Appraising Science Textbooks through Quantitative Text Analysis and Psychometric Results of Students' Reading Skills

Teiko Arai (arai-teiko@g.ecc.u-tokyo.ac.jp),
Interfaculty Initiative in Information Studies, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Hidenao Iwane (iwane@rstest.co.jp),
Research & IT Division, Reading Skill Test, Inc., 2-1-4 Shinkawa, Chuo-ku, Tokyo 104-0033, Japan
Takuya Matsuzaki *1, *2 (matuzaki@rs.tus.ac.jp)
*1 Department of Applied Mathematics, Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan
*2 JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

Abstract

The "primary-secondary learning gap" has long been discussed in Japan. Many students suddenly have difficulties in understanding subjects when they enter junior high school (7th grade in Japan). Despite the fact that textbooks are one of the most important learning instruments, the qualitative and quantitative change in the content of textbooks has not been examined in light of the primary-secondary learning gap. In this paper, we show that students are overloaded with the steep increase in the definitions of scientific concepts in textbooks. While the number of definition expressions in textbooks increases rapidly toward junior high school, students' skills in understanding definitions develop only gradually. We demonstrated this through a quantitative linguistic analysis of textbooks and psychometric results of students' reading skills.

Keywords: Reading Comprehension; Textbook; Definition

Introduction

Modern science textbooks usually begin by defining a set of concepts on which a theory is built. This reflects the fact that science is an activity of capturing the infinite diversity of the real world with a finite system of signs, which includes both natural language and mathematics. The defined concepts serve as the contact points between the system of signs and the objects and phenomena in the real world. Thus, we can say definitions are the entrance point to scientific knowledge. From a pedagogical point of view, a fundamental question is, "Can students read and understand definitions in textbooks?"

Numerous studies in the field of cognitive and developmental psychology have explored how children's reading comprehension develops and what skills are involved. With the exception of some text comprehension studies (e.g., Bransford & Johnson, 1972), the most common material in these studies has been the narrative text. Understanding definitions apparently involves a different skill than recognizing a state-of-affairs described in a text. To understand a definition means to be able to *apply* a definition to an object or phenomenon to tell whether it is in the scope of that definition. A recent large-scale study on students' reading skills suggests that reading definitions is much more difficult than other types of language processing (Arai, Todo, Arai, Bunji, Sugawara, Inuzuka, Matsuzaki, & Ozaki, 2017).

A textbook should be written so that a student can learn on his/her own even if it is mainly used as a material in oral lectures. However, we argue that students are overloaded with many definitions in textbooks. That is, there is a gap between the increase rate of the number of definitions in textbooks and the development rate of their reading skills.

Read the following text.

Sugar water is made by mixing sugar with water. Like sugar in sugar water, the substance which is dissolved in a liquid is called solute, and, like water in sugar water, the liquid in which solute is dissolved is called solvent. The mixture of a solute and a solvent is called a solution. When the solvent is water, the mixture is called an aqueous solution.

Select all items that are examples of a solute:

- ☐ Ice in iced water
- ☐ Water in iced water
- ☐ Salt in salted water
- ☐ Water in salted water

Grade	6th	7th	8th	9th
%Correct Answers	12.7%	16.4%	18.8%	23.1%

Figure 1: Instantiation question and percentage of correct answer (Correct answer: "Salt in salted water")

We demonstrated this through a quantitative text analysis of school textbooks and psychometric results of students' reading skills. We first analyzed one of the most widely used science textbooks for 5th through 8th grades in Japan focusing on definition expressions. We then investigated the basic reading skills of students by using the results of a language skill test developed by Arai et al. (2017), which includes a type of question called instantiation, wherein a test taker must choose all proper examples of a scientific or mathematical concept described in a question. Figure 1 presents an example of an instantiation question, which is based on a text taken from a science textbook for 7th grade. The topic of an instantiation question includes both those that are taught in school (e.g., solute and solvent) and those usually not (e.g., sexy primes). Note that it is not suitable to evaluate students' skills in reading textbooks based on the scores of an achievement test because problem solving is different from reading text to acquire a concept.

