Investigating the validity of KFL text difficulty as defined by the ILR Reading Scales

Sun-Young Shin
Indiana University

Abstract

The purpose of this study is to explore the validity of text difficulty as defined by the Interagency Language Roundtable (ILR) in the context of the Korean Flagship Admission Test (KFAT) administered at a university on the west coast in the U.S. To do so, this study used four external concurrent data sources: a survey and interviews with five college-level Korean teachers and 39 KFAT test takers; an analysis of item difficulty; and an analysis of linguistic features in reading texts. The correlations of these external measures with ILR text difficulty rankings indicate that text difficulty as described by the ILR is not always consistent with experts' and test takers' perceptions of text difficulty, test taker performance, and common linguistic indices of text difficulty. These findings raise doubts about the appropriateness of using the generic ILR reading descriptors to determine text difficulty in the development of the KFAT reading test.

Introduction

Flagship programs are the result of a partnership between the United States (U.S.) government and higher education institutions, which prepare university students for the global professional market place by providing them with the opportunity to achieve high levels of proficiency in less commonly taught languages such as Russian, Arabic, Chinese, and Korean (Spring, 2012). In that vein, the goal of the Korean Flagship program was to bring Korean learners from middle levels of proficiency to high levels of proficiency over the course of their studies. Flagship programs measure student proficiency using a common scale developed by the Interagency Language Roundtable (ILR). Students in the Korean Flagship program must enter the program with an ILR proficiency level of 2 or 2+ and are expected to achieve Korean proficiency at ILR level 3 or 3+ by the end of the program. Applicants' proficiency is tested in reading, listening, and writing using the Korean Flagship Admission Test (KFAT). The present study focuses on the reading section of this test, though the KFAT test battery also includes writing, listening, and grammar subsections.

A critical step in designing any reading test is for developers to establish text difficulty, especially when they must choose texts at appropriate levels for target groups and create comparable sets of reading passages for different forms of the test. In the context of KFAT reading test, applicants' reading proficiency is determined by how they perform on passages reflecting each level of the ILR. Thus it is important that each text accurately reflects its ILR level.

The ILR descriptors for reading, writing, speaking, and listening at levels 0 through 5, including the "plus" values, were originally established in the early 1980s by the Testing Committee of the ILR, with representation from all government agencies concerned. These descriptions were ratified and disseminated as the official standards for documenting language proficiency within the U.S. government (Leaver & Shekhtman, 2002) and have been regularly updated ever since. The concept of reading proficiency in the ILR, as shown in the method section in this paper, is defined in terms of text type, reading skill, and task-based performance (Galloway, 1986). In other words, a specific developmental level is associated with text and reading skill variables in the ILR. Such standards are often viewed as "correct, accurate, exact descriptions of development levels of second-language reading proficiency" (Lee and Musumeci, 1988, p.173). Yet, few empirical studies have been conducted to validate those descriptions. Furthermore, although there are a number of studies related to validating a hierarchy of reading skills (e.g., Alderson, 1990a, 1990b; Lumley, 1993), very little research has been conducted on validating text difficulty in Korean based on the ILR or the ACTFL (American Council on the Teaching of Foreign Languages) Proficiency Guidelines (American Council on the Teaching of Foreign Languages, 2012).

The assumption made in the ILR reading scales is that generic level descriptors can be used to determine text difficulty in all foreign languages. However, some evidence suggests that this assumption needs to be further tested. The ACTFL Proficiency Guidelines (ACTFL, 2012), which are based on the ILR scale, have been criticized as being a closed system (Lantolf & Frawley, 1985), which may suggest that text difficulty has been predetermined by a limited number of factors without considering actual learners' performance on given texts. Similarly, Lee and Musumeci (1988) found that actual test taker performance on different text types was not in line with the hierarchy of text types based on the ACTFL construct of reading proficiency. The need to find evidence for the hierarchical construct of text as defined by the ILR reading scales is a fundamental validity issue, and must come from outside the ILR scales or tests based on them (Kenyon, 1998). If the ILR proficiency guidelines are to be supported as an official norm for documenting language proficiency, then they should be validated by sources external to them.

