NIRA OPINION PAPER
No.31 | July 2017

Cultivating Human Resources for the Era of Digitalization

Noriko Arai
Professor, Information and Society Research Division, National Institute of Informatics

Koken Ozaki
Associate Professor, Graduate School of Business Sciences, University of Tsukuba

The 21st century will be a period of enormous transformation in the technology known as digitalization, and this will have a huge impact on human work. All information around us is being transformed into data which can be read and processed by machines, and already jobs exclusively involving the transmission of data and jobs characterized by routine and standardized procedures are being taken over by machines. In such an era, what types of skills should human workers attempt to acquire?

The results of research conducted in the AI project "Can a Robot get into The University of Tokyo?," commenced in 2011, led to the conclusion that AI had difficulty in understanding the meaning of words and judging situations. This is to say that the role of humans will not disappear in jobs involving problem-solving, jobs which demand understanding of contexts and judgment of situations. In jobs of this type, humans capable of appropriate communication are able to display greater ability than machines. However, reading skills tests conducted by the authors indicated that many students have not mastered the ability to read with an understanding of the meaning of words and contexts, and are able to read only in a superficial manner similar to that of an AI. Based on these results, the authors suggest that there is value in considering a shift in Japanese language education towards logical activities which cultivate reading ability.

Section 1

The Difference between Robots and People

1. The Technological Revolution of Digitalization

What is Digitalization?

In the 21st century, we are faced directly with a major technological revolution known as digitalization. "Digitalization" refers to the distribution of all information – information which was formerly distributed in an unencoded form – in a format which enables it to be read and processed by computers. While an analogue text might be scanned and input to a computer, this would not represent digitalization if the machine was unable to read the scan as a digital text. The Internet of Things (IoT), which has attracted considerable attention in recent years, is a technology which seeks, by attaching sensors to a variety of objects, to convert events, situations, statuses, and the changes in these into data able to be analyzed by a computer. Another focus of attention at present, AI technologies should also actually be viewed as technologies seeking to enable machines to "understand" things by means of data from sensors attached to objects, rather than technologies which look toward the creation of machines able to "think" like humans. Put another way, digitalization is an essential infrastructure that will enable information processing which has previously been conducted person-to-person to be taken over by machines.

What Types of Labor will be taken over by Machines?

Digitalization will have a huge impact on human labor. It can be predicted that jobs exclusively involving the transmission of data will be the first to be taken over by machines. We may take systems for dealing with forms at banks or government offices as an example. If the practice of having people fill in forms by hand is done away with, and information is input digitally from the very beginning, the labor involved in the simple information transmission procedure of transcribing the information on hand-written forms will become unnecessary. Routine and standardized processing will also be taken over by machines. "Routine and standardized" here means "routine and standardized from the perspective of machines." If the form of input and output is mathematically determined and the type of processing which should be applied when the input satisfies specific conditions can be described mathematically, this type of processing can be taken over by machines. For example, in the case of the digitalization of medical patients' "drug history handbooks," the issuing of alerts when a drug is over-administered or when it is necessary to exercise caution in taking drugs together would represent model types of routine and standardized processing.

The development in recent years of the methodology known as "machine learning," which models human judgment statistically, has broadened the scope of the procedures that can be undertaken by machines. There are numerous procedures in which the information obtained as

input is standardized and the judgment to be output has only two values (Yes/No), such as, for example, the inspection of power lines or the diagnosis of tuberculosis via an X-ray photograph. Further along the continuum are procedures in which, rather than two values, input can be classified into multiple types. Judgment in the case of identifying an object in a photograph, or distinguishing the words that handwritten words represent or the words that a voice is speaking, are representative examples. In this matter of classification decision, as it is termed, if provided with a sufficiently high volume of input data and data concerning correct and incorrect responses to that input, machines often perform with greater accuracy than humans. For example, machines achieve greater accuracy than humans in cases such as ascertaining the genuineness of works by Van Gogh and diagnosing illnesses that are infrequently encountered[1].

When AI make classification decisions, a technology known as "ranking" is important. This can be considered using the example of mechanizing the process of referring to a FAQ table and returning the most appropriate response to an inquiry to a call center. This would necessitate measurement of degrees of difference in meaning in order to determine which of the questions on the FAQ sheet the customer's inquiry most resembles. In order for a machine to calculate these degrees of difference, it would be necessary for them to be able to be defined mathematically. However, we do not possess a methodology for treating meaning in the real world mathematically, and it is not possible to directly measure closeness in meaning. An alternative strategy would therefore be to measure the superficial closeness of the sentences (strings of words) uttered by the customer to the questions on the FAQ sheet. In many cases, it would be possible to express strings of letters or strings of words as vectors, and for a machine to learn which of the questions on a FAQ sheet they were closest to from past data, enabling it to output a ranked order of differences (or closeness) in meaning. If the probability of the question positioned at number one in this ranking being the correct answer was 99%, the machine could be expected to provide the correct answer to 99 callers out of 100. However, if the probability was only 85%, there would be cause for concern over customer satisfaction. The next-best approach that has been suggested based on this consideration is to introduce AI as tools to support judgment regarding classification. In this case, multiple top-ranking choices would be displayed on a screen, allowing the call center employee to respond to the inquiry. This would eliminate the necessity to memorize 1000s of FAQs and reduce the labor involved in looking up FAQs. If the probability of the correct answer being in the top five choices was 99%, the technology would function extremely well as a support system; it has been indicated that the use of such a system could increase the efficiency of call center procedures by 30%. In fact, technologies of this type are already being introduced in a variety of scenarios.

