten items – Items 16, 19, 43, 18, 17, 45, 44, 20, 42, and 46 – between IH and A were constructed in order to examine how skill differentiates reading performance based on the same texts. Similarly, five items – Item 29, 32, 31, 30, and 28 – between AH and S were also constructed. Items 16, 43, 17, 44, and 42 were aimed at IH, and Items 19, 18, 45, 20, and 46 were targeting A.

Items 28 and 29 were intended for AH, and Items 30, 31, and 32 were designed for S. As the order of the items demonstrates, no consistency in order was found with items between the two pairs of levels. Such a finding indicates that skill is not a significant factor that affected the reading performance. Therefore, Lee and Musumeci's (1988) findings regarding reading proficiency differentiations by the text characteristics appear legitimate, as the results of this study also confirm.

VI. Conclusions

Throughout this study, how to achieve level-specific test construction based solely on the ACTFL Reading Guidelines was examined. Also, the procedures enabled observation of how the two major facets of the Guidelines – i.e., text and skill, in addition to their interaction, can be operationalized in performing on the test items by SL learners. Using the item-level information generated by IRT, a thorough investigation was performed of reading proficiency as a progressive and linear developmental process in SL learning.

For Research question 1, Cronbach's alpha was calculated in order to examine the reliability coefficients of the two instruments for pilot and main study purposes, and the results were .81 and .87, respectively. In terms of their internal consistency, both instruments were found to be relatively reliable measures. Furthermore, it was noted that through the piloting procedures, the reliability of the measurement was actually improved. Therefore, it was found that it is feasible to construct a reliable test for measuring reading proficiency at least statistically.

For Research question 2, IRT procedure revealed a general progression in the right direction of SL reading proficiency on the difficulty continuum of items (Figure 3). However, for the following two reasons, the entire characterizations of the ACTFL Reading Guidelines do not seem to be legitimate in their presentation of SL reading proficiency as linear or progressive.

1) Frequent insensitivity was revealed by relatively large number of items within and across levels presented in the Guidelines.
2) As revealed by IRT (Figure 3) as well as by the ANOVA analyses, the distinction between the Advanced and Superior levels does not seem clear. If this is to be the actual case with the Guidelines, the addition of Distinguished Level becomes especially questionable.

In summary of the research findings of the current study in relation to the research questions:

1) Constructing a reliable measure for SL reading proficiency appears feasible, although, a significant amount of effort has to be put forward for the development process (Research question 1).
2) From Figure 3, a general progression of SL reading proficiency was observed; yet, frequent insensitivity was revealed among items within and between adjacent levels (Research question 2).

3) The level specific characterization based on the skill and text hierarchy of the ACTFL Reading Guidelines appears to be generally acceptable; yet still not entirely valid due to inconsistency demonstrated by relatively large number of the items (Research question 2).

As has been argued by previous researchers, the ACTFL Reading Guidelines need further empirical research in order to support the validity of the Guidelines. Particularly, the sensitivity in describing reading proficiency for some of the levels such as Intermediate Low, Intermediate Hi, Advanced High and Superior needs to be reexamined. Therefore, when one intends to develop a level-specific test based on the ACTFL Reading Guidelines, doing so together with the research findings of prior studies as well as that of the current one will suggest better guidance for approximating SL reading proficiency development in the test development process.

VII. References

Bachman, L.F. 1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F. and **Savignon, S. J.** 1986: The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *The Modern Language Journal* 70, 380-390.

Brown, J.D. 1996: *Testing in language programs*. Englewood Cliffs, NJ: Prentice Hall Regents.

Canale, M. 1988: The measurement of communicative competence. *Annual Review of Applied Linguistics* 8, 67-84.

Chapelle, C. and **Douglas, D.** 1993: Foundations and directions for a new decade of language testing. In Douglas D. and Chapelle C., editors, *A new decade of language testing research*. Washington, DC: TESOL.

Child, J.R. 1987: Language proficiency levels and the typology of texts. In Byrnes H. and Canale M., editors, *Defining and developing proficiency: Guidelines, implementations, and concept*. Lincolnwood, IL: National Textbook Co, 97–106.

Clark, J.L. 1985: Curriculum renewal in second language learning: An overview. *Canadian Modern Language Review* 42, 342-360.

Dandonoli, P. and **Henning, G.** 1990: An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals* 23, 11-22.

Hambleton, R.K. and **Swaminathan, H.** 1985: *Item response theory: Principles and application.* Boston: Kluwer.

Henning, G. 1984: Advantages of latent trait measurement in language testing. *Language Testing* 1, 123-133.

Henning, G. 1987: *A guide to language testing: development, evaluation, research.* Cambridge, MA: Newbury House.

