Evaluating Reading Support Systems through Reading Skill Test

Teiko Arai (arai-teiko@g.ecc.u-tokyo.ac.jp),
Kyosuke Bunji (bunji@p.u-tokyo.ac.jp)
University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Naoya Todo (ntodo@human.tsukuba.ac.jp)
University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki 305-8577, Japan
Noriko H. Arai (arai@nii.ac.jp)
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
Takuya Matsuzaki (matuzaki@nuee.nagoya-u.ac.jp)
Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

Abstract

We propose a computer-based testing environment, Reading Skill Test, to measure the effects of various types of systematic reading support systems. We prove its validity, reliability and one-dimensionality using 31,000 subjects. The effects of *furigana* system on the 5th to 8th grade students are analyzed using this environment. Furigana is a widely used Japanese reading support system that has been believed to be beneficial especially for pupils. Despite our expectation, we have to conclude that furigana failed to improve pupils' reading significantly, and discuss why it did so.

Keywords: reading comprehension, reading support systems, item response theory

1. Introduction

Refinement of typesetting styles and improvement of fundamental writing systems have been investigated for enhancing the legibility and readability of text for adult readers as well as providing an aid for children and dyslexic people. Most notably, the effect of typesetting parameters (e.g., font type, character size, and line spacing) on the reading speed of adult readers has been extensively studied since the 19th century (Roethlein, 1912; Tinker, 1963). Electronic text allows a more radical enhancement of the text presentation style. Among others, the effect of hyperlinking on text comprehension attracted much attention during the 1990s (DeStefano & LeFevre, 2007).

In most studies, the effect of these text presentation styles was measured either by a simple metric such as reading speed or a very high-level metric such as the score of a posttest. A more fine-grained and highly replicable metric is desirable for both the evaluation of the practical utility of a text presentation style and the cognitive science of reading using the enhanced text presentation as a research material.

In the current paper, we take up *furigana*-attached text as the research material and investigated its effect on short-text comprehension utilizing the results of our past study on a large-scale language skill test (Arai et al., 2017). Furigana-attached text is a style of Japanese text presentation, in which the pronunciation of Chinese characters (kanji) in the text is provided in the hiragana characters (syllabic letters) that are rendered in parallel with the kanji characters (Fig. 1). It is considered effective for educational purposes and is commonly used in texts for children including primary school

Figure 1: Furigana-attached text

textbooks. As such, the effect of furiganas on reading comprehension is of great interest both from the viewpoint of cognitive science of reading (i.e., phonetic coding and reading) and educational psychology.

We conducted an experiment in which 198 subjects (5th to 8th grade students) participated. Half of them answered six kinds of reading skill test questions provided in furigana-attached format and the rest answered the same questions without furigana. The results were analyzed utilizing the statistics about the same test questions that we previously collected from more than 31,000 subjects.

Our main findings are summarized as follows:

- The response time (RT) of the 8th grade students was longer on the furigana-attached test but there was no difference in the RT of lower grade students, which indicates, contrary to common expectations, the lower graders do not utilize furiganas.
- No positive effect of furigana was found on the test scores across all grades and all component skill types.

The rest of the paper is organized as follows. Section 2 provides an overview of the Japanese writing system and a review on reading support systems. Section 3 describes the design of the reading skill test and demonstrates its validity and reliability. Section 4 summarizes the hypothesis verified in the experiment. Section 5 describes the test material used in the current study and the experimental procedure. Section 6 presents the experimental results and Section 7 discusses the results. Section 8 concludes the paper.

2. Background

This section provides an overview of the Japanese writing system, the furigana-attached text presentation style, and its relation to phonological coding. A short review on other kinds of reading support systems is also provided.

2-1. Japanese Writing System

Three kinds of character are used in the Japanese writing system: kanji (Chinese characters), hiragana, and katakana.

Kanji characters are logographic. The 2,000 most frequent kanji characters account for approximately 99% of kanji in Japanese. Primary school students learn 1,000 kanji characters in six years. Hiragana and katakana are syllabic characters. Hiragana characters are used for conjugative suffixes and postpositions. Katakana characters are used for loan and foreign words.