Work will not disappear from Human Society

The point to bear in mind here is that in terms of listening to a customer and coming up with a solution to their problem, consulting with a call center and a handyman, which to humans seem similar, involve entirely different procedures for a machine. In the former case, understanding of

[1] At present, research in AI technology is seeking to determine what level of diversity of data and what volume of data are necessary in relation to specific types of input (image, voice, natural languages, numerical data, etc.) to control the incidence of errors to the same level as human error.

meaning is not required, and the procedure is no more than the solution of a classification problem. The latter case involves problem-solving, and demands genuine understanding of meaning and judgment of the situation. While the introduction of machines will optimize procedures in the former case towards the realization of zero cost, the added value represented by humans performing jobs which require flexibility in grasping meanings, as in the latter case, is likely to continue to increase.

There are also situations in which it is difficult to determine how much data are necessary for a decision, what the results of a decision will be, and whether or not a decision was correct, such as management decisions, the provision of nursing care or caring for a child, the establishment of a venture company or the provision of consultation regarding regional revitalization. In addition, it will be impossible to prevent machines from making errors in cases for which there are insufficient similar data. In cases in which humans are able to naturally arrive at the correct solution based on alternative information (context, etc.) when there are insufficient past data, machines, relying on statistical methods, will remain unable to respond flexibly.

As long as such jobs continue to exist, work will not disappear from human society. The value of human resources who possess skills enabling them to perform work that cannot be performed by machines will further increase, while human resources not possessing such skills will be forced to cling to low-wage jobs. It is also of concern that disparities will be prone to arise between scarce, high-value jobs and other jobs, and that the latter will be likely to involve long hours.

This provides grounds for an urgent, concrete consideration of education in the era of digitalization. As has been shown above, the skills which will have value in the labor market will change significantly following digitalization. For workers to enter the contemporary labor market, it will be essential for them to equip themselves with skills suited to the era of digitalization through education.

2. Can a Robot get into The University of Tokyo?

The "Todai Robot": Helping to determine the Ideal Direction for the Fostering of Human Resources

The Ministry of Education, Culture, Sports, Science and Technology's new curriculum guidelines, set to go into effect from fiscal 2020, were published in March 2017. The guidelines specify how to educate children born in the era of digitalization. However, when discussions towards the guidelines commenced in 2011, we had no clear image of the 2020s, when we will be living with AI against a background of digitalization. There was no shared image regarding the extent to which white collar jobs would be replaced and what type of jobs would remain for humans with the unavoidable introduction of AI to society.

Against this background, the project "Can a Robot get into The University of Tokyo?" (nicknamed the "Todai Robot") was commenced in 2011. This project was an unprecedented attempt to compare 500,000 third-year senior high-school students looking towards taking university entrance examinations, who could be considered to be seeking to secure white collar

positions, with current and near-future AI performance. As a result of the traumatic failure of the Fifth-Generation Computer Systems initiative in the 1990s, large-scale AI projects were non-existent in Japan at that time. In part because of this, in Japan, experience-based intuition regarding how much accumulated granular data, and of what type, would produce the conditions for high-accuracy judgment in a machine was lacking, and researchers were sent this way and that by events such as the advent of the Google Car and the victory of IBM's Watson in the quiz show Jeopardy. Responding to this situation, industry and university research institutions came together to clarify the potential and the limitations of near-future AI with university entrance examinations as the benchmark, seeking to accurately delineate what types of business Japan should focus on and how the nation should attempt to foster its human resources. Thus, the "Todai Robot" project was born.

After five years of research, the project concluded that "Given the state of theory and near-future data and technologies that can be projected on this basis, it is not possible to create an AI that is able to understand its interlocutor, accurately judge situations, and solve problems in cooperation with humans." The group of technologies known as "deep learning" are unable to generalize from limited examples like humans, and are unable to deal with abstract concepts such as "meaning" and "technology" which cannot be expressed as binary code or images. This is because, as indicated above, no mathematical framework exists to deal with concepts of this type. Some will counter that we cannot predict the future, but one thing is clear. If, hypothetically, an AI capable of understanding meaning is created in the future, it will not be the result of a revolution in the worlds of AI technology or hardware. It will come as a result of mathematics researchers creating theory able to support this achievement. Without this grounding of theory, the sudden appearance one day of a perfected AI would be a miracle of the type that only occurs in science fiction.