Previous researches on textbooks have mainly focused on

what is written in textbook, how it is expressed, and how textbooks are produced. Pedagogy have traditionally put emphasis on historical studies of institutional issues and historical change of contents (e.g., Kaigo, Naka, & Terazaki, 2017). Whereas in the interdisciplinary field of education and cognitive science there have been studies on the means of expression, for instance, specially designed fonts and universal designs (UD) for learners with *a specific cognitive tendency*, such as dyslexia and color blindness (Zhu & Kageura, 2019; Pisha & Coyne, 2001). Our research belongs to these cognitive science studies in the sense of evaluating textbooks based on psychological results but at the same time the present study delved deeper into the content of text. Besides, this paper is an attempt to evaluate textbooks for *all* students with no special attention to a certain cognitive tendency.

The rest of the paper is organized as follows. Section 2 provides an overview of the Japanese textbook-system and explanation of the primary-secondary learning gap. An overview of Arai et al.'s language skill test is also provided. Section 3 summarizes our hypotheses, Section 4 describes the methods and results, Section 5 provides a discussion of the results, and Section 6 concludes the paper.

Background

We explain the Japanese administrative system concerning textbooks used in compulsory education and introduce a known issue in the Japanese education system called "primary-secondary learning gap." We then describe a large-scale experimental study on the reading skills of Japanese students, on the basis of which we argue there is a discrepancy between their reading skills and textbook content.

Textbooks in Japanese School

In Japan, textbooks used in compulsory education are created by nongovernmental publishers and then submitted for official examination by the Ministry of Education. The contents must follow a national curriculum standard, but the publishers may include their own learning method and ideas in the material. The final decision on which textbooks to use rests with each school.

Primary-Secondary Learning Gap and Textbooks

The primary-secondary learning gap has long been pointed out as a serious problem in the Japanese education system. Many students suddenly have difficulties in understanding subjects when they enter junior high school (7^{th} grade in Japan). In the Japanese school system, 5^{th} and 6^{th} grades are the last two years of primary school, and junior high school is from 7^{th} through 9^{th} grades. People have given different explanations to this phenomenon. For instance, sudden changes in the style of teaching and increased difficulties in subject content have been suspected (Itou, 2013).

Most students do not know the content of the subject ahead of learning. Note that we can not understand the content of knowledge directly, but only through language or symbolic expressions: there is no substance such as "knowledge" itself. Nevertheless, the qualitative and quantitative change in the expressions of the textbooks have not yet been examined in light of the primary-secondary learning gap.

Reading Skill Test

Arai et al. (2017) developed a test for assessing the reading skills of children and adults called the Reading Skill Test (RST). The RST is a computer-based test and all the questions are in multiple-choice style. The skills of the test-takers are assessed by applying item response theory (IRT; Lord & Novick, 1968; Hambleton & Swaminathan, 1985). In IRT, the chance of a subject i correctly answering question j is modeled by a function on the subject's ability θ_i and difficulty of the question b_j.

The RST consists of seven question types: dependency (DEP), anaphora (ANA), paraphrase (PARA), inference (INF), representation (REP), instantiation-dictionary (INSTd), and instantiation-mathematics-and-science (INSTm). DEP and ANA questions test the ability of recognizing the syntactic and semantic structure of a sentence. PARA and INF questions test the ability of recognizing the logical relation between two text segments. REP questions test the ability of recognizing the relation between a text and schematic representation (diagrams and illustrations). INSTd and INSTm questions test the ability of recognizing the relation between a definition and its examples. Specifically, in INSTd questions, a definition of a word in a dictionary is presented and the test takers must choose its proper usages (example sentences) from the options. In INSTm questions, a definition of a mathematical or scientific concept is presented and the test takers must choose concrete examples of the concept.

Most of the texts used in the RST are taken from textbooks approved by the Ministry of Education and are being used in Japanese junior high and high schools. Other texts are taken from newspapers and dictionaries.

More than 120,000 people have taken the RST since 2014. The reliability and one-dimensionality of the test was shown by statistical analyses of the results (Arai et al., 2017). The test takers' ability parameters θ assessed from the RST, when averaged within the school they belong to, were shown to correlate with their socioeconomic status (Arai et al., 2017) and level of their scholastic ability (Arai, Bunji, Todo, Arai, & Matsuzaki, 2018).