More than 70% of kanji characters have two or more pronunciations. One can guess some of these pronunciations from a component of the character but other pronunciations, especially ones based on traditional Japanese words, are difficult to guess without knowing them in advance.

Due to the irregular correspondence of kanji characters and their pronunciations, it has often been assumed that the meaning of a word written in kanji is accessed directly by its orthography rather than through a phonological code. In fact, there is an experimental result that suggests it (Kimura, 1984) but there are also results indicating parallel access through both orthographic and phonological representations (Wydell et al., 1993; Sakuma et al., 1998).

Furigana-attached style is a way to compensate for the irregularity of the kanji pronunciations. In furigana-attached style, the pronunciation of each kanji character in a sentence is provided with small hiaragana characters (furiganas) on top of the kanji. Because there is a one-to-one correspondence between hiragana symbols and sounds, the reader knows the pronunciation of the kanji by the small hiragana without ambiguities. Similar methods are used in Taiwanese Mandarin (bopomofo) and Arabic (harakat).

Furigana-attached style is commonly used in reading materials for children, including textbooks. They are hence accustomed to it. Furigana-attached text naturally encourages phonological coding or subvocalization. Therefore, its effect on reading comprehension is of great interest from the viewpoint of the psychology of reading, especially about the role of sound in silent reading.

2-2. Systematic Reading Support through Improved Text Presentation

Furigana is a systematic way to help readers' comprehension through a special form of text presentation. Depending on the language and the presentation media, various other forms of systematic reading assistance are conceivable.

At a more fundamental level, there have been a vast number of studies pursuing the ideal style of typesetting for text legibility, through the choice and tuning of the parameters such as font type, font size, line spacing, and line width. The primary metric of the goodness of these parameters has been reading speed. Many studies have also tested readers' understanding of text, but the tests mostly concerned recall of content by the readers (e.g., Dyson & Haselgrove, 2001; Gasser et al., 2005). It is not straightforward to identify what aspect of the text processing is most affected by the typesetting factors only through coarse-grained metrics such as reading speed or high-level tests on the content of the text.

Hypertext is a versatile text presentation system, on which the reader can traverse across a document through clickable links that connect semantically related sections. There has been an expectation that hypertext is an effective media for education materials by virtue of its interactive nature. Its effect on document comprehension has hence been investigated by many researchers. These studies are however mostly based on tests that only check overall understanding of the document (e.g., Plass et al., 2003; Eveland et al., 2004) or the efficiency of information retrieval (e.g., McDonald & Stevenson, 1996; Lin, 2004). While it was observed across these studies that text comprehension is impaired by too many hyperlinks (DeStefano & LeFevre, 2007), it is not entirely clear what aspect of the document processing is hindered by them.

It is highly desirable to examine the effect of all these reading support systems on the basis of a fine-grained and replicable test framework. Consideration of such effect would be especially mandatory when a system is introduced to primary education classrooms. Our reading skill test offers such an environment by providing a suite of test questions organized along a hierarchical reading skill model and statistics about the test collected across different age groups with a wide range of scholastic ability.

3. Design of RST and its Reliability

3.1 Six Component Skills and their Measurement

We defined six component skills relevant to reading and designed a Reading Skill Test (RST) in (Arai et al, 2017).

1. Dependency Analysis (DEP): The skill of recognizing the dependency relation between words and phrases in a given sentence.
2. Anaphora Resolution (ANA): The skill of anaphora resolution. ANA is comprised of two elements, which are Demonstrative Anaphora Recognition (DANA) and Zero Anaphora Restoration (ZANA).
3. Paraphrasing (PARA): The skill of recognizing that a sentence is the same as another one. PARA is comprised of three elements: Structural Paraphrasing (PARA1), Lexical Paraphrasing (PARA2) and Logical Paraphrasing (PARA3).
4. Logical inference (INF): The skill of reading a sentence and determining what can be inferred from the sentence, what conflicts with it, and what does not relate to it.
5. Representation (REP): The skill to represent an image (figure or table) by comprehending a sentence of the textbook.
6. Instantiation (INST): The skill of understanding how to use a term correctly according to a given definition of the term. INST is comprised of two elements, INST1 (definitions taken from dictionary) and INST2 (definitions taken from mathematics and science).