At the same time, the Todai Robot demonstrated greater facility than human examinees in writing plausibly argued short essays and taking fill-in-the-blanks tests, based on rote memorization of textbooks and Wikipedia, and conducting searches and cutting and pasting. The robot's performance did not fall behind that of human examinees in 600-character essay problems on a mock-up University of Tokyo world history test (National Institute of Informatics press release, 2016b). As this indicates, the issue is not whether the robot is faced with essay questions or answer sheets. The issue is whether or not problems necessitating understanding of meaning are presented.

In 2013, an Oxford University research group issued a prediction that the jobs of half of America's workers would be taken over by machines by 2030 (Frey and Osborne, 2013). The jobs that were predicted to disappear in this paper, among them conducting bank and insurance reviews, accounting, and writing sports articles, can be characterized as not necessitating the manipulation of abstract concepts and cooperation with others, and as involving an overwhelming proportion of procedures able to be performed by finding similar examples in past data and following prescribed formats.

Table 1: Results for Todai Robot in fiscal 2016 National Center for University Entrance Examinations Mockup Test

《Overview of results: Fiscal 2016 Shinken Mockup Test – Comprehensive Academic Ability Mockup Test (June)》

	Japanese (200)	Mathematics IA (100)	Mathematics IIB (100)	English (Writing) (200)	English (Listening) (50)	Physics (100)	Japanese History B (100)	World History B (100)	Total for five subjects (950)
National average score	96.8	54.4	46.5	92.9	26.3	45.8	47.3	44.8	437.8
Todai Robot score	96	70	59	95	14	62	52	77	525
Todai Robot T-score	49.7	57.8	55.5	50.5	36.2	59.0	52.9	66.3	57.1

(Note) T-scores aggregated from 120,582 examinees (total number of examinees: 264,604) for eight courses in five subjects (humanities) – Japanese, two Mathematics courses, English (writing and listening), two History courses, and one Science course. The figures in parentheses below the names of the courses are the points allotted to that course.

3. The Importance of the Skill of Understanding Meaning

We must avoid reaching for "Pie in the Sky"

What types of human resources, and what types of skills, will be demanded by the labor market in the era of digitalization? Among the top contenders for replacement by machines as specified by the Oxford University research mentioned above were jobs requiring a good education and qualifications, including loan officer, insurance appraiser, paralegal, accountant, and sports book writer. The fact that the jobs predicted to be significantly affected by digitalization cover a wide range (from the human perspective) invites the misapprehension that AI are omnipotent. But as it happens, all of these jobs are characterized by conditions making it easy for them to be replaced by AI, as discussed in Section 1 above.

Because AI are dependent on theory and mathematical languages such as probability and statistics, they would be powerless if applied to work that cannot be processed using mathematical languages, encompassing 1) Non-standardized work in which input and output is not fixed; 2) Work in which it is difficult to fix standards for the judgment of correct or incorrect answers; 3) Work which necessitates correct judgment based on learning from an extremely limited range of examples; and 4) Work which creates businesses which have not previously existed. Given this, we can consider that the demand for labor power to perform work of these types will not decrease but will rather increase.

However, there is a pitfall here. Work which necessitates correct judgment based on learning from an extremely limited range of examples and work which creates businesses which have not previously existed (3) and 4) above) represent sophisticated and creative types of work. There has been a common awareness throughout society since the 20th century that this type of work is particularly rare. If it was possible to develop these abilities via a standardized educational

program, no doubt it would already be being done. “Active learning,” which is receiving so much attention today, is very similar to the form of education known as “Seikatsu Tangen Gakushu,” introduced in Japan during the occupation following the Second World War. Seikatsu Tangen Gakushu emphasized investigation and discussion, and the solving of problems corresponding to real-world phenomena. However, this type of learning sparked the criticism that it was causing a decline in academic ability, and in the 1960s, government academic guidelines were revised to stress systematic education. The feeling among educators was that the ideal of Seikatsu Tangen Gakushu was pie in the sky, and that while already able children might increase in ability, children who were not would be unable to acquire skills under this system. For example, one might ask students to run 100 meters in less than 10 seconds, but the chances of them developing this skill are entirely based in probability. If the targets of an educational system were nevertheless set on this basis, the ultimate result would be enormous educational costs for almost no return. We must learn the lesson of Seikatsu Tangen Gakushu, and not repeat the same mistakes.

In the case of non-standardized work in which input and output is not fixed and work in which it is difficult to fix standards for the judgment of correct or incorrect answers, 1) and 2) above, if they are able to judge situations and understand meaning, and are able communicate appropriately via words and gestures, it is certainly not difficult for humans to display superior abilities to machines. In concrete terms, if everyone is able to acquire the basic skills that are the essential aim of elementary and high school education – the ability to look, read, hear, write, and speak well – there is no need to fear the AI era, because AI are not able to understand meaning, and are not able to use words in a true sense.