In the current study, we used the data collected by Arai et al. to estimate the development rate of the skills in understanding definitions of scientific concepts and the speed of reading them. By combining these two quantities, i.e., the accuracy and speed of reading definitions, we demonstrated that the difficulty of commonly used science textbooks, measured in terms of the number of definitions, is far above the expected skill level of junior high school students.

Hypothesis

Two hypotheses are examined in the following section. First, we hypothesize that it is more difficult to understand the meaning of a sentence that defines a mathematical or

Table 1: Number of participants of RST by grade

Grade	Number of Participants
≤ 5	783
6	15,035
7	23,558
8	21,625
9	16,078
≥ 10	49,661

Table 2: Basic statistics on science textbooks

Grade	# of sentences	# of word tokens	# of word types
P5	1,067	16,962	1,621
P6	1,214	20,092	1,782
S1	1,168	23,904	1,989
S2	1,292	27,486	2,069

scientific concept than to recognize the syntactic and semantic structure of a sentence that describes some state-of-affairs. It is examined by comparing the difficulty parameter of the INSTm questions of the RST with those of the other question types.

Second, we hypothesize that the development of the skill in reading definitions is not as fast as necessitated for comprehending science textbooks used in Japanese junior high schools. We verify this in three steps. First, we examine the number of definition sentences in elementary and junior high school science textbooks. Second, we quantify the development of the skill in understanding definitions, measured from the accuracy and speed of the responses to the INSTm questions of the RST. Finally, we compare the rate of increase in the number of definitions in the textbooks with the expected rate of the development of the skill in understanding definitions to show the former is far more steep than the latter; this is quantitative evidence of the primary-secondary learning gap.

Methods and Results

Difficulty of Definition Understanding

We applied the commonly used three parameter IRT model (3PL IRT) to the results of the RST. The number of participants is shown in Table 1. The 3PL IRT model involves the random guessing parameter c_j for each question j in addition to the participants' ability parameters θ_i and difficulty parameters b_j. The c_j parameter measures the inessential difference in the difficulty of the questions mainly due to the multiple-choice style (i.e., the effect of random guess among the choices by the participants having low reading skill). We can thus compare the difficulty of the different types of RST questions by comparing the b_j obtained from 3PL IRT despite the fact that the number of

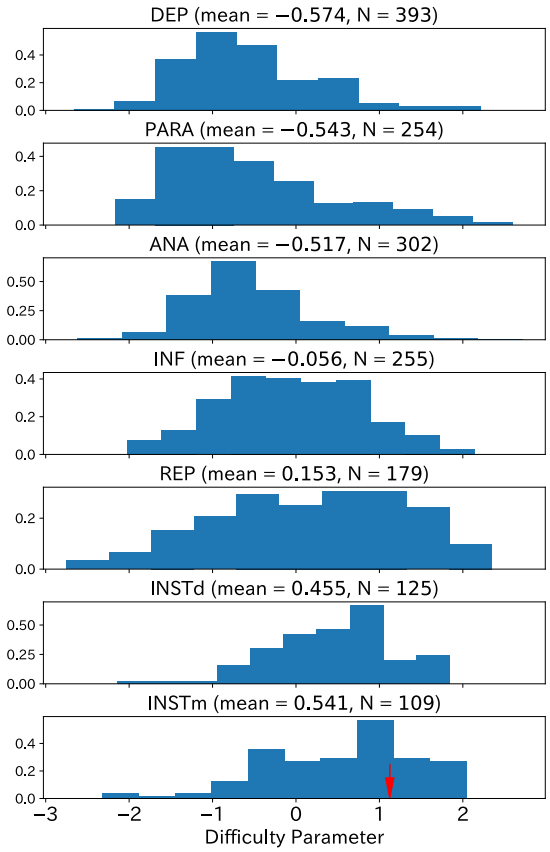

Figure 2: Distribution of difficulty parameters. Red arrow in plot for INSTm indicates difficulty of example question presented in Figure 1 ($b = 1.124$).

options (two to four) and style of multiple-choice (single or multiple selection) differ across the questions.