Therefore, the present study aims to validate the text hierarchy constructed according to the ILR Reading Scales through four concurrent external sources of information. First, the ILR descriptors of text difficulty are compared with impressions of text difficulty held by Korean as a foreign language (KFL) teachers and

KFAT test-takers who have no experience using ILR descriptors. In addition, test takers' actual performance is compared with text difficulty based on the ILR. Finally, text difficulty is validated by comparing descriptions of text difficulty from the ILR scales with an external framework of text difficulty developed by Chapelle, Jamieson, and Hegelheimer (2003). In their study, they examined if specific texts were appropriately selected for a given level by professional judges. The researcher compared the judges' designated levels of text difficulty to lexico-grammatical and semantic features of a text, which includes type-token ratios, word and sentence length, and sentence complexity. Some of these features were adapted in this study for Korean.

This study thus seeks to answer the following questions: 1) Do KFL experts and test takers perceive text difficulty in the same manner as described in the ILR? 2) What criteria do KFL experts and test takers use to rate text difficulty? Do these criteria reflect the ILR level descriptors? 3) Is there a relationship between text difficulty as judged by the ILR and test taker performance? 4) Are ILR and Chapelle, et al.'s (2003) ratings of text difficulty consistent?

Method

KFAT Reading Test

The reading section of the KFAT contains ten passages and 21 questions consisting of 13 multiple-choice (MC) items with four options and eight open-ended questions. The test takers have a total of 50 minutes to complete the test. Each question is designed to tap into the characteristics of ILR levels 2, 2+, 3 or 3+ as stated in the Skill Level Descriptions (Interagency Language Roundtable, n.d.). The main characteristics are: 1) understanding and locating the main ideas and details in each passage; and 2) recognizing vocabulary in the passages. All 10 passages in the reading test come from authentic newspaper, magazines, and academic textbooks in Korean and are meant to represent ILR levels 2 through 3+. Note that the texts are not presented in a sequence reflecting the hierarchy of text types because of a potential bias in presentation.

The actual difficulty of a test task relies on both text difficulty and task difficulty (Chapelle, Jamieson, and Hegelheimer, 2003). Thus, in order to examine the effects of text difficulty, not item difficulty, the present study holds item type constant, only including multiple choice items related to understanding and locating main ideas. These ten scores were compared with the levels of text as described in the ILR.

The ILR Scales

The ILR Scales posit levels of proficiency from 0 to 5 including additional sublevels (e.g., 2+), In the KFAT, two test developers determine text difficulty based only on the ILR criteria not on their intuitions about text difficulty. Since this study focuses on the ILR text difficulty level from 2 to 3+, the basic features of texts associated with levels 2 to 3+ described in the ILR are summarized as follows (Interagency Language Roundtable, n.d.):

> ***Reading(R)-2 Limited Working Proficiency***
>
> The Level R-2 level passage is characterized by simple, authentic written material in a form equivalent to usual printing or typescript on subjects within a familiar context. Texts may include descriptions and narrations with a clear underlying structure in context such as news items describing frequently occurring events, simple biological information, social notices, formulaic business letters, and simple technical material written for the general reader.
>
> **R-2+ Limited Working Proficiency, Plus**
>
> The Level R-2+ passage is associated with factual material in non-technical prose as well as some discussions of concrete topics related to special professional interests.
>
> **R-3 General Professional Proficiency**
>
> The Level R-3 passage includes news reports or news items in major periodicals, routine correspondence, general reports and technical material in his or her professional field;

all of these may include hypothesis, argumentation, and supported opinions.

R-3+ General Professional Proficiency, Plus

The Level R-3+ passage is typically related to contemporary expository, technical, or literary texts not containing slang and uncommon idioms.