The Importance of the Ability to read with an understanding of Meaning

With the development of an advanced information society connected by the Internet, in which a huge amount of information has been digitalized, a transition is occurring in which the majority of human communication will no longer be face-to-face, but will be conducted via the exchange of email and other documents. And because the skills demanded by the advancement of digitalization are changing rapidly, it is necessary for workers to continue to absorb new knowledge. However, because the transmission of knowledge by means of the traditional student/teacher or apprenticeship method cannot keep up with the pace of change, many companies are requiring their employees to undertake independent study, for example via e-learning. When starting a company, it is necessary to learn about matters such as how to read a contract and legal compliance from documents obtained from the Internet. As this indicates, rather than being taught in a class, or by a member of staff who fulfills the role of teacher in the relevant department, workers will unavoidably be required to learn independently, from documents. The absolute requirement in these cases is the ability to read while understanding the meaning of the text. In a digitalized society, as we cooperate with machines in tasks corresponding to categories 1) and 2) above, the ability to look, read, hear, write, and speak well will be essential to distinguishing humans from machines.

We tend to think that anyone who has received a standard high school education, and anyone who has proceeded to higher education, will possess these skills. In a country like Japan, which

achieves excellent results in OECD surveys of academic ability (PISA and TIMMS), this tends to be taken for granted[2].

However, there is a considerable gulf between this expectation and the reality. In 2011, the authors, in collaboration with the Mathematical Society of Japan, conducted a survey involving more than 5,000 university students, in order to determine their level of understanding of mathematics taught up to the first year of senior high school, in particular basic items (Mathematical Society of Japan Education Committee, 2013). The results of this survey demonstrated a number of problems. For example, only three-quarters of the survey examinees correctly understood the meaning of "average." While almost 100% of the students knew that "The average can be determined by adding together all the figures and dividing by the number of figures" (in this case the number of humans), and were able to perform the procedure, they did not understand what conclusions could be derived from the average. Again, a result which attracted particular attention was that less than 20% of the students were able to correctly answer the question "Why, when odd and even numbers are added together, is the result always odd?" (This figure increased to 34% when semi-correct answers were included). The correct answer is "If the even number is expressed as 2n and the odd number as 2m+1, the sum of these numbers is 2n+2m+1=2(n+m)+1, which is an odd number." This is something that all university students will have learned at both junior and senior high school levels, and does not demand a particularly high level of mathematical knowledge or computational accuracy. Why, then, could more than 60% of university students not even provide a semi-correct answer? It is because they have not learned how to correctly use variables, as they should have in the first year of junior high school. The 20th century labor market demanded workers possessing as much knowledge as possible and capable of accurate processing of information, and the school system had no choice but to respond to those demands. However, it is possible that the attempt to efficiently provide students with this knowledge and these skills in the 12 years of elementary and high school education produces human resources who may be characterized by the attitude "I don't understand what it means, but at least I've learned it, and if I'm told to do it, at least I can do it." In other words, the survey results generated the concern that students are not learning the ability to look, read, hear, write, and speak well which should be the greatest factors in distinguishing humans from AI, and are only acquiring surface-level skills which make them easily replaceable by contemporary AI. The fact that the Todai Robot, which does not understand meaning and is only able to solve problems on a superficial level, achieved a T-score of more than 57, and placed in the top 25% of senior high school students for two consecutive years in 2015 and 2016, suggested that this hypothesis may very well be correct.

The authors therefore developed a method of directly measuring the extent to which junior and senior high school students were actually able to understand basic sentences in their textbooks. This was termed the "Reading Skills Test."

[2] In "Basic Thinking regarding Future Educational Reform," published in 2016, the Japan Business Federation also assumed that Japanese students possessed a high level of basic reading comprehension and skills, and concluded that the nation should work to advance English language and programming education and active learning (Japan Business Federation, 2016).

NIRA NIPPON INSTITUTE FOR RESEARCH ADVANCEMENT

Section 2

Necessary Skills for the Era of Digitalization

1. The Reading Skills Test

Types of Reading to which AI are Suited and Not Suited

The Reading Skills Test (RST) is a test which measures the extent to which the examinee is able to quickly and accurately understand the meaning and intention of text in documents including textbooks, newspapers, manuals, and contracts (National Institute of Informatics press release, 2016a). Rather than testing reading comprehension of long Japanese sentences, the RST tests whether the examinee is able to correctly interpret short sentences of between 50 and 200 characters in length, drawn from sources such as textbooks and newspapers. The examinees are instructed to accurately interpret as many of the sentences as possible in the allotted time. Currently, questions might either require the examinee to select one answer from two or more options, or multiple answers from multiple options, and no time limit is set for each question. The questions are divided into two types: One measures whether examinees are able to interpret the superficial information that the sentence is communicating; the other measures whether they are able to understand the meaning of the sentence and make inferences on this basis. The first type includes questions related to (1) Dependency analysis, (2) Anaphora resolution, (3) Paraphrasing, (4) Logical inference, (5) Representation, and (6) Instantiation.