In short, DEP and ANA assess a person's skill in recognizing the syntactic and the semantic structure of sentences. PARA and INF check the ability of recognizing the logical relation between sentences. REP and INST require "deeper" semantic processing. Furigana-attached tests may have different effects on the results because of the

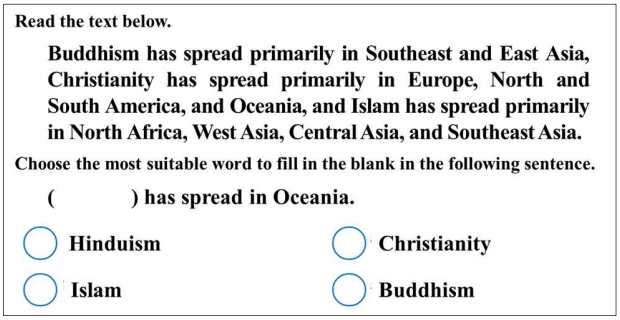

Figure 2. A sample item of DEP, of which text taken from a junior high school textbook

difference of the skills tested in them. We verify this possibility through the experiment.

RST is a computer-based test in a multiple choice style. We created all of the items on the basis of dictionaries, and textbooks that have been approved by the Ministry of Education, Culture, Sports, Science and Technology and are being used in Japanese elementary schools and high and junior high schools. It is designed so that each question (item) is shown on one screen without any scrolling (Figure 2). Each examinee answers items randomly chosen from the item pool one after another within a time limit: one examinee may respond to only 10 items while another responds to 40. Item Response Theory is used to determine an item *j*'s difficulty parameter, denoted by b_j, and simultaneously an examinee *i*'s characteristics denoted by θ_i as described in (Arai et al, 2017). We do not take the number of items an examinee *i* responded to into consideration in calculating θ_i. Note that b_j is only estimated after at least 1,000 examinees have solved it.

3.2 Psychometric Properties of RST

To examine the validity, reliability, and one-dimensionality of each test, correlations between the θ's, reliability coefficients based on average item information functions and number of responses, and average factor loadings in categorical factor analysis were estimated using responses from more than 31,000 examinees, which are from elementary-school students to adults. Table 1 shows the results. The correlation coefficients of θs of all the types are positive, and range from 0.398 to 0.660, which shows that they are all closely related, but measure different aspects of reading skills. It shows the validity of our reading skill test. Note that it turned out that INST1 and INST2 measure different aspects of reading skills, because their correlation coefficient is only 0.419. This is interesting since both ask if an examinee understood the given definitions. Although all reliability coefficients (ρ) are below 0.7, RST is scheduled to be a computerized adaptive testing (CAT). In CAT, appropriate items for each examinee will be selected from the item pool and CAT will increase measurement precisions of each participant's θs (van der Linden, 2010). Therefore, this means that RST has a practically sufficient level of reliability (Nunnally, 1978). The means of the factor loadings (λ) are over 0.35. This shows the one-dimensionality of each test, which is required in IRT (Yong & Pearce, 2013).

Unlike in the US or EU countries, one has to pass an entrance exam to enter even a public high school in Japan. Examinees' scholastic abilities are examined by their performance on paper tests and GPA. Each school is weighted by T-Score, ranging from 20 to 80, which represents the difficulty to enter the school. Two major cram schools (Hensachi.net and Kateikyoshi-no-Torai (Torai)) put the list of T-scores of high schools on the Internet. We calculated the correlations between the T-Scores of high schools and the means of the seven θs of the students. Table 2 shows the results. The correlation coefficients range from 0.799 to 0.999, which are extremely high. It shows that how well an examinee does in our reading test is closely related his/her comprehensive scholastic ability. This also shows that our test has enough validity.