Each of the ten reading passages is assigned to an ILR level between 2 and 3+ according to the topic and text type, as shown in the Table 1 below. There are four R-2, two R-2+, two R-3, and two 3+ reading passages in the KFAT reading test.

Table 1. The topics, text type, and ILR levels of the selected reading passages

Passages	Topics	Text Type	The ILR levels
Passage 1	Origin of the word, *ul-jjang*, 'a handsome face'	Factual reports on current events	2
Passage 2	Reason why sufficient sleep is important	Factual reports on familiar topics	2
Passage 3	Danger of rapid spread of AIDS in Korea	Factual reports on current events	2
Passage 4	Benefits of walking to help lose weight	Factual reports on familiar topics	2
Passage 5	Korean college entrance exam	Editorials with evaluative comments	3
Passage 6	Relationship between carbon dioxide and food quality	Factual reports on unfamiliar topics	2+
Passage 7	Avoidance tendency in	Editorials with	3+

	science and engineering college in Korea	critical information	
Passage 8	Concerns with increased spending on overseas travel in Korea	Factual reports on unfamiliar topics	2+
Passage 9	Need for crackdown on illegal immigrants in Korea	Editorials with evaluative comments	3
Passage 10	Unemployment problems across all age groups in Korea	Editorials with critical information	3+

Participants

Participants of the study were five KFL teachers from U.S. universities and 39 KFAT test takers. Five KFL teachers had taught Korean as a foreign language at the college level in the U.S. for at least three years. KFAT test takers were all applicants for the Korean Flagship. They ranged in age from 19 to 34 years; there were 12 males and 27 females and, except for three participants, all participants were Korean heritage speakers.

Concurrent Measures

Four concurrent measures were used to obtain evidence about the validity of the text hierarchy constructed in accordance with the ILR scales. First, the self-assessment survey for KFL teachers and KFAT test takers was developed to get their opinion on the hierarchical levels of the texts. This instrument asked experienced Korean teachers from test takers' perspectives not for themselves, and test takers to rate the text difficulty of each passage on a four-point scale which is equivalent to the ILR levels (from 2 to 3+).

The second concurrent measure is an individual interview with KFL teachers and test takers to explore the specific criteria they used to rank overall text difficulty. This provided an opportunity to verify the survey data with a more qualitative measure of how these groups determined the hierarchical ordering of reading the passages.

The interviewees were asked about their standards to rank text difficulty.

The third measure applied to this study is comparing the results of item difficulty of the reading test with the text difficulty of the ILR. To empirically validate the ILR scale, simple comparisons were conducted between item difficulty and texts aligned with the ILR Reading Scales. Item difficulty was calculated for all MC test questions and analyzed alongside the ILR level of its corresponding text to determine the extent to which these correlated

The fourth measure is quantitative linguistic analysis of text difficulty based on Chapelle et al.'s (2003) framework in which lexical, syntactic, and semantic features of English texts are used to justify their level assignments. The variables in the framework are: type-token ratio, average word and sentence lengths, information density, noun-noun sequences, passives, and instances of several types of sentence clauses. However, some English grammatical constructions such as information density (noun + attributive adjective + preposition frequency), noun-noun sequences, and passives, are not applicable to Korean. Thus, two adaptations were made to the framework to reflect Korean, specifically to the variables type-token ratio and the number of embedded clauses per sentence, known as T-units, were applied to all passages in this context. Type-token ratio is a measure of the diversity of words in a text, calculated as the ratio of the number of different words in a text to the number of total words (Chapelle et al., 2003). Horning (1993) shows that redundant words make texts easier to read and there is additional evidence that high type-token ratios are associated with texts that are difficult to process because of the higher information load (Conrad, 1996). However, given that function words such as particles and conjunctions in Korean do not seem to affect the information load, only content words were counted in this study.