This paper cannot provide commentary on all of these question types, and readers are urged to refer to National Institute of Informatics press release, 2016a and Arai et al, 2017 for further details. Broadly speaking, question types (1) to (3) are questions which are, comparatively speaking, possible for AI to solve, and are already the subject of research concerning analysis using AI. Question types (4) to (6), by contrast, can be considered difficult for AI to solve unless there is discontinuous and dramatic innovation in the field. Type (4), Logical; inference, tests whether, when provided for example with the information that because Europe is located at a higher latitude than Japan, summer days are longer, the examinee is able to judge what the length of summer nights in Europe is like in comparison to Japan. By means of logical inference based on the commonsense notion that a day is made up of day and night, humans are able to see that because Europe is located at a higher latitude than Japan, the nights in summer are shorter. Lacking commonsense notions, such inferences are difficult for an AI[3]. Through inference, humans are able to create a rich and elaborate picture of the world based on limited knowledge.

Here, we will consider questions related to (1) Dependency analysis, and (5) Representation, and look at the rate of correct answers. To begin, we will look at the question shown in Figure 1.

[3] The fifth-generation computer project, which sought to universally realize inference of this type, was a failure. However, the project did realize inference within a limited scope, via so-called "expert systems."

This question tested whether the examinee was able to correctly recognize the dependent relationships in the sentence.

Figure 1: Example of question testing (1) Dependency analysis

Read the following sentence:

Buddhism spread mainly to Southeast Asia and East Asia, Christianity to Europe, North and South America and Oceania, and Islam to North Africa, West Asia, Central Asia, and Southeast Asia.

Choose the most appropriate answer from the given choices that correctly fill the blank in the following sentence.

( ) has spread to Oceania is.
A. Hinduism
B. Christianity
C. Islam
D. Buddhism

(Source) Tokyo Shoseki Co., Ltd., New Society: Geography (a junior high school Social Studies textbook), P. 36

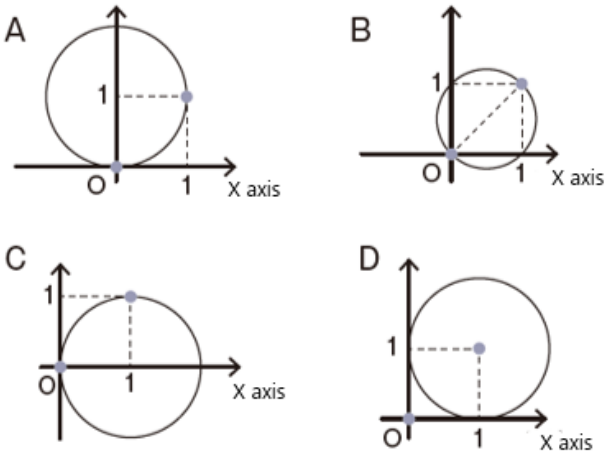
This type of question can be correctly answered by syntactic analysis programs which have already been developed. The correct answer is "B."

The question shown in Figure 2 is classified into the group (5) Representation, and involves selecting the image which correctly represents the sentence.

Figure 2 Example of question testing image identification

Select the figure or figures from A to D that express the content of the following sentence:

The circle that passes through the origin point, 0, and point (1,1) is tangent to the X axis.



The correct answer is "A." This type of question is difficult for a machine to solve unless the frame is radically restricted.

Humans also display an Insufficient Level of Understanding

Figure 2 shows the rate of correct answers to the two questions discussed above for high school students taking the test. The rate of correct answers among third-year senior high school students to the problem involving recognition of dependency shown in Figure 1 was high at 94%. However, the rate of correct answers among the same third-year high school students to the representation question in Figure 2 was only 45%. The students who sat the test were almost all third-year senior high school students intending to advance to university, and yet were unable to understand a question text able to be answered based on mathematical knowledge which should have been learned in the second year of junior high school. This is not a simple problem of a difference between humanities- and sciences-oriented students.

Table 2 Rate of correct answers to the questions shown in Figures 1 and 2 for government junior and senior high schools (Unit: %)

Academic year	Figure 1	Figure 2
1st-year junior high	58.2	10.7
2nd-year junior high	48.8	22.2
3rd-year junior high	64.6	25.4
1st-year senior high	71.8	29.0
2nd-year senior high	84.2	30.0
3rd-year senior high	94.1	45.5

It is often the case that people doubt, in surveys of this type which are not directly related to grades or university entrance, that the students were actually making a serious effort. However, if the majority of the students had selected answers at random, quite a large number could be expected to have selected “A. Hinduism” as the response for the question shown in Figure 1, but as it was, only 1% selected this answer. In other words, at the very least the students who were attending schools preparing them for university admission seriously read the questions, and seriously selected their answers. All of the questions in the survey were of the type that the better the student’s grades, the more likely they were to do well, and the worse the student’s grades, the more likely they were to make mistakes, but even under these conditions, the rate of correct answers for the question shown in Figure 2 was less than 50%. There is no point in teaching students who cannot answer these questions trigonometric functions. And yet, oddly enough, we may consider that the majority of these students will by some means be able to superficially go through the motions of solving trigonometric functions (or in fact not actually solve them) and get through the university entrance examinations.