Table 1. Correlations, reliability coefficients, and means of the factor loadings

Type	ANA	DEP	INF	INST1	INST2	PARA	REP
ANA	1.00	.660	.529	.567	.533	.610	.569
DEP		1.00	.486	.471	.461	.575	.523
INF			1.00	.398	.416	.467	.502
INST1				1.00	.419	.465	.488
INST2					1.00	.504	.568
PARA						1.00	.572
REP							1.00
ρ	.650	.603	.503	.553	.494	.583	.639
λ	.547	.576	.359	.521	.579	.408	.512

Table 2. Correlations between seven θs and T-Scores

Type	ANA	DEP	INF	INST1	INST2	PARA	REP
H-net	.869	.857	.882	.836	.799	.986	.863
Torai	.866	.846	.866	.839	.803	.999	.865

4. Hypothesis

There are two possible effects of furigana. Firstly, it aids silent reading by facilitating subvocalization. Secondly, the pronunciations provided in furiganas help the subjects identify a word that they know by sound but not by the orthography in kanji. In other words, the second effect takes place when a student, especially of lower grades, does not recognize a word written in kanji but she/he actually knows the word and recognizes it in hiragana-format.

Our main interest was whether the two effects of furigana improve reading performance. In addition, as a prerequisite, we needed to verify whether the subjects actually used furiganas. We thus wanted to see:

A) Whether the students actually use furiganas,
B) Whether the effect of assisting phonological, processing of text improves reading performance, and
C) Whether the word identification assisted by furiganas improves reading comprehension performance.

Table 3: Test times of each section

	5-6th grade	7-8th grade
DEP	180 sec.	372 sec.
PARA	180 sec.	317 sec.
REP	180 sec.	392 sec.
(break)		
ANA	180 sec.	377 sec.
INST	180 sec.	412 sec.
INF	180 sec.	325 sec.
Total	1080 sec. (18 min.)	2195 sec. (37 min.)

Table 4: Number of students participated in the test

Test \ Grade	5th	6th	7th	8th
Furigana-attached	22	27	23	24
Normal	23	33	23	23

However, it is difficult to verify A to C as is. As for A), there is no direct way to check whether the subjects use furiganas. As for B), it is difficult to observe phonological processing directly. As for C), there is no way to check whether a subject knows a word in the test question (by sound and/or orthography) without interfering the test itself.

We reinterpret A to C as verifiable hypotheses as follows. In the case of A, if a subject uses furiganas, he/she has to process more information and hence the reading speed must be slower than when he/she doesn't use furiganas. Therefore, the response time of the subject who takes the furigana-attached test must be longer than the average of the rest of the subjects, and the difference must be larger for lower grade students who are supposed to use furiganas more frequently.

Both B and C should result in higher scores on the furigana-attached version of the test. A possible difference between the results of B and C is as follows: If C is correct, the positive effect of the furiganas is more prominent on the text including difficult kanji, which is likely to be unknown in the kanji-format but possibly known by sound. Accordingly, B and C were rewritten as c) and d) as follows.

a. The response time of the subjects who took the furigana-attached test is longer than that of the subjects who took the test without furiganas
b. The difference in the response time is larger in the lower grade students than in the higher grade students
c. The subjects who took the furigana-attached test get higher scores
d. The positive effect of the furiganas on the test score is more prominent on the questions including difficult kanji

By using RST, we verified a)-d) to see whether furiganas have educational effects; and if so, for which age group.

5. Method

Six component skills are measured separately in the RST. We selected 10 questions for each skill such that:

- The difficulty of the questions was evenly distributed.
- The questions included relatively difficult kanji.
- The statements of the questions were not too short.
- The topics of the questions (i.e., natural or social science) were evenly distributed.

We then made the furigana-attached version of the same questions.

Experiments were conducted at a public school attended mostly by children living in the area. Thus, the subjects could be regarded as unbiased samples. 198 students, from 5th graders to 8th graders, took the test: half of the students of each grade solved the furigana-attached test, and the other half solved the test without furiganas.

The students were instructed to answer the questions as accurately and quickly as possible. A sample question was provided at the beginning of each section. Table 3 summarizes the test times of each section and Table 4 shows the number of students who participated in the test.

