EFFECTS OF SUBSKILLS AND TEXT TYPES ON KOREAN EFL READING SCORES

SUNYOUNG SHIN

University of Hawai'i at Mānoa

ABSTRACT

Recent research has shown that the subskills and text types affect reader performance considerably (Lumley, 1993, Dennis, 1982). Despite the fact that most reading tests consist of a variety of subskills and text types at the same time, the studies about the relative effects of both subskills and text types on readers' performance are rare. In that vein, this article reports the relative effects of subskills and text types on reader performance simultaneously. A reading test was designed to provide equal numbers of items and texts representing four different subskills (Inference, Skimming, Scanning, and Coherence) and three text types (Narrative, Expository, and Argumentative). The participants in this study were 157 Korean male 12th graders attending Changwon Nam High School in Korea. For the analysis of these data, a Generalizability study (G-study) and a Decision study (D-study) (Brennen, 1983) were applied. Results show the effects of having various numbers of text types and subskills on the reliability of scores on this reading test.

INTRODUCTION

Although the processes of reading are often too dynamic and varied for different readers on different texts to be investigated, it is generally accepted that the interaction between readers and text variables is key to understanding the reading process. As a result, it has become common practice to divide reading-related research into two separate factors: the reader and the text (Alderson, 2000).

Reader Variables

When it comes to reader variables, the state of the reader's knowledge, broadly speaking, constitutes one significant reader variable, as does the reader's motivation to read. It is clear that the nature of the knowledge that readers bring to the reading process will affect the way they process and understand text. The development of schema theory

Second Language Studies, 20(2), Spring 2002, pp. 107-130

has attempted to determine the degree to which readers' knowledge affects what they understand. Schemata are seen as interlocking mental structures representing readers' knowledge (Anderson, 2000). When readers process a text, they integrate the new information from the text into their schemata. Schemata are often divided into formal schemata and content schemata. The former refers to knowledge of the language and linguistic conventions, including the organization of the text. The latter pertains to knowledge of the world, including the subject matter of the text.

In L1 reading research, since it is assumed that first-language readers already have basic syntactic and semantic knowledge, the effect of linguistic features in text is more commonly investigated than the knowledge of such features that readers have. Thus, research into linguistic knowledge has concentrated on vocabulary size and metalingustic knowledge (Read, 2000).

In second language reading research, it has been presupposed that learners must acquire linguistic knowledge before they can read. In particular, lexical knowledge of the text has been seen as essential for second language readers to process texts. According to Cooper (1984), without sufficient lexical knowledge, L2 readers showed substantial inability to use linguistic cues in the large context in order to deduce meaning and to recognize lexical relationships and meaning relationships between sentences.

In addition to L2 linguistic knowledge, the transfer of L1 reading ability into L2 reading ability has also been investigated. Bernhardt and Kamil (1995) claimed that L1 reading ability was a strong predictor of L2 ability based on their findings that L1 reading ability accounted for 20% of the variability in test-taker's reading scores. However, they argued that L2 linguistic knowledge was a consistently more powerful predictor, accounting for more than 30% of the variance.

With regard to the effects of the reader's content schemata on reading ability, Rumelhart (1985) showed that readers need knowledge about the content of the passage in order to be able to understand it. In other words, L2 readers are able to integrate new information with their previous knowledge related to it, but they find it difficult to integrate new information with non-existent information in processing texts.

Text Variables

The other side of the reader-text interaction is the text itself. In text variables, many aspects of text might facilitate or make the reading process difficult. Although the language of the text is known to be the major variable, there are other factors ranging from aspects of text content, to text types, text organization, and sentence structures. Simply put, it is generally assumed that abstract texts will be harder to understand than texts describing real objects since the former require more exacting inferencing skills than the latter. The text would be more readable if it were more concrete, imaginable, and interesting. In that vein, texts about familiar settings tend to be easier to process than those that are not.

Nevertheless, research has shown that lexical and syntactic knowledge in L2 are the strongest predictors in L2 reading performance among other factors (Bernhardt & Kamil, 1995; Cooper, 1984). However, when the linguistic variables, for example, lexis and sentence structures, are controlled in a reading test, such variables as readers' subskills and text types are more likely to influence their reading performance. For instance, the K-SAT (Korea Scholastic Aptitude Test) for English, consisting of a listening section (30%) and a reading section (70%), limits the range of vocabulary and sentence structures to those found within Korean EFL textbooks. Hence, in this kind of test, the subskills measured by each item and text type may be crucial parts of the variability in the students' reading test scores.