2. Theoretical Concept of the Reading Skills Test

Estimation of Difficulty of Questions

Prior to advancing the discussion any further, this section will consider the theoretical concept of the RST.

The goal in the creation of the RST was to produce an adaptive test in which questions were drawn from groups of more than 100 questions, depending on the status of the answers provided by the examinee. This means that the questions being answered by examinees sitting next to each other would be different. In the surveys conducted up to the present, offered to 15,290 examinees ranging from sixth-year elementary school students to workers, the questions have been presented at random. Why has this method not caused any problems, and why was this method adopted in the first place?

Normally, test results are expressed as a total of scores for each question. However, even if, for example, a student scores 90 out of 100, this does not necessarily indicate that this student possesses a high level of ability, because the test may be composed exclusively of easy questions. The total score obtained on a test is dependent not only on the ability of the student, but also on the level of difficulty of the questions. In order to more accurately measure reading ability, the RST therefore uses the method of evaluating the student's ability based on the difficulty of the question presented. The level of difficulty of the questions is estimated in advance, based on past answer data for the examinee. Having examinees solve questions the difficulty of which has already been estimated makes it possible to estimate ability independently of the difficulty of the question, and to appropriately evaluate each examinee despite the fact that they solve different questions. (See the box below for further details).

The method employed to estimate the difficulty of test questions will be explained using the most basic model. The rate of correct answers to each question is expressed as the function θi-bj. Here, θi is examinee i's ability, and bj is the difficulty of test item j. In this model, the rate of correct answers is more than 50% whenθi>bj, 50% whenθi=bj and less than 50% whenθi<bj. This model incorporates the concept that the difference between examinee i's ability and the level of difficulty of test item j is an important factor, and estimates the level of difficulty of each question from data for students' past answers. The higher the value of b, the more difficult the question. This approach is called item response theory.

(Reference)
Lord and Novick (1968), *Statistical Theories of Mental Test Scores*

But why does the RST deliberately present different questions to each examinee? Our aim in creating the RST was to realize a test which, each time a examinee answered a question, presented the appropriate question to the examinee based on the answer data for the previous question[4]. In order to do so, it was necessary to prepare in advance a large number of questions the difficulty of

[4] This is termed an adaptive test. For further details, see Linden and Glas (2010).

which had already been estimated. We therefore presented our 15,290 examinees with different questions and compiled answer data, enabling us to estimate the difficulty of a large number of questions. Had we presented the same questions to all the examinees, we would have been unable to obtain this large number of items for which difficulty has been estimated.

Correction of the Content of Questions

Looking towards the realization of an adaptive test, the RST project is currently proceeding with the formulation of questions, the implementation of tests, the estimation of the level of difficulty of questions, and the correction of the content of questions.

Figure 3 divides the ability of examinees in the recognition of grammatical dependency into four groups, and shows the percentage of each group which selected options A, B, C, or D for the question shown in Figure 1. Because option B, "Christianity," was the correct option for this question, the line for option B is shown as bold. Figure 3 shows that the percentage of correct answers for this question increases as ability in recognizing grammatical dependency increases. This question can therefore be considered a question which appropriately reflects ability in recognizing grammatical dependency.

Figure 3 Differences in rate of selection of options for the question shown in Figure 1 by level of ability in grammatical dependency problems

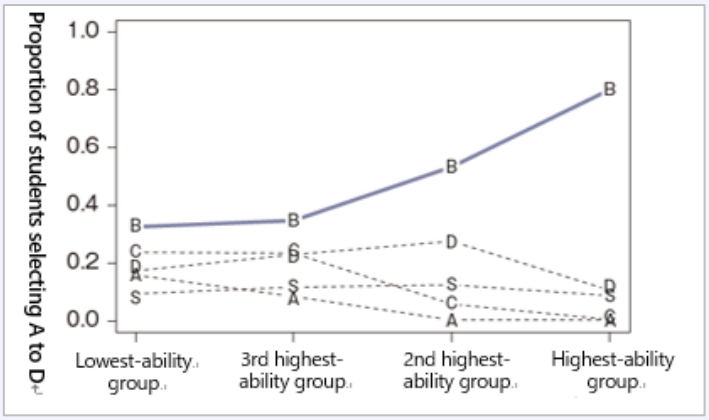


Figure 4 Differences in rate of selection of options by level of ability in grammatical dependency problems – Recognition of problem requiring correction –