T-unit is the minimal unit constructing a complete sentence, consisting of one independent clause and any dependent clauses connected to it. The number of clauses per T-unit is often used as a measure of the structural complexity of a sentence. Korean is different from English in features of syntactic complexity. Thus, instead of causative adverbial subordination and relative clauses from Chapelle et al.'s (2003) framework, which do not exist in Korean grammar, a list of Korean embedded clauses from Sohn (1999) was used in this study.

Data analysis

To explore the relationship between teachers' and test takers' text difficulty ratings and the ILR hierarchical ordering, correlation coefficients were calculated between participants' survey responses and the ILR-based hierarchical ordering. KFL teachers' and test takers' interview data explaining their rating criteria were also analyzed.

Each test taker received a total of 21 scores, yet, as discussed above, in order to control for task effect, only the item difficulty scores (p-values) of the ten questions related to finding the main idea were computed and compared with ILR text ratings. Type-token ratios and T-units of each passage were calculated by hand and the Spearman rank correlation coefficient was calculated using SPSS 20 (SPSS, 2011) between these two linguistic variables and text difficulty rankings based on the ILR.

Results

The results of the analysis on KFL teachers' and test takers' impressions of text difficulty indicate that there is a strong consistency in judging text difficulty among the two groups. The correlation coefficient (estimated by Cronbach alpha) for text difficulty ratings are 0.82 for five KFL teachers and 0.86 for the 39 test takers. However, the ratings assigned by KFL teachers and test takers are quite different from the ILR text ratings. The Spearman rank correlation coefficient between KFL teachers and the ILR text ratings is as moderate as .62 but not statistically significant ($p=0.06$), indicating that there is not strong consistency between KFL teachers' text ratings and the ILR text ratings. Thus, the KFL teachers did not seem to perceive text difficulty in the same manner as the ILR hierarchical ordering of reading passages. On the other hand, the Spearman rank correlation coefficient between test takers and the ILR is as moderate as 0.71, but statistically significant (p=0.02), indicating that there seems to be some positive linear relationship between student text difficulty ratings and ILR ratings. Post-survey interviews with teachers and test takers indicated that the criteria most commonly used by KFL teachers were vocabulary and content, whereas test takers tended to use vocabulary and grammar, which are not specified in the ILR reading scale.

As a third concurrent measure of text hierarchy, item difficulty was calculated for each item. As can be seen in Table 2 below, in terms of item difficulty, the most difficult item is the one related to Text 8, and the easiest items for test takers were from Text 2. It appears that test taker performance was consistent with low ILR level texts (R-2) but not consistent with high ILR level (R-3+) texts. For instance, Text 10 was labeled 3+, but did not have high item difficulty (p=0.67). However, Text 8 was labeled 2+ but was one of the most difficult texts for test takers (p=0.28). The Spearman rank correlation coefficient between item difficulty and the ILR text difficulty rating was not statistically significant (r=0.-45, p=0.19).

Table 2. Item difficulty & ILR text difficulty for the 10 KFAT texts

Passage	1	2	3	4	5	6	7	8	9	10
ILR level	2	2	2	2	3	2+	3+	2+	3	3+
Item Difficulty (*p*)	0.59	0.87	0.72	0.79	0.56	0.67	0.67	0.28	0.51	0.67

Table 3 below illustrates how the ILR level is related to the linguistic variables identified for the adapted text difficulty framework. In terms of type-token ratio (TTR), Text 4 is the lowest (0.61), which means many words are redundant in this passage and the text 2 and 10 have the highest values of type-token ratio (0.82) indicating that these passages contain few redundant words. The Spearman rank correlation coefficient between TTRs and ILR levels variables is 0.42 (p=0.22), indicating that the relationship between two variables is not statistically significant. The t-unit, another linguistic variable known to affect text difficulty, was calculated as a measure of the structural complexity of sentences. Text 10 has the lowest value (0.83), showing that this passage contains fewer subordinate clauses and is less syntactically complex than other passages, whereas the highest value of Text 8 (1.45) indicates that this passage is more syntactically complex than any other passages in this reading test. Like TTR, the t-unit does not seem to have any positive linear relationship with ILR levels, as can be seen in the very low Spearman rank correlation coefficient (r=0.01, p=0.99). This suggests that these linguistic variables do not necessarily correlate with text difficulty defined by the ILR reading skill scale.