Figure 2 presents the distribution of b_j of each question type. It reveals that INSTm questions are the most difficult among the seven question types. The shape of the distributions as well as the mean of b_j are quite different between the questions assessing basic syntactic and semantic processing skills (DEP, PARA, and ANA) and the two question types

Table 3: Example of definitions

Grade	Example
P5	A liquid in which something is dissolved is called **a solution**.
P6	**Breathing** is to take in oxygen and expel carbon dioxide.
S1	The difference in the arrival time of the primary and secondary wave is called **the duration of preliminary tremors.**
S2	**Oxidation** is a chemical reaction that occurs when a substance reacts with oxygen to form an oxide, whereas **reduction** is a chemical reaction in which oxygen is removed from an oxide.

Table 4: Number of definition expressions by subject

	Grade			
	P5	P6	S1	S2
Physics	2	6	37	36
Chemistry	2	8	42	39
Biology	9	11	23	52
Earth science	7	10	47	45
Appendix	2	2	3	2
Total	22	37	152	174

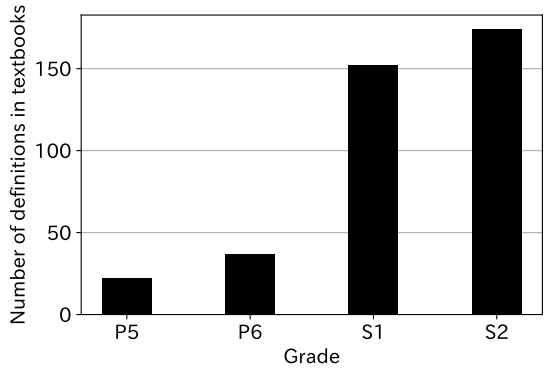

Figure 3: Number of definitions in science textbooks

assessing the skill in understanding definitions (INSTd and INSTm).

Analysis of Science Textbooks

We examined one of the most commonly used science textbooks for 5th through 8th grades in Japan, i.e., for the fifth and sixth years of primary school and the first and the second years of secondary school, as we are concerned with the primary-secondary learning gap. We used P5, P6, S1, and S2 to refer to the textbooks of 5th, 6th, 7th, and 8th grades, respectively. All the textbooks we examined were published by Tokyo Shoseki Co., Ltd (Mouri & Kuroda, 2016a, 2016b; Okamura & Fujishima, 2016a, 2016b) and used in approximately 35% of public schools in Japan.

Table 2 shows the number of sentences, word tokens, and word types in the textbooks. We counted the sentences and sentential parts in the main body of the texts and the chapter and section titles. Sentences were decomposed into words by using the Japanese morphological analyzer MeCab (Kudo, Yamamoto, & Matsumoto, 2004).

We searched definition expressions from each textbook and imparted annotations on them. Table 3 lists several examples of the definitions. The annotation policy was as follows:

- We counted only definitions that are given completely in one sentence; there are few cases where a concept is defined with multiple sentences or a paragraph.

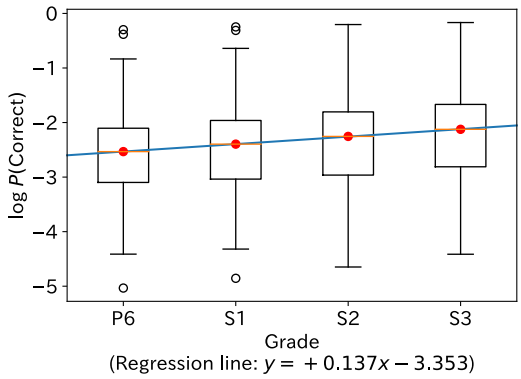

Figure 4: Distribution of expected rate of correct answers to INSTm questions by students in each grade

- When more than one concepts are defined in one sentence, we annotated them separately (e.g., we consider there are two definitions in S2 in Table 3).

Most concepts were defined with expressions that explicitly introduce a new concept such as "... is called ..." and "... is defined as ...", and the names of the defined concepts were often written in bold face font.

We then counted the number of definition expressions by grade and subject. Table 4 summarizes the results. There is a noticeable gap between primary and junior high school in the number of definition expressions. Figure 3 presents the plot of the total number of definitions for each grade.