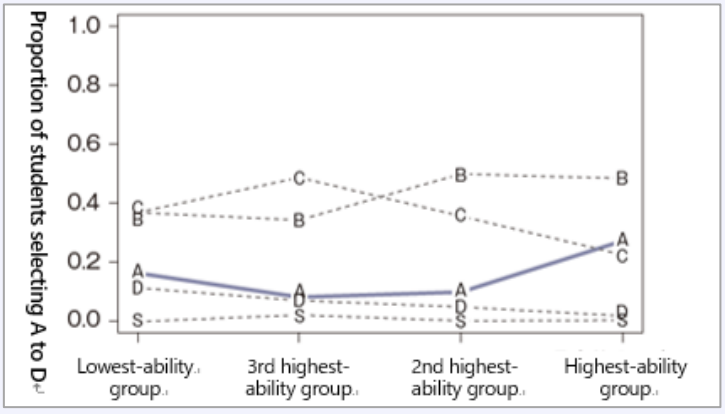


Figure 4 shows the selection of responses for a different question, classified by ability in recognizing grammatical dependency. The correct option for this question was A. Because the percentage of examinees selecting A as the answer in the group displaying the highest level of ability was higher than that of the other groups, the question can be considered to be one which distinguishes high-ability examinees from low- to medium-ability examinees. However, the group displaying the highest level of reading ability also features the highest percentage of selection of the incorrect option B. This leads to a number of conjectures. One of these is the possibility that this is a question for which not only the actual correct answer, but also an incorrect answer, can be taken to be the correct answer. If this was the case, it would be necessary to correct the content of the question and the options for the answer. Alternatively, this may be a question which distinguishes examinees of extremely high ability (a group of even higher ability within the group displaying the highest level of ability). In this case, it would be desirable to offer the question at schools, etc. at which it was considered that the rate of correct answers would be high. This provides an indication of how the RST project is proceeding with the correction of the content of the questions offered on the test.

3. Results of the Reading Skills Test

Students who read in the Same Way as AI

Because all the questions on the RST either request examinees to select one answer from two or more options or multiple answers from multiple options, there is some chance of selecting the correct answer by chance. Table 3 shows the rate of correct answers for all question types for first- to third-year government junior high school students and first- to third-year senior high school students intending to advance to university, and the percentage of students whose rate of correct answers could not be considered any better than random answers (these answers are termed "random answers" below), as determined statistically. For all question types, the rate of correct answers largely increases, and the percentage of students offering random answers largely decreases, as the academic level of the students increases. In addition to indicating that the junior and senior high school students are extending their abilities in each type of reading at school and in their daily lives, this also demonstrates the validity of the RST[5].

What should be given attention here is the fact that the rate of correct answers for question types (4) to (6), which are difficult to solve without an understanding of the meaning of the sentences employed, is lower, and the percentage of students offering random answers is higher, than they are for question types (1) to (3), which can be answered by AI. It can be considered that quite a few students acquire a level of reading ability which makes them prone to replacement by AI, as discussed above.

[5] However, rather than being the result of development in the students, the difference in results between third-year junior high school students and first-year senior high school students should probably be interpreted as the result of senior high school entrance examinations having screened students with significant reading comprehension problems. (We intend to publish results for senior high school students with a lower level of ability in the near future.)

Table 3 Rate of correct answers for each question type and proportion of students giving random answers (in parentheses) (Unit: %)

	(1) Dependency analysis	(2) Anaphora resolution	(3) Paraphrasing
1st-year junior high	56.4 (44.6)	55.4 (46.5)	67.3 (58.1)
2nd-year junior high	58.4 (40.5)	55.9 (42.6)	71.8 (52.5)
3rd-year junior high	64.5 (29.3)	66.8 (24.4)	74.7 (41.8)
1st-year senior high	82.0 (6.7)	82.1 (5.8)	87.5 (17.0)
2nd-year senior high	86.4 (3.5)	81.3 (10.4)	88.7 (12.5)
3rd-year senior high	87.5 (3.0)	86.9 (2.2)	90.5 (8.5)

	(4) Logical inference	(5) Representation	(6) Instantiation
1st-year junior high	50.6 (68.4)	26.4 (57.7)	23.6 (62.1)
2nd-year junior high	54.0 (58.9)	28.6 (45.4)	25.4 (56.6)
3rd-year junior high	57.9 (49.9)	36.3 (34.5)	33.7 (41.4)
1st-year senior high	67.9 (32.9)	51.1 (16.8)	49.1 (19.3)
2nd-year senior high	69.7 (29.5)	57.7 (10.4)	50.0 (8.5)
3rd-year senior high	74.9 (19.1)	53.2 (4.4)	51.3 (14.3)

Abilities which should be developed for the Digital Era

Until the authors conducted the survey discussed in this paper, neither the government nor the academic world nor private enterprise had conducted a single survey in order to answer the question "Can Japanese junior and senior high school students read Japanese at the level of their textbooks?" This despite the fact that criticism to the effect that the quality of written Japanese among today's young people is poor and that young people cannot even adequately read a newspaper are frequently heard, and that unified tests of academic ability are conducted every year at a cost of ¥5 billion. This indicates a definitive failure on the part of educational reforms.