Subskill variables. Readers may be able to get the literal meaning of sentences but be unable to infer unstated assumptions made by the writer if readers have relevant linguistic knowledge of the text but might simply not possess the ability to process text. It has been suggested that reading ability can be divided into various subskills, and this notion is common in ESL teaching and testing. Munby (1978) identified several reading subskills for specifying ESP syllabus content, and there have been several studies about the degree to which it is possible to identify and label these separate skills in reading.

However, much controversy surrounds such research. There is contradictory evidence regarding whether subskills are separately identifiable. Different analyses of the same databases of skills have resulted in more or fewer factors that appear to underlie

adequate understanding of texts. Spearitt (1972), claimed there were four separate factors: recalling word meanings; drawing inferences from the content; recognizing a writer's purpose, tone and mood; and following the structure of the passage. Thorndike (1974) reanalyzed the same data and claimed that only one skill (word knowledge) could be distinguished from other skills. Several other studies on this issue of reading skills show no evidence of the existence of separate reading subskills. One study (Lunzer *et al*, 1979) suggests that reading consists of one single, global, and integrated aptitude, not distinguishable microskills. Similarly, Alderson concludes:

> Answering a test question is likely to involve a variety of interrelated skills, rather than one skill only or even mainly due to the fact that analyses of test performance do not reveal separability of skills, nor even a hierarchy of skill difficulty (1990, p. 436).

Nevertheless, since specifying subskills in language tests is clearly a widespread practice, there should be some way of investigating whether or not they are in fact being tested in the items themselves. In that vein, Alderson and Lukmani (1989) examined the questions of the existence of identifiably separate subskills and the idea of a hierarchy of subskills according to level of cognitive ability. Their study was based on items from a test used to assess the English reading ability of students at the end of their first year of undergraduate study—students who had completed a course in language and communication skills.

The study suggested that: (a) teachers showed relatively little agreement about the subskills tested by a range of reading comprehension test items, leading the researchers to question the possibility of relating individual test items to identified subskills; (b) the teachers disagreed considerably over the order of cognitive abilities required by the same items; and (c) students with lower English language proficiency performed as well as stronger students on items classified by the teachers as demanding a high level of cognitive skills, suggesting that cognitive levels were unrelated to levels of linguistic proficiency.

However, their study had the following three potential weaknesses as they admitted. First, there were some problems in the choice of items used in the study. They claimed that advanced level students performed on average no better than the low level students

on the items classified as measuring skills identified as higher order. This result would seem to be obvious if the items show poor discriminability, as nearly half the items analyzed did (six of the 14 items examined show discrimination values of 0.18, 0.08, 0.24, 0.10, 0.08, and 0.24). Since the establishment of adequate discriminability is a fundamental aspect of the reliability of norm-referenced test items and since this test was a norm-referenced language test, such items should have been eliminated from the study. Second, examination of the texts and the questions in the test suggested these low discriminability levels are not entirely surprising, as many items appear either to rely on background or cultural knowledge or to be answerable without reference to the text, suggesting they are testing things other than reading skills. Third, there was no exploration of why the raters made their choices about the skills tested by test items, and no attempt was made to examine where sources of disagreement existed. This highlights the need for making explicit the interpretations of the subskills described.

In contrast to the findings of Alderson and Lukmani (1989), a study by Brutten, Perkins, and Upshur (1991), investigating whether certain ESL reading comprehension skills were shown to lag behind others, as measured by performance on the TOEFL, found a high level of agreement among four raters about the skills tested by individual test items, using the Iowa test of basic skills taxonomy of reading skills (Hieronymus, Hoover, & Lindquist, 1986).