Table 3. Linguistic variables & ILR text difficulty

Passage	1	2	3	4	5	6	7	8	9	10
ILR level	2	2	2	2	3	2+	3+	2+	3	3+
Type-Token Ratio	.78	.82	.65	.61	.77	.65	.74	.75	.81	.82
T-unit*	1.11	1.27	1.25	.88	1.33	.77	1.44	1.45	.90	.83

* The t-unit as a measure of sentence complexity

DISCUSSION

The purpose of this study is to examine the validity of the text difficulty assumption used to develop the ILR hierarchy of text difficulty. This is particularly important for the KFAT, which relies on this assumption to determine text difficulty for the reading test. The results of the four concurrent measures in this study provide information on the construct validity of text difficulty as defined by the ILR.

It has been shown that there is little evidence from outside the ILR that supports the use of the ILR text hierarchy. First, KFL teachers do not seem to agree with the ILR reading scales regarding text difficulty. KFL teachers' and test takers' perception of text difficulty are not the same as the ILR level of text difficulty; the ILR emphasizes text type as the main factor determining text difficulty

(Lee & Musumeci, 1988), whereas content and vocabulary for KFL teachers on one hand, and vocabulary and grammar for test takers on the other hand, were determining factors of text difficulty. Interview data suggested that these impressions reflected teachers' observations and intuitions about how students learn, or what distinguishes one text from another. For instance, one of the KFL teachers arguing for the importance of content said, "Especially for Korean heritage students, if they are provided with the familiar text topic, they can get the meaning right away without resorting to information in the text." Another teacher discussing the importance of vocabulary noted, "I have seen my students always struggling with the text when given by numerous unknown words." The discrepancies between these teachers' comments and the ILR descriptors suggest that the ILR guidelines may not capture some important aspects of text difficulty. Indeed, the ILR reading scales are based on the belief that reading proficiency increases according to particular functions and text types (Allen et al, 1988), without a great focus on the components identified by the KFL teachers.

It is also important to note that test takers' level of overall reading ability seems to be related to how these variables in the text affect text difficulty. One of the low-level test takers said during the interview, "As for me, all the texts are equally difficult." On the contrary, an advanced test taker said, "All the texts seem equally easy for me that it's hard for me to rank these passages in difficulty." This evidence appears to support the Bachman's (2002) view of the interaction between the test taker and the task. The text alone itself rarely accounts for all the variance in test taker performance on a reading test. As Bachman (2002, p.466) points out, "difficulty is not a separate factor at all, but resides in the interactions among all of these components involved in assessment." Future research studies involving a larger sample would allow for a more in-depth examination of the issue of proficiency. Indeed, a larger sample would also allow for an IRT analysis of texts, which has been argued to be a thorough and appropriate tool for investigating validity (Lumley, 1993).

As mentioned in the descriptions of the ILR Reading Scales, types of text have been categorized, advancing from simple to complex - for instance, from a friendly letter to newspaper editorials on serious issues. However, text type as described in the ILR was not found to correlate significantly with actual test taker performance. This study showed test takers performing better on some texts with high ILR levels and worse on some texts with lower ILR levels. Text 10, which is labeled ILR level 3+, deals with unemployment problems facing people of all ages in Korea. Although this text is argumentative, it does not seem to be syntactically challenging to test takers; this passage has the lowest T-unit value. It appears that not all factual types of text are equally easy to understand and not all argumentative text types are equally hard to understand. With regard to more linguistic-oriented analyses, type token ratio and T-unit are not strongly associated with ILR text difficulty scores, which may be due to the fact that no clear account was made of vocabulary difficulty and structural complexity in framing text difficulty in the ILR reading scale.