The recent revision of the government's educational guidelines has positioned moral education as a formal subject. Irrespective of the merit or otherwise of this move, it has been pointed out that there is a considerable amount of overlap between elementary school Japanese language education and moral education. This point can be grasped by comparing the methods of teaching Japanese and English in Japanese schools. The English language is logically broken down and taught as an external object called "English." By contrast, it is assumed that students will absorb the Japanese language naturally, because it is their mother tongue, and there is little rigorously logical guidance in its use. For example, it is almost never the case that such logical constructions as "If "If A then B" is true, "If B then A" is not necessarily true," or "If "Everyone is A" is true, then "There is no-one who is not A" is true, and "There are also people who are A" is false" are taught in Japanese language education.

Vastly less attention is paid in Japanese language education to logical activities such as learning how to correctly read graphs and tables, learning how to present information in itemized form, and examining the validity of materials than to conjecturing about the feelings of the author of a text and

making presentations. If moral education is made a formal subject, there would be merit in considering transferring (not all, but perhaps around half of) activities in which students attempt to conjecture about the feelings of authors from the Japanese language curriculum, and shifting towards the abovementioned logical activities.

4. The Necessity for Evidence-based Education Reform

The attempt to institute reforms in the absence of evidence is doomed to failure. It is no doubt unnecessary to even mention the reforms undertaken in the former Soviet Union or in Cambodia under the Pol Pot regime. But has the necessary evidence been obtained in the case of the education reforms which Japan is now attempting to undertake, and was it in the case of reforms undertaken in the past? Education is an experience which everyone undergoes, and everyone has their own opinion regarding the subject. However, while the people concerned may have the best of intentions, proposals made in the absence of evidence are often harmful. As was emphasized in the proposal by the Japan Business Federation quoted above, in the future educational reform based on evidence will be necessary.

The authors would like to posit the following as particularly urgent tasks: 1) Surveys conducted to determine the potential and the limits of digitalization; 2) Detailed analysis of the effect of digitalization on the labor market; 3) Quantitative simulations studying what will be necessary to ensure a soft landing for Japanese society against the background of rapidly advancing digitalization; 4) Study, with consideration of the effect of digitalization on the labor market, of the extent to which it will be necessary to secure workers with specific skill sets, and what those skill sets will consist of; 5) Design of compulsory and higher education necessary to securing these workers; and 6) Study of the policies which will be necessary in order to achieve this goal (free education, etc.).

Class time will have to somehow be set aside for education to ensure that at least 80% of junior high school students are able, by the time of their graduation, to accurately read junior high school textbooks and newspaper-level Japanese without difficulty. Given the current reading ability of junior high school students, attention to subjects such as programming and English can be considered not only unreasonable but also futile, necessitating their reconsideration. At any rate, time is running out for discussion of reform of education.

References

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

Frey, C. B. and Osborne, M. A. (2013), "The Future of Employment: How Susceptible are Jobs to Computerisation?",
http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf
van der Linden, W. J. and Glas, C. A. W. Eds. (2010), Elements of Adaptive Testing, Springer.

National Institute of Informatics press release (2016a), "Development of a Test enabling Scientific Measurement of the Ability to Accurately read Sentences: Industry-Academia Collaboration in Accelerating Research towards Improving Reading Ability," July 26, http://www. nii.ac.jp/userimg/press_ 20160726 .pdf

National Institute of Informatics press release (2016b), "NII Artificial Intelligence Project "Can a Robot get into the University of Tokyo?": Machine realizes T-score of 50 or Higher in Six Subjects on National Center for University Entrance Examinations Mockup Test / Exceeds T-score of 65 in World History for Two Consecutive Years / Significantly Improves T-score for Physics to 59.0 / Achieves T-score of 76.2 for Essay-type Mathematics Questions (Sciences) with Entirely Automated Responses," November 14, http://www.nii.ac.jp/userimg/press_ 20161114 .pdf

Japan Business Federation (2016), "Basic Thinking regarding Future Educational Reform – Towards the Formulation of the Third Basic Plan for the Promotion of Education," April 19, http://www.keidanren.or.jp/policy/ 2016 / 030 _honbun.pdf

Mathematical Society of Japan Education Committee (2013), "Report concerning the First Basic Survey of Mathematical Ability in University Students," March 14,
mathsoc.jp/publication/tushin/ 1801 /chousa-houkoku.pdf

Noriko Arai
Professor, Information and Society Research Division, and Director, Research Center for Community Knowledge, National Institute of Informatics. Holds a Ph.D. in science (Tokyo Institute of Technology). Specializes in mathematical logic, artificial intelligence, and educational technology.

Koken Ozaki
Associate Professor, Graduate School of Business Sciences, University of Tsukuba, and Visiting Associate Professor, Institute of Statistical Mathematics. Holds a Ph.D. in literature (Waseda University). Specializes in statistical science, behavioral genetics, and social surveys.

This is a translation of a paper originally published in Japanese.
NIRA bears full responsibility for the translation presented here. Translated by Michael Faul.