Likewise, Lumley (1993) examined the place of subskills in ESL syllabus and test design. His study showed that five readers, as a result of discussion of items and clarification of the meaning of subskill descriptions, were able to match subskills to individual test items in the reading comprehension test items. He also used Rasch analysis in analyzing reading comprehension test items to help validate teachers' perceptions about reading subskills. His study showed a significant correlation ($r = .716$) between the teachers' consensus regarding subskill difficulty levels and the Rasch analysis of item difficulty, providing some empirical validation for the teachers' perceptions.

However, his study also had the following drawbacks: First, question types (such as short answer, cloze, multiple choice, matching, true/false, completing a flow-chart, and labeling maps) were so wide-ranging that this factor alone might greatly affect the

variability of test scores; second, although the test consisted of a common text topic (environmental issues) and had 58 items based on two texts with a total length of approximately 1500 words, the test did not control the effect of text types which might influence the variance of test scores. Finally, since the judges rated only the selected 22 items, the number of items was too small to generalize the representativeness of items for each subskill.

Text type variables. Compared to the studies about the effects of reading subskills on reader performance, few studies have been conducted on the effects of text genre and type. Some studies have been conducted on the effect of text structure on reader performance (Carrell, 1984). The results of these studies showed that certain more highly structured English rhetorical patterns were more facilitative of meaningful recall for non-native readers in general, indicating an interaction between a reader's prior knowledge of and processing strategies for text structure and the rhetorical organization of the text. However, since the effects were exclusively examined for only the expository text type, those results cannot be applied to the effects of other text types.

There is a long tradition of research into the differences between expository and narrative texts. Generally, it has been suggested that expository texts are harder to process than narrative texts, perhaps because of the greater variety of relationships among text units, or possibly due to greater variety of content types (Alderson, 2000). A large number of empirical studies has demonstrated that narratives typically have a hierarchical structure, that readers are sensitive to such structure, and when the structure is used to guide comprehension and recall, both are facilitated (Glenn, 1978; Mandler, 1978; Carroll, 1985). In addition, narrative texts are more likely to induce visualization in the reader as part of the reading process than expository texts (Dennis, 1982).

Despite the fact that most reading tests consist of a variety of subskills and text types applied at the same time, studies of the relative effects of both subskills and text types on readers' performance are rare. In that sense, it is worth investigating the relative effects of item and text types simultaneously in this study. Thus, the following questions are examined in this study:

1. Do the different kinds of reading subskills and text types contribute to the variance among reading items?

2. If they do, to what extent do both reading subskills and text types affect the reliability of reading test scores?

Generalizability theory (G-theory). In order to examine the relative contribution to test variance of separate subskills and text types, a Generalizability study (G-study) was applied in this paper since G-theory allows the investigator to decide which facets will be of relevance to the assessment context of interest. A follow-up D-study was then designed to estimate the relative effects of these facets on test performance data. This estimation was expressed in terms of variance components, obtained from the expected mean squares in an analysis of variance where the main effects were persons and the facets.

The estimated variance components from the G-study were then used for making decisions about how the measurement procedure can best be improved. This involved designing a Decision study (D-study), or series of D-studies, which use the same data as the G-study and introduces the concept of the universe of generalization. The universe of generalization specifies a particular set of conditions for each facet to which the researcher would like to generalize. The variance components calculated for a D-study are meant to show the relative effects of specific numbers of conditions for each facet, not limiting them to single observations.

Application of G theory to language testing situations was first discussed in Bolus, Hinofotis, and Bailey (1982), who further iterated the usefulness of this systematic approach to the study of measurement error. Brown (1984) applied G theory to study the relative effects of numbers of items and passages in measuring engineering English reading ability. Then Brown and Bailey (1984) studied the effects of numbers of raters and scoring categories on the dependability of writing scores. Stansfield and Kenyon (1992) applied G theory to study the effects of numbers of tests and raters on oral proficiency interview scores. Brown (1990, 1993) applied G theory to the problems of estimating score dependability in criterion-referenced language tests. Kunnan (1992) also applied G-theory to a criterion-referenced test at UCLA. Bachman *et al*. (1995) used G-theory to investigate variability in test tasks and rater judgments on a speaking test. Recently, Brown and Ross (1996) examined the relative contributions of items types,

sections, and tests to the dependability of norm-referenced TOEFL scores. Most recently, Brown (1999) conducted a series of G studies to explore the relative contributions to TOEFL score dependability of various numbers of persons, items, subtests, languages, and their various interactions.