Quantifying Development of Skill of Definition Understanding

Development of $\boldsymbol{\theta_i}$ Figure 5 presents the boxplots of $\boldsymbol{\theta_i}$ of the test-takers in each grade (6th to 9th grades) and for five question types including INSTm. The other two question types (PARA and INF) exhibited similar trends. It also presents the regression lines of $\boldsymbol{y}$ on $\boldsymbol{x}$ where $\boldsymbol{y}$ is the mean of $\boldsymbol{\theta_i}$ over the participants in a grade and $\boldsymbol{x}$ is the grade. The equation of the regression line is shown in the headings. The similar slopes of the regression lines (between 0.169 and 0.178) indicate that, on average, the seven types of abilities develop similarly from 6th to 9th grades.

Figure 4 presents the boxplot of the expected rate of correct answers in each grade (in logarithmic scale) for the 109 INSTm questions, obtained by substituting the mean of the θ of each grade for the θ of 3PL IRT. The regression line is of the mean of the log P(Correct response) on the grade, which means the expected rate of correct responses on typical INSTm questions with a participant having average ability. The slope of the regression line indicates that the expected rate of correct answers increases by approximately $\exp(0.137) \approx 1.15$ times each year.

Increase in Reading Speed We analyzed the response time to the RST questions to estimate the increase rate of reading speed. At least two factors are involved in reading speed: the characteristics of the text (e.g., length and difficulty) and those of the reader. To reduce the effect of the variation in the questions' characteristics, we used a linear