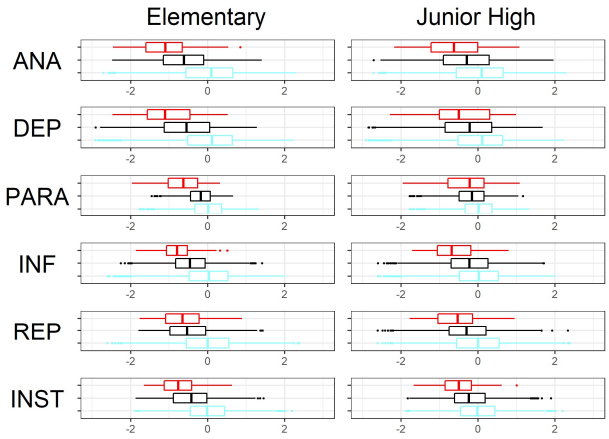

Figure 3: Box Plots of the Subjects' θ_i

6. Results

First, we measured the characteristics, θs, of the subjects who participated in this experiment and compared them with those of the 31,000 subjects collected previously. Figure 3 shows the result. The horizontal axis stands for the characteristics, where 0 is the mean of all the 31,000 subjects. Red plot boxes represent the θs of the subjects in each grade of this experiment whereas black ones are those of all the subjects of the corresponding grades. It indicates that the θs of all the six skill types of the subjects in this experiment are lower than the others. This is appropriate, as furigana is expected to be more useful for students with lower reading skill. Blue plot boxes show the θs of all the examinees across age groups.

To investigate hypotheses (a) and (b), we used a linear mixed-effect model with the formula,

$$\ln(RT_{ij}) = (\beta_0 + u_i + u_j) + \beta_1 * x_{Furigana} + e_{ij} \qquad (1)$$

where RT_{ij} is the response time of subject i to item j. We included a dummy variable, indicating whether furigana is attached, as the fixed effect. In addition, we included random intercepts of subject i (u_i) and item j (u_j).

We concluded that there is no consistent difference regarding the response time between the presence and absence of furigana across the grades.

Table 5: Fixed Effect Estimates (β_1) in Linear Mixed-Effect

Type	all	5th	6th	7th	8th
ANA	0.097	0.057	0.053	0.012	0.272**
DEP	0.055	−0.024	0.110	−0.134	0.231**
PARA	0.068	−0.020	0.093	−0.187	0.362*
INF	0.236**	0.314	0.222	0.078	0.365*
REP	0.131	−0.001	0.181	−0.163	0.485**
INST1	0.090	0.096	0.004	−0.117	0.371*
INST2	0.079	−0.099	0.076	−0.078	0.410*

Note: *p<0.05; **p<0.01

We scrutinized the effects in each grade and found that all of the coefficients are significant in the 8th grade. By contrast, there is no significant coefficient in lower (5-7th) grades. This is surprising because furigana is meant to support lower-grade pupils. However, the result showed that only the highest-grade students in the experiment used furigana.

Next, we tested hypothesis (c) with a mixed-effect logistic regression. The model formula was

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \left(\beta_0 + u_i + u_j\right) + \beta_1 * x_{Furigana} + e_{ij} \qquad (2)$$

where p_{ij} is the probability of subject i answering correct for item j. We used the same predictor as Eq. (1).

Table 6: Fixed Effect estimates (β_1) in Mixed-Effect Logistic Regression

Type	all	5th	6th	7th	8th
ANA	−0.174	−0.156	−0.440	−0.420	0.215
DEP	0.084	0.473	0.124	−0.391	−0.046
PARA	0.046	0.358	0.081	−0.502	0.021
INF	0.029	−0.295	−0.100	0.328	0.130
REP	−0.074	0.431	−0.025	−0.802*	0.127
INST1	0.021	0.583	−0.628	−0.201	0.234
INST2	0.627*	0.917	0.714	0.823	0.215

Note: *p<0.05

Table 6 summarizes the estimates of β_1 in Eq. (2). There is no consistent difference between the results on the test with and without furigana in the 5-7th grades. This is plausible from the previous result showing that they seemed not to use furigana. Then, how about 8th grade students who seemed to use furigana? It is surprising that no significant estimates are found even in the 8th grade. This result clearly shows that the presence of furigana does not make any significant difference in reading comprehension, even if they use it.