Simply put, G-theory allows the researcher to take all the various facets of a measurement procedure into account, and to differentiate their effects, via the estimated variance components, on the dependability of decisions or interpretations made from the test scores. In this vein, application of G-theory to my study is appropriate to investigate the relative effects of items, subskills, and text types in the reading comprehension test.

For the present study, a *p* x (*i:s:t*) (persons by items nested within subskills nested within texts) design with t (text) a fixed facet was applied in order to deal with the fact that each item occurs in one of the four subskills also nested within one of the three texts, while all of them are crossed with persons (p). The concept of random effects was important here. In a random-effects model, all levels in the experiment must be randomly selected from the much larger population of possible levels. The items and subskills facets in this study are considered random variables here, not because they are randomly selected from the population of all possible items and subtests, but rather on the basis of the concept of exchangeability (Shavelson & Webb, 1991). The perspective taken in this G-study, then, is that the various items and subskills are exchangeable and, therefore, are considered random effects. In contrast, the text facet in this study is considered fixed in the sense that it cannot necessarily be replaced by other types of texts.

METHODS

Participants

The participants in this study were 157 Korean male 12th graders attending Changwon Nam High School in Korea. They had all learned English for six years and their level of English proficiency varied as measured on the K-SAT for English as follows: the test scores of 41 students ranged from 60 to 80 (perfect score), those of 82 students ranged from 40 to 60, and those of 34 students were below 40.

In this school, three levels of EFL classes are offered to the students: basic, intermediate, and advanced level. Students are placed by their results on a mock K-SAT for English at the beginning of the semester. They take English classes five hours a week in school. Ten Korean EFL teachers are in charge of those classes. None of those teachers is a native speaker of English.

Materials

The reading test was designed to provide equal numbers of items and texts representing four different subskills (Inference, Skimming, Scanning, and Coherence) and three text types (Narrative, Expository, and Argumentative) as shown in Table 1.

Table 1

The General Description of Items and Text Types

	Subskills			
Test types	Inference	Skimming	Scanning	Cohesion
Narrative	1, 13, 25, 37	4, 16, 28, 40	7, 19, 31, 43	10, 22, 34, 46
Expository	2, 14, 26, 38	5, 17, 29, 41	8, 20, 32, 44	11, 23, 35, 47
Argumentative	3, 15, 27, 39	6, 18, 30, 42	9, 21, 33, 45	12, 24, 36, 48

The subskills and text types in this test were identified by analyzing the item specifications for the K-SAT for English. Three high school teachers were involved in rating the relationship between items and subskills and also between passages and text types. This test included 48 items based on multiple-choice format with five answer options within the range of vocabulary in the glossary of Korean EFL textbooks. All instructions in each item were written in Korean to make sure that the students should only be tested on understanding the passages. An example of an item in the test is as follows:

29. What is the main idea of the following passage? (Originally written in Korean)

> Do you become unhappy when clouds appear? Are you more cheerful on a sunny day than on a rainy day? Does the weather really affect your moods? Most of us feel that stormy weather can bring on sadness. This feeling may be caused by having to stay indoors for too long during bad weather. In contrast, a sunny day can make people happy and optimistic. When the weather is pleasant, people are friendlier and more willing to help each other.

① How the weather is changing
② Forecasting of the weather
③ The weather is good for exercise
④ The weather and human health
⑤ The weather and your feelings

Procedures

The reading test was administered for 110 minutes under standard conditions in April 2001. The students were all proctored by teachers and were not allowed to have anything other than the testing materials on their desks during the test.

After the administration, answer sheets were returned to me for scoring. The raw scores were calculated on the basis of the total number of correctly answered questions. There was no penalty for guessing.

Analyses

The analyses in this study began with descriptive statistics, item analysis statistics, and traditional reliability estimates (K-R20) were provided to describe the general classical theory nature of the test. Interrater reliability among the three raters was also calculated to estimate the degree of agreement among the three raters on matching subskills and text types to individual test items and passages. The Spearman-Brown Prophecy formula used to adjust for three-rater reliability. To investigate the effects of two variables in the reading test, a G-study (p x $(i:s:t)$ with t a fixed facet) was conducted. Finally, a D-study was conducted and G-coefficients (based on lower case delta for norm referenced error) were calculated for various numbers of items and

subskills so that the reader can directly observe the effect on reliability of these two random facets in various combinations of numbers of items and subskills in the reading test.