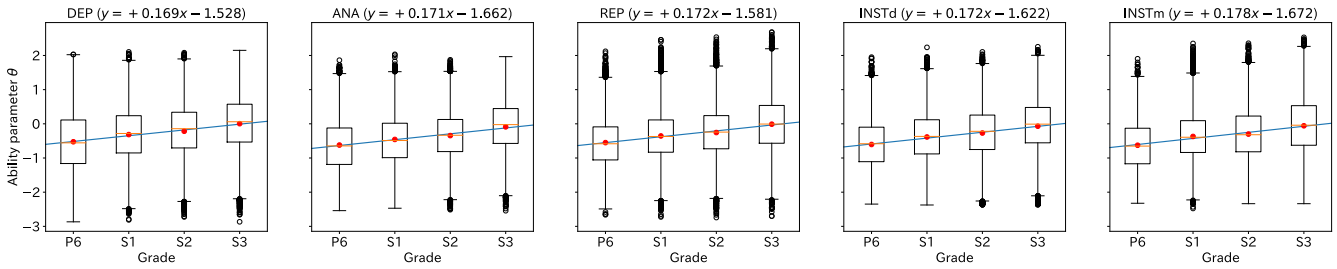

Figure 5: Distribution of ability parameters

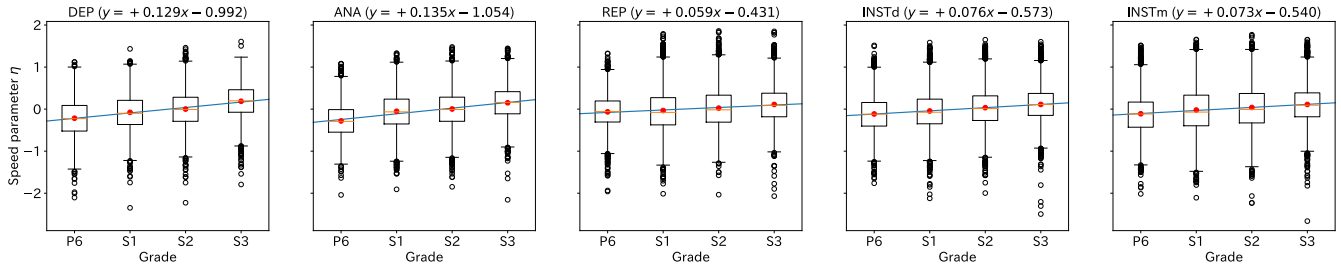

Figure 6: Distribution of reading speed parameters

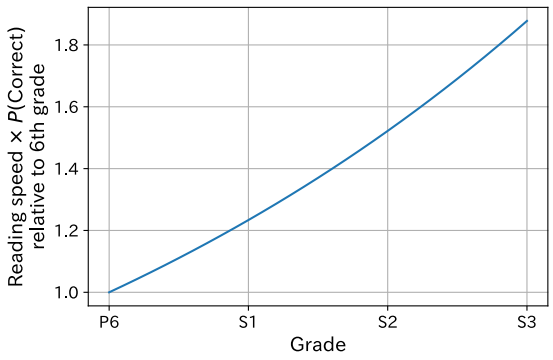

Figure 7: Expected throughput of definition reading

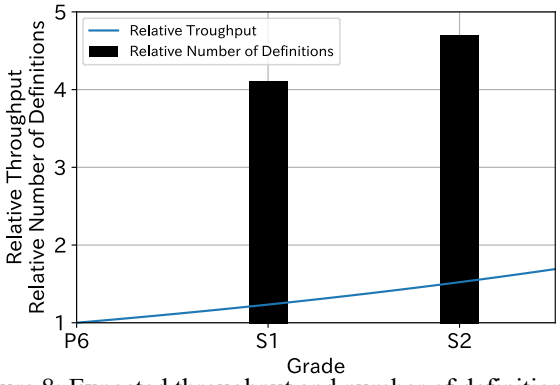

Figure 8: Expected throughput and number of definitions in textbooks relative to those of 6th grade

model of the logarithm of the response time, in which the above two factors are included as two sorts of independent variables:

$$\log t_{ij} = \beta_j - \eta_i + \epsilon_{ij}$$

where $\boldsymbol{t_{ij}}$ is the response time on question $\boldsymbol{j}$ by participant $\boldsymbol{i}$, $\boldsymbol{\beta_j}$ is the question-dependent parameter that represents the logarithm of the response time to question $\boldsymbol{j}$ by a participant with average reading speed, $\boldsymbol{\eta_i}$ represents the average reading speed of participant $\boldsymbol{i}$, and $\boldsymbol{\epsilon_{ij}}$ is the error term that is assumed to be normally distributed. Parameters $\boldsymbol{\beta_j}$ and $\boldsymbol{\eta_i}$ were estimated by least square fitting, assuming that the mean of $\boldsymbol{\eta_i}$ is zero (to avoid colinearity). Similar models of response time have been proposed in the psychometric literature (e.g., van der Linden, 2007). Although it is often assumed that $\boldsymbol{\eta_i}$ correlates with $\boldsymbol{\theta_i}$, no such correlation was observed from the results of the RST.

Figure 6 presents the distribution of $\boldsymbol{\eta_i}$ for the participants in each grade. The regression line is of the mean of $\boldsymbol{\eta_i}$ on the grade. We can see that the reading speed on every question type increases by grade but at a different rate. Specifically, on DEP and ANA questions, the reading speed increases by approximately $\boldsymbol{\exp(0.13) \approx 1.14}$ times by grade, which means an average 9th grader answers to the DEP and ANA questions $\boldsymbol{1.14^3 = 1.48}$ times faster than an average 6th grader. Meanwhile, on INSTm questions, the rate of increase is $\boldsymbol{\exp(0.073) \approx 1.08}$ by grade, which amounts to the factor of only $\boldsymbol{1.08^3 = 1.26}$ between the reading speed of an average 6th and 9th grader.

Development of Throughput of Definition Understanding

We define the "throughput" of definition understanding as the expected number of definitions correctly understood per unit of time by a student. We can estimate this by multiplying the accuracy of definition understanding and the reading speed of them. Specifically, by adding the growth rate of $\boldsymbol{\log P(\text{Correct response})}$ (Figure 4) and that of the logarithm of reading speed (Figure 6) for the INSTm questions, we obtained the expected increase rate of throughput $\boldsymbol{\exp(0.137 + 0.073) \approx 1.23}$ by grade. This means, for example, an average 7th grader can understand

1.23 times more definitions than an average 6th grader in the same amount of time and at the same level of understanding. Figure 7 presents the curve of the expected throughput at each grade relative to the 6th grade.

Although this throughput is, of course, a crude estimate of the true ability of the students' science learning, it enables us to quantitatively compare the development of their ability against the increasing complexity of textbooks measured by the number of definitions in them. We accomplish this by comparing both quantities relative to those in a grade (e.g., 6th grade) as the baseline. If the relative increase in the number of definitions in the textbooks is significantly faster than the development of throughput, there would be a high risk for the students not to be able to keep up with the pace of their science course. Unfortunately, this is the case shown in Figure 8, which presents the development curve of throughput (Figure 7) overlaid with the number of definitions in the textbooks (Figure 3).