Hudson, T. 1991: A content comprehension approach to reading English for science and technology. *TESOL Quarterly* 25, 77-104.

Hudson, T. 1998: Theoretical perspective on reading. *Annual Review of Applied Linguistics* 18, 43-60.

Ingram, D.E. 1985: Assessing proficiency: An overview on some aspects of testing. In Hyltenstam K. and Pienemann M., editors, *Modelling and assessing second language acquisition*. Great Britain: Multilingual Matters Ltd.

Kaya-Carton, E. and **Carton, A.S.** 1986: Multidimensionality of foreign language reading proficiency: Preliminary considerations in assessment. *Foreign Language Annals* 18, 95-102.

Lange, D.L. and **Lowe, P.** 1987: Grading reading passages according to the ACTFL/ILR reading proficiency standard: Can it be learned? *Selected Papers from the 1986 Language Testing Research Colloquium*. Monterey, CA: Defense Language Institute.

Lantolf, J.P. and **Frawley, W.** 1985: Oral proficiency testing: A critical analysis. *Modern Language Journal* 69, 337-345.

Lantolf, J.P. and **Frawley, W.** 1988: Proficiency: Understanding the construct. *Studies in Second Language Acquisition* 10, 181-195.

Lee, J.F. and **Musumeci, D.** 1988: On hierarchies of reading skills and text types. *The Modern Language Journal* 72, 173-187.

Long, M. 1985: A role for instruction in second language acquisition: task-based language training. In Hyltenstam K. and Pienemann M., editors, *Modelling and assessing second language acquisition*. Great Britain: Multilingual Matters Ltd.

Mecartty, F.H. 1998: The effects of proficiency level and passage content on reading skills assessment. *Foreign language Annals* 31, 517-534.

Mislevy, R.J. 1995: Test theory and language-learning assessment. *Language Testing* 12, 341-369.

Mislevy, R.J. and **Bock, R.D.** 1990: *PC-BILOG 3: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software, Inc.

North, B. 2000: *The development of a common framework scale of language proficiency*. New York: Peter Lang Publishing, Inc.

North, B. and **Schneider, G.** 1998: Scaling descriptors for language proficiency scales. *Language Testing* 15, 217-263.

Nunnally, J.C. and **Bernstein, I.H.** 1994: *Psychometric theory*. NY: McGraw-Hill, Inc.

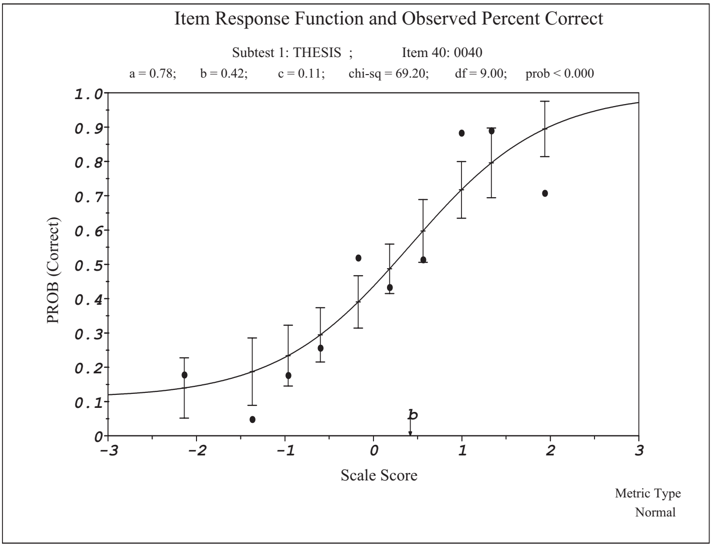
Park, S. 2004: Validation of the text and skill hierarchy of the ACTFL Reading Guidelines. *English Education* 59, 143-164.

Perkins, K. and **Miller, L.D.** 1984: Comparative analyses of ESL second language reading comprehension data: Classical test theory and latent trait measurement. *Language Testing* 1, 21-32.

Pienemann, M. and **Johnston, M.** 1987: Factors affecting the development of language proficiency. In Nunan D., editor, *Applying second language acquisition research*. Adelaide: National Curriculum Resource Centre, 45-141.

Spolsky, B. 1986: A multiple choice for language testers. *Language Testing* 3, 147-58.

Appendix **The item characteristic curve (ICC) for Item 40**


Item Response Function and Observed Percent Correct
Subtest 1: THESIS ; Item 40: 0040
a = 0.78; b = 0.42; c = 0.11; chi-sq = 69.20; df = 9.00; prob < 0.000
PROB (Correct)
b
Scale Score
Metric Type
Normal