RESULTS

First, the descriptive statistics for the test are presented in Table 2. As shown in the table, the distribution of scores on the reading test was relatively normal: three indicators of central tendency, the mean (28.87), median (28), and mode (31), are very similar. In addition, skewedness is just a little positively skewed (.34) and kurtosis (-.71) also shows that the distribution is not too peaked. The test reliability was also quite acceptable (K-R20 = .85), considering the relatively small number of items and examinees. Apparently, the test itself was not too difficult for students as indicated by the mean of 28.87 out of 48.

Table 2

Descriptive Statistics (*N*=157, *k*=48)

Mean	*SD*	Median	Mode	Max	Min	S.E	Range	Skewedness	Kurtosis	K-R20
28.87	7.83	28	31	46	13	.63	33	.34	-.71	.85

Item facility and item discrimination coefficients (shown in Table 3) were also calculated to determine how well the items were functioning for this group of students.

Table 3

Item Analysis

Item	IF	ID	Item	IF	ID	Item	IF	ID	Item	IF	ID
1	.89	.14	13	.60	.37	25	.36	.27	37	.62	.57
2	.44	.48	14	.46	.20	26	.66	.32	38	.69	.50
3	.89	.25	15	.61	.33	27	.81	.42	39	.58	.21
4	.46	.47	16	.52	.38	28	.82	.29	40	.69	.33
5	.44	.11	17	.60	.23	29	.42	.39	41	.90	.35
6	.44	.25	18	.54	.52	30	.60	.52	42	.82	.31
7	.79	.33	19	.55	.33	31	.59	.48	43	.54	.28
8	.55	.55	20	.31	.32	32	.60	.28	44	.36	.34
9	.82	.37	21	.68	.31	33	.46	.29	45	.59	.40
10	.70	.30	22	.73	.31	34	.71	.41	46	.52	.35
11	.28	.38	23	.89	.29	35	.35	.24	47	.61	.47
12	.61	.36	24	.69	.45	36	.59	.36	48	.57	.39
									Mean	.60	.35

As can be seen from the table, the mean IF for the test is higher than the ideal IF of .50 and the average ID, .35, is also high, indicating that the items are functioning fairly well. According to Ebel's (1974) guidelines for making decisions based on ID, as can be seen from Table 4, all but two of the items would be considered acceptable (item 1 and item 5).

Table 4

Item Discrimination (after Ebel, 1974)

ID	Number of items
.40 and up (very good)	13
.30 to .39 (reasonably good)	21
.20 to .29 (marginal)	12
Below .19 (poor)	2

In order to validate the content of the test, I asked three raters to match each item and passage to specific subskills and text types. The interrater reliability among these three raters was calculated and adjusted using the Spearman-Brown Prophecy formula, and it suggested strong agreement among the three raters on matching subskills and text types to individual test items and passages as shown in Table 5.

Table 5 *Interrater Reliability* (n = 3)

	Interrater Reliability
Subskills	.97
Text	.90

In order to examine the persons, items, subskills, and text types variance components for the test results being examined in this study, a G-study was conducted. An analysis of variance (ANOVA) procedure was run using GENOVA for a persons by items nested within subskills nested again within texts (with texts as a fixed facet), or *p x (i:s:t)*. In other words, this G-study investigated the effects on the total reading test score variance of items, subskills, and texts (with the text facet viewed as a fixed effect based on the following three fixed text types: Narrative, Expository, and Argumentative). Based on the mean squares obtained in the ANOVA procedures, variance components were estimated as shown in Table 6.