Conclusion

The results of the present study suggest that text difficulty may be better evaluated using multiple indicators, instead of relying solely on the ILR text difficulty hierarchy when it applies to Korean passages. This study suggests that levels of text difficulty are not sufficiently captured by the ILR descriptions, which focus mainly on the text type. It also shows that test taker performance on each passage is not consistent with the ILR hierarchy of text difficulty. The ILR reading scale does not seem to hold all the aspects of text difficulty and might be misleading when used for text selection, with unknown and potentially negative effects when extrapolating test scores to make decisions admission decisions. Overall, this study confirms previous findings that text difficulty does not seem to reside inherently in the print but in the interaction with a reader (Alderson, 2000; Lee & Musumeci, 1988). Based on only the guidelines of the ILR, it is hard to convince a test taker why they are a level 2 or 3 Korean reader. The construct of text difficulty as defined in the ILR needs to be refined to reflect the interaction between the text and the reader, based on actual test taker performance. This study also calls for the development of a Korean-specific reading proficiency scale that will guide curriculum design and test development in KFL or KSL contexts. More detailed level descriptors for text difficulty in Korean language need to be developed based on multiple resources such as the specific constructions of Korean language, various existing proficiency guidelines and standards for foreign languages, and the framework of language use and language ability. Once scale descriptors for varying text difficulty in Korean has been established, they should then been validated by different validity evidence including teachers' and students' perceptions of given text difficulty level descriptors, and actual students' performance on reading tasks constructed based on them.

References

Alderson, J. C. (1990a). Testing reading comprehension skills (Part One). *Reading in a Foreign Language 6*, 425-438.

Alderson, J. C. (1990b). Testing reading comprehension skills (Part Two). *Reading in a Foreign Language 7*, 465-503.

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Allen, E. D., Bernhardt, E.B., Berry, M.T., & Demel, M. (1988). Comprehension and Text Genre: An analysis of secondary school foreign language readers. *The Modern Language Journal, 72*, 163-172.

American Council on the Teaching of Foreign Languages. *(ACTFL)* (2012). *ACTFL proficiency guidelines. Retrieved from* January 19, 2015 http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*, 453-476.

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20,* 409-439.

Conrad, S. (1996). Investigating academic texts with corpus-based techniques: An example from biology. *Linguistics and Education, 8*, 299-326.

Galloway, V. (1986). From defining to developing proficiency: A look at the decisions. In H. Byrnes, & M. Canale (Eds.), *Defining and Developing Proficiency: Guidelines, Implementations and Concepts* (pp.25-74). Lincolnwood, IL: National Textbook Company.

Horning, A. (1993). *The psycholinguistics of readable writing.* Norwood, NJ: Ablex

Interagency Language Roundtable. (n.d.). *Interagency Language Roundtable Language Skill Level Descriptions – Reading.* Retrieved January 19, 2015, from http://www.govtilr.org/skills/ILRscale4.htm.

Kenyon, D. (1998). An investigation of the validity of task demands on performance-based tests of oral proficiency. In A. J. Kunnan (Ed.), *Validation in language assessment: selected papers* *from the 17th Language Testing Research Colloquium, Long Beach* (pp. 19-40). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Lantolf, J. P., & Frawley, W. (1985). Oral Proficiency testing: A critical analysis. *The Modern Language Journal, 69,* 337-345.

Leaver, B. L., & Shekhtman, B. (2002). *Developing Professional-Level Language Proficiency.* Cambridge: Cambridge University Press.

Lee, J. F., & Musumeci, D. (1988). On Hierarchies of Reading Skills and Text Types. *The Modern Language Journal, 72*, 173-187.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing, 3,* 211-234.

Sohn, H. M. (1999). *The Korean Language.* Cambridge: Cambridge University Press.

Spring, M. K. (2012). Languages for specific purposes curriculum in the context of Chinese-Language Flagship programs. *The Modern Language Journal, 96*, 140-157.

SPSS Ins. (2011). *PASW Statistics for Windows, Version 20.0.* Chicago: SPSS Inc.