Of course, a textbook is not necessarily written so that a reader can understand it by reading it only once. It may thus appear too simplistic to directly compare the throughput with the number of definitions. However, it does not affect our conclusion much unless the chance of having complete understanding of a concept is superlinear (i.e., more than proportional) to the number of times one reads the definition. The low average score for the INSTm questions suggests such a superlinear effect is unlikely because takers of the RST can read the definitions as many times as they like.

Discussion

School hours are given to every student equally regardless of their throughput of understanding. Highly skilled students can understand all definitions given in the classroom. In contrast, there would be some unintelligible definitions for other students. As scientific knowledge is organized hierarchically, if such unintelligible definitions accumulate, the risk of dropout from the course would sharply increase.

One can compensate for the increasing gap between the ability of learning and the difficulty of textbook by spending more time studying. However, a national survey (Japanese Ministry of Education, 2017) showed that Japanese junior high school students spend only slightly more time studying than elementary school students, which is apparently not enough for filling the gap exhibited in Figure 8.

Different amendments to the current situation are conceivable. First, we could make the curriculum less crowded. Second, we could teach using various methods besides through text. For instance, movies and other digital teaching materials might be good complements to the textbooks. Specialized lessons for reading definitions, such as exercise of enumerating and checking the conditions in a definition expression, might also be effective.

Conclusion

We verified two hypotheses:

1) It is more difficult to understand the meaning of a sentence that defines a mathematical or scientific concept than to recognize the syntactic and semantic structure of a sentence.
2) The development of the skill of reading definitions is not as fast as necessitated for comprehending science textbooks used in Japanese junior high schools.

Hypothesis 1) was verified by the result that INSTm questions are the most difficult among the seven question types of the RST. Hypothesis 2) was examined by the gap between the growth rate of the number of definitions in textbooks and the throughput of students. We suspect this discrepancy is one of the major reasons for the primary-secondary learning gap in science and mathematics.

The number of definition expressions in textbooks increases rapidly toward junior high school. In 7th grade, the number of definition expressions suddenly becomes four times more than that of the previous year. Some countermeasures such as special training curriculum for reading definitions are clearly needed.

Acknowledgement

This research is supported by MEXT/JSPS KAKENHI. Grant Number JP 16H01819 and JST, PRESTO Grant Number JPMJPR175A.

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*, (pp. 1556-1561).

Arai, T., Bunji, K., Todo, N., Arai, N.H., & Matsuzaki, T. (2018). Evaluating Reading Support Systems through Reading Skill Test. *Proceedings of the 40th Annual Cognitive Science Society Meeting*, (pp. 100-105).

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding. *Journal of Verbal Learning & Verbal Behavior*, 11, (pp. 717-726).

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhof.

Itou, J. (2013). Study on collaboration between elementary and junior high schools to overcome primary-junior high school gap: Focusing on the view from junior high school. *Yamagata University Graduate School of Practice Annual Report,* 4, (pp.268-271), (in Japanese).

Japanese Ministry of Education. (2019). National Assessment of Academic Ability and Learning Activities. http://www.nier.go.jp/19chousakekkahoukoku/index.html

Kaigo, S., Naka, A., & Terazaki, M. (2017). *Modern Japanese education in textbooks.* Tokyo Shoseki. (in Japanese)

Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. *Proceedings of the 2004 Conference on*

Empirical Methods in Natural Language Processing (pp. 230-237).
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287-308.
Mouri, M., & Kuroda, R. (ed.) (2016a). *New Science: 5th grade.* Tokyo Shoseki. (in Japanese)
Mouri, M., & Kuroda, R. (ed.) (2016b). *New Science: 6th grade.* Tokyo Shoseki. (in Japanese)
Okamura, S., & Fujishima, A. (ed.) (2016a). *New Science: 7th grade.* Tokyo Shoseki. (in Japanese)
Okamura, S., & Fujishima, A. (ed.) (2016b). *New Science: 8th grade.* Tokyo Shoseki. (in Japanese)
Pisha, B., & Coyne, P. (2001). Smart from the start: The promise of Universal Design for Learning. *Remedial and special education*, 22(4), 197-203.
Zhu, X., & Kageura, K. (2019). Research on Japanese Typefaces and Typeface Customisation System Designed for Readers with Developmental Dyslexia. *International Association of Societies of Design Research Conference*.