Table 6

Variance Components for G-study

Sources	*df*	SS	Mean Squares	Variance Component	Percent of Variance
Persons (p)	156	2928.87500	1.27312	.02249	9.26
Texts (t)	2	2731.08678	.40963	*.00000	*.00
s:t	9	2767.14013	4.00593	*.00000	*.00
i:s:t	36	2919.31210	4.22700	.02569	10.57
pt	312	2996.12500	.21292	.00121	.50
ps:t	1404	3296.50000	.18826	*.00000	*.00
pi:s:t	5616	4536.00000	.19361	.19361	79.68

* This value was a negative variance component, which was rounded to zero (after Brennan, 1983, pp. 47-48)

The results of the G-study indicate that subskills and text types alone do not affect the test scores (i.e., the variance components for subskills and text facets are zero). The persons variance component (the object of measurement) only accounts for 9.26 % of the variance, whereas the interaction of persons with items nested within subskills nested within text facets accounts for almost 80 % of the variance and the items (nested within subskills nested within text) facet accounts for 10.57 %.

Using the variance components from the G-study, I also conducted a D-study to investigate the relative effects of varying the numbers of the fixed facet, text types, and two random facets, subskills, and items. Summaries of the statistics found in the D-study are presented in Tables 7, 8, and 9.

Table 7

G-Coefficients for D-Study with Three Text Types

	Subskills						
Items	1	2	3	4	5	6	7
1	.26	.41	.51	.58	.64	.68	.71
2	.41	.58	.68	.74	.78	.81	.83
3	.51	.68	.76	.81	.84	.86	.88
4	.58	.74	.81	.85	.88	.89	.91
5	.64	.78	.84	.88	.90	.91	.92
6	.68	.81	.86	.89	.91	.93	.94
7	.71	.85	.88	.91	.93	.94	.95
8	.74	.86	.89	.92	.94	.94	.95
9	.76	.88	.90	.93	.94	.95	.96
10	.78	.91	.91	.93	.95	.95	.96
15	.84	.93	.93	.95	.96	.97	.97
20	.87	.95	.95	.97	.97	.98	.98
30	.91	.95	.96	.98	.98	.98	.99
40	.93	.97	.98	.98	.99	.99	.99
48	.94	.97	.98	.99	.99	.99	.99

Generally, the results show that there is considerable reliability gained from having the various text types and subskills rather than having one long, homogeneous test. In other words, there is an increase in reliability due to increases in the numbers of text types and subskills involved while holding the number of items constant. For instance, a test configured with the same 10 items and four subskills but in one text type is estimated to be dependable at .79; with two text types, it is predicted to be .88; and, with three text types, it would be .93.

Table 8

G-Coefficients for D-Study with Two Text Types

Items	Subskills						
	1	2	3	4	5	6	7
1	.19	.31	.41	.48	.53	.57	.61
2	.31	.48	.57	.64	.69	.72	.75
3	.41	.57	.66	.72	.76	.79	.81
4	.48	.64	.72	.77	.81	.83	.85
5	.53	.69	.76	.81	.83	.85	.87
6	.57	.72	.79	.83	.85	.87	.89
7	.61	.75	.81	.85	.87	.89	.90
8	.64	.77	.83	.86	.88	.90	.91
9	.66	.79	.84	.87	.89	.90	.91
10	.69	.81	.85	.88	.90	.91	.92
15	.76	.85	89	.91	.92	.93	.94
20	.81	.88	.91	.93	.93	.94	.95
30	.85	.91	.93	.94	.95	.95	.95
40	.88	.93	.94	.95	.95	.96	.96
48	.90	.93	.95	.95	.96	.96	.96

Tables 7, 8, and 9 also allow considering other potential combinations of numbers of items, text types, and subskills as part of the D-study to help in deciding what is the optimal number of items, text types, and subskills to include in future versions of this and other tests. For example, by looking in Table 7 at the point where four subskills with three text types intersect with seven items (for a total of 28 items), the table reveals that a G-coefficient of .91 is predicted. In my study, I actually used a combination of four subskills and four items (for a total of 16 items) with three text types and G-coefficient turned out to be .85, which is exactly equivalent to the K-R20 reliability estimate of .85 as would be expected.