Finally, we checked hypothesis (d) with the formula,

$$\ln\left(\frac{p}{1-p}\right) = \left(\beta + u_i + u_j\right) + \beta_1 * x_{Furigana} + \beta_2 * x_{Kanji} + \beta_3 * x_{Furigana} * x_{Kanji} + e_{ij} \qquad (3)$$

In this analysis, we added the number of difficult kanji (centered at the grand mean) and an interaction term of kanji and furigana. We counted the kanji characters that are not taught in primary school as "difficult."

Hypothesis (d) will be supported if the interaction term is positively significant. However, only one estimate (INST2) was significant; the other estimates were very small or even negative. From the results, we concluded there is no difference in the effectiveness of furigana considering the number of difficult kanji.

Table 7: Coefficient estimates of kanji (β_1), furigana (β_2), and interaction term (β_3)

Type	Furigana	Kanji	Furigana*Kanji
ANA	−0.174	0.108	−0.113
DEP	0.062	−0.048	0.092
PARA	−0.083	−0.059	0.010
INF	0.032	0.062	0.039
REP	−0.068	0.234	−0.108
INST1	0.628*	0.033	−0.059
INST2	−0.030	−0.225	0.458*

Note: *p<0.05

7. Discussion

Our finding in the experiment is summarized as follows:

a. The response time increased significantly in the furigana-attached test only in the 8th grade.
b. The response time did not increase with the use of furigana in the lower grades.
c. The presence or absence of furiganas did not make significant difference in the score of RST.
d. Even on the questions including difficult kanji, no positive effect of furiganas was found in the score of RST.

Therefore, hypothesis (a) was supported only partially and (b), (c), and (d) were rejected. We discuss these results, focusing on the use of furiganas (hypothesis (a) and (b)) and the effect of furiganas (hypothesis (c) and (d)).

First, we discuss the results on the use of furiganas. The reason why only the 8th grade students used furiganas is unknown. Furigana is believed to assist reading comprehension, and it is widely practiced in Japan. Nonetheless, we revealed for the first time that elementary school students do not use furiganas as we expected. Further investigation is necessary to clarify who use furigana on what kind of text. We shall experiment with different conditions and test environments for it.

Next, we discuss the effect of furiganas. Why did the presence of furiganas make no significant difference in the score of the test? As explained in Section 4, we expected that the score of the RST would increase by the furigana's effect on phonological processing and/or word identification. The experimental results suggest our basic premise on the effect of furigana was wrong. We consider two explanations:

c-1. Furigana does not assist phonological processing.
c-2. Furigana does not assist word identification.

Regarding c-1, firstly, there is a possibility that the assistance of phonological processing indeed helps reading comprehension, but furigana is not a right way to do it. For example, screen readers (i.e., reading while hearing) may assist phonological processing more directly through sound instead of letters, through it would incur additional cognitive load due to the multimodality. We need another experiment comparing different assistance methods. Secondly, there is a possibility that the subjects had no custom to subvocalize while silent reading.

There are several possibilities for c-2. First of all, note that there are four types of words in relation to a subject' auditory and visual vocabulary, which is determined by whether or not a subject knows a word by sound and by its kanji notation. Four scenarios are thus conceivable:

Case 1-1. If there are many words that the subjects know by sounds but not by spelling, furiganas would be most effective.

Case 1-2. If there are only a few words known by sounds but not by spelling, furiganas would not be effective.

Case 2. If there are a large number of words that the subjects know by spelling, furiganas would not be effective whether or not the words are known by sound.

Case 3. If there are many words that the subjects do not know by either sounds or spelling, furiganas would not be effective.

That is, furiganas have no effect on word identification in the three scenarios except for Case 1-1.

Since the subjects were 5-8th graders, it is not plausible to assume that most subjects knew the majority of the words by spelling, including those spelled with difficult kanji. Case 2 is hence rejected. Meanwhile, Case 1-2 should be held as a possibility since it is difficult to directly examine how many words are known only by sound.