Table 9

G-Coefficients for D-Study With One Text Type

Items	Subskills						
	1	2	3	4	5	6	7
1	.10	.19	.25	.31	.36	.40	.44
2	.19	.31	.40	.47	.52	.56	.60
3	.25	.40	.50	.56	.61	.65	.68
4	.31	.47	.56	.63	.67	.71	.73
5	.36	.52	.61	.67	.72	.75	.77
6	.40	.56	.65	.71	.75	.77	.79
7	.44	.60	.68	.73	.77	.79	.81
8	.47	.63	.71	.76	.79	.81	.83
9	.50	.65	.73	.77	.80	.82	.84
10	.52	.67	.75	.79	.82	.84	.85
15	.61	.75	.80	.84	.86	.87	.88
20	.67	.79	.84	.86	.88	.89	.90
30	.75	.84	.87	.89	.90	.91	.91
40	.79	.86	.89	.90	.91	.92	.92
48	.81	.87	.90	.91	.92	.92	.93

DISCUSSION

All in all, the results of my study consistently provide direct answers to the research questions about the relative effects of numbers of items, subskills, and text types on Korean EFL reading test scores. Examining the G-study variance components shown in Table 6 in terms of their relative magnitude reveals the relative contributions of persons, items, subskills, and texts as a fixed facet, as well as their interactions. The variance of the test scores is mostly accounted for by persons and items (nested within subskills

nested within texts) and the interaction of persons with items (nested within subskills nested within texts). The relative contributions to variance of the subskills and text types facets are so minimal that these two facets alone do not seem to affect the variance of reading test scores. This result appears to support Alderson's (1990) view of reading skills as rather interrelated skills, not separable ones.

However, it should be noted that the variance component for the persons by text interaction also accounts for a small amount of the total variance (0.5%), whereas the variance component for the persons by subskills nested within texts interaction is zero. These findings go with the findings of previous studies about the effects of subskills and text types on reading performance. Particularly, with regard to the effects of subskills on the test results, although I chose the most common four reading subskills usually measured in the K-SAT for English exam, students' reading performances do not appear to vary over the subskills measured in this reading test (see Table 6). Similarly, Rost (1993) found only one broad factor, general reading competence, for German first language readers. He concludes that, as found in studies using other reading tests, the test he used "cannot measure several clearly distinguishable components of reading comprehension" (p. 80).

Compared to subskills, the text types are known to have an influence on readers' performance as shown in several studies (Brown, 1984; Carrell, 1984). Especially, readers seem to perform differently depending on whether they work on narrative or expository texts. Denis (1982) argued that readers could recall narrative texts more easily and accurately than expository texts because of 'visualization' in the reader as part of the reading process. In addition, as Carrell (1984) suggested, this difference might be due to the fact that those two different texts types consist of different rhetorical organization patterns.

However, as represented by the test types on the test, argumentative texts share plenty of common features with expository texts other than the purpose of the texts. Therefore, this commonality between expository and argumentative texts may have lead to the text types facet not taking up much of the test score variance in my study (see Table 6).

With regard to the results of the D-study, although the two factors, subskills and texts types, alone hardly affected the variance of entire test scores according to G-study,

because they were involved in substantial interaction effects, the text types and subskills facets influenced the predicted reliability indices. For example, the reliability was not the same for one subskill and text type and more than one subskill and text types with the number of items held constant (see Tables 7, 8, and 9). In other words, the reliability was enhanced by having an increased number of text types and subskills even if the number of items were to be kept the same. Thus, the expansion of the number of text types and subskills may have a substantial effect on the reliability of the scores on the reading test and this study has demonstrated how that would work.

CONCLUSION

In general, this study indicates the following about the effects of numbers of items, subskills, and texts types on reading ability: First, in a real test situation in which all diverse subskills and text types exist at the same time, their relative contributions to the total variance of reading test scores can be estimated. Second, to some degree, having various text types and subskills in a reading test appears to have a strong beneficial effect on the reliability of scores for the test. In other words, including a variety of reading factors like subskills and text types might be a sound policy decision from the reliability point of view. In addition, based on Tables 7, 8, and 9, further policy decisions can be made about the relative merits of various numbers of items, text types, and subskills.

However, this study also has the following limitations: First, the text types on this test were not representative of all possible text types. They are tentative text type categories designed for this particular test; secondly, the results cannot be generalized beyond the particular group of Korean EFL 12th graders in Korea tested here, especially given the relatively small number (n = 157) and homogeneity on the test (all Korean male 12th graders) of the participants in this study.