Case 3 is plausible enough: even if the pronunciation of an unknown word is provided, one cannot understand the word. Regarding Case 3, we plan to introduce a vocabulary quantity test as an additional experiment and to match the result with the result of the furigana experiment.

Finally, there is a possibility that the role of phonological processing and word identification was relatively small in the whole process of "reading comprehension," which is comprised of complicated steps; if it is the case, it is not surprising that the subjects' ability parameter θs did not increase significantly with furiganas.

We will continue to examine the effectiveness of various text presentation methods using the RST as the metric.

8. Conclusion

We proposed a computer-based testing environment, Reading Skill Test, to measure the effects of various types of systematic reading support systems. We proved its validity, reliability, and one-dimensionality using 31,000 subjects. The effects of a major Japanese systematic reading support system, furigana, on the 5^{th} to 8^{th} grade students are analyzed using this environment. Despite our expectation, the 5^{th} to 7^{th} grade students did not seem to enhance their reading with the support of furiganas. The 8^{th} grade students seemed to utilize furigana in reading in vain: their characteristics θs did not increase significantly. We discussed several possibilities why furigana system, which is widely used in Japanese society, does not seem to improve students' reading. Further studies are needed because of the number of samples is small.

Acknowledgement

This research is supported by MEXT/JSPS KAKENHI Grant Number JP 16H01819.

References

Arai, N. H., Todo, N., Arai, T., Bunji, K., Sugawara, S., Inuzuka, M., Matsuzaki, T. & Ozaki, K. (2017). Reading Skill Test to Diagnose Basic Language Skills in Comparison to Machines. *Proceedings of the 39th Annual Cognitive Science Society Meeting*, 1556-1561.

DeStefano, D. & LeFevre, J.-A. (2007). Cognitive load in hypertext reading: A review. *Computers in Human Behavior,* 23, 1616–1641.

Dyson, M.C. & Haselgrove, M. (2001). The influence of reading speed and line length on the effectiveness of reading from a screen. *International Journal of Human-Computer Studies*, 54, 585-612.

Eveland, W. P., Cortese, J., Park, H. & Dunwoody, S. (2004). How Web site organization influences free recall, factual knowledge, and knowledge structure density. *Human Communication Research*, 30(2), 208–233.

Gasser,B., Boeke, J., Haffernan, M. & Tan, R. (2005). The influence of font type on information recall. *North American Journal of Psychology*, 7(2), 181-188.

Kimura, Y. (1984). Concurrent vocal interference: Its effects on kana and kanji. *Quarterly Journal of Experimental Psychology: Section A. Human Experimental Psychology*, 36, 117-131.

Lin, D. M. (2004). Evaluating older adults' retention in hypertext perusal: Impacts of presentation media as a function of text topology. *Computers in Human Behavior*, 20(4), 491–503.

van der Linden, W. J. (2010). *Elements of adaptive testing*. C. A. Glas (Ed.). New York, NY: Springer.

McDonald, S. & Stevenson, R. J. (1996). Disorientation in hypertext: the effects of three text structures on navigation performance. *Applied Ergonomics*, 27(1), 61–68.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Plass, J. L., Chun, D. M., Mayer, R. E. & Leutner, D. (2003). Cognitive load in reading a foreign language textwith multimedia aids and the influence of verbal and spatial abilities. *Computers in Human Behavior*, 19(2), 221–243.

Roethlein, B. E. (1912). The Relative Legibility of Different Faces of Printing Types. *The American Journal of Psychology*, 23 (1), 1-36.

Sakuma, N., Sasanuma, S., Tatsumi, I.F., & Masaki, S. (1998). Orthography and phonology in reading Japanese kanji words: Evidence from the semantic decision task with homophones. *Memory & Cognition,* 26(1), 75-87.

Tinker, M. (1963). *Legibility of Print*. Iowa State Univ. Press.

Wydell, T. N., Patterson, K. E., & Humphreys, G. W. (1993). Phonologically mediated access to meaning for Kanji: Is a rows still a rose in Japanese Kanji? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(3), 491-514.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, 9(2), 79-94.