In the course of conducting this study, a number of questions occurred to me that I was unable to address. They are presented here in the hope that they will be investigated in the future:

1. Would similar results be obtained if this study were replicated using other students in different ESL or EFL contexts?

2. What can G-theory tell us about the separate effects of text types and subskills on the reading test separately when the other variable is controlled?
3. How would the predicted and actual reliability results match if the test were actually redesigned and administered again?
4. What new information would be garnered from using a multifaceted Rasch analysis of the same facets for these data?

REFERENCES

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5*, 253-270.

Alderson, J. C. (1990). Testing reading comprehension skills. *Reading in a Foreign Language, 6*(2), 425-438.

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

Bachman, L. F., Lynch, B. K, & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*, 238-257.

Bernhardt, E. B., & Kamil, M. L. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypothesis. *Applied Linguistics, 16*(1), 15 –34.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning,* 32, 245-258.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa city, IA: American College Testing Program.

Brown, J. D. (1984). A norm-referenced engineering reading test. In A. K. Pugh & J. M. Ulijn (Eds.), *Reading for professional purposes: Studies and practices in native and foreign languages London*: Heinemann Educational Books.

Brown, J. D. (1990). Short-cut estimators of criterion-referenced test consistency. *Language Testing, 7*, 77-97.

Brown, J. D. (1993). A comprehensive criterion-referenced testing project. In. D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research.* Alexandra, VA: TESOL

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16*, 217-238

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*(4), 21-42.

Brown, J. D., & Ross, J. A. (1996). Decision dependability of item types, sections, tests, and the overall TOEFL test battery. In Milanovic, M. & Seville, N. (Eds.), *Performance testing, cognition and assessment.* Cambridge: Cambridge University Press

Brutten, S. R., Perkins, K., & Upshur, J. A. (1991). *Measuring growth in ESL reading*. Paper presented at the Thirteenth Annual Language Testing Research Colloquium, Princeton, NJ.

Carrell, P. (1984). The effects of rhetorical organization on ESL readers. *TESOL Quarterly, 18*(3), 441- 469.

Carrell, P. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly, 19*(4), 727-753.

Cooper, M. (1984). Linguistic competence of practiced and unpracticed nonnative readers of English*. Reading in a foreign language.* London: Longman.

Dennis, M. (1982). Imaging while reading text: A study of individual differences. *Memory and Cognition, 10*(6), 540-545.

Glenn, C. (1978). The role of episodic structure and of story length in children's recall of simple stories. *Journal of Verbal Learning and Verbal Behavior, 17*(2), 229-247.

Hieronymous, A. N., Hoover, H. D., & Lindquist, E. F. (1986). *Preliminary teacher's guide. Multilevel battery levels 9-14.* Chicago, IL: Riverside Publishing.

Jordan, R. R. (1997). *English for academic purposes.* Cambridge: Cambridge University Press.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor analysis. *Language Testing, 9*(1), 30-49.

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

Lunzer, E., Waite, M., & Dolan, T. (1979). Comprehension and comprehension tests. In E. Lunzer & K. Garner (Eds.), *The effective use of reading.* London: Heinemann Educational Books.

Mandler, J. (1978). A code in the node: The use of story schemata in retrieval. *Discourse Processes, 1*(1), 14-35.

Munby, J. (1978). *Communicative syllabus design.* Cambridge: Cambridge University Press.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press

Rost, D. (1993). Assessing the different components of reading comprehension: Fact or fiction? *Language Testing, 10*(1), 79-92.

Rumelhart, D. E. (1985). Towards an interactive model of reading. In H. Singer & R. B. Ruddell (Eds.), *Theoretical models and process of reading.* Newark, DE: International Reading Association.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage

Spearitt, D. (1972). Identification of subskills of reading comprehension by maximum likelihood factor analysis. *Reading Research Quarterly, 8*, 92-111.

Stansfield, C. W., & Kenyon, D. M. (1992). Research of the compatibility of the oral proficiency interview and the simulated oral proficiency interview, *System, 20*, 347-364.

Thorndike, R. L. (1974). Reading as reasoning. *Reading Research Quarterly, 9*, 135-147.

Sunyoung Shin
Department of Second Language Studies
1890 East-West Road
Honolulu, HI 96822

suns@hawaii.edu