上越教育大学研究紀要　第38巻第１号　平成30年９月
Bull. Joetsu Univ. Educ., Vol. 38(1), Sep. 2018

Why Japanese students omit reading items in PISA
A comparison with Finnish students via tree-based analysis

Taichi OKUMURA*

(Received January 15, 2018; Accepted May 18, 2018)

ABSTRACT

Previous studies on PISA have argued that Japanese students have difficulty expressing their own opinions in sentences based on written information because they omit reading items requiring such higher cognitive processing. Applying extended tree-based items response models to the data of PISA 2009, this study examined the relationships between item properties and response categories of not-reached, omitted, and answered in comparing Japanese with Finnish students. Japanese students were more likely to omit advanced and open-ended items than Finnish students. However, the increments of the item difficulty for Japanese students owing to the requirements of the advanced processing were smaller than for Finnish students. Therefore, Japanese students tended to decide to omit items based not only on the difficulty but also on their superficial impressions of the advanced items.

KEY WORDS
Japanese students, omission of items, PISA, reading literacy, tree-based analysis

1　Introduction

1．1　Omission tendency of Japanese students in PISA

In Japan, researchers have been interested not only in rankings and changes in scores of the PISA (Programme for International Student Assessment) but also in the relatively high rate of missing responses in the tests. Concerning the reading assessment, Arimoto (2006) indicated that Japanese students especially tended to omit open-ended questions compared with the average of the Organization for Economic Co-operation and Development (OECD) countries. When he examined the per requisite aspect of cognition, he also found that the rate of omission was especially high in the items that required students to reflect and evaluate written texts. Similar features were identified in results of the mathematics and science assessments. He argued that the omissions occurred because Japanese students had difficulty in logically expressing their own opinions in sentences. He also indicated that students were not encouraged in such activities by their teachers very much and that they read fewer books for enjoyment in their daily life.

According to the technical report of PISA 2009, Japanese students omitted on average 5.01 items in a booklet (OECD, 2012). This is the 32nd largest score among the 65 participant countries (−0.18 in the standardized score) and the 14th largest among the 34 OECD countries (0.31 in the standardized score). The rankings themselves do not seem to be so extreme as to require discussion. However, the major feature of the Japanese scores is that the number of the non-responses is much higher than expected from the achievement scores of the students. A residual value can be obtained per country by regressing the average numbers of the omitted items on the country means of the reading scores. These scores describe the differences between the observed numbers of the omitted items and the predicted values. According to this analysis, Japan shows the third largest residual value among the OECD countries following France and Israel, whose reading scores are much lower than those of Japan. In fact, Japanese students omitted many more items than New Zealand students (3.27 items on average) and Canadian students (2.69 items

*School Education

on average), which are the closest to Japan in their reading achievements.

1．2 Scaling methods in PISA

The PISA differentiates not-reached items from simply omitted items when categorizing missing values (OECD, 2012). All consecutive missing values clustered at the end of a test session are coded as the not-reached items, because students could not solve them in the first place for lack of time. However, items are regarded as intentionally omitted when no responses were provided even though the students were expected to solve them. Test scores in the PISA are estimated by using the mixed coefficients multinomial logit model described by Adams, Wilson, and Wang (1997), which was developed for treating both binary responses and multiple responses simultaneously (OECD, 2012). The model describes the probabilities of response categories (not-reached, omitted, and answered) for an item as a linear combination of item parameters that are defined by design vectors. That is, the model assumes that all response categories for an item reflect a common ability. Although the model is multidimensional, it assumes different dimensions between items but not among response categories within an item.

However, there are no guarantees that this assumption is valid. If some students solve problems more carefully, they will finish fewer items within a limited time, but the proportion of correct answers may be high if the slowness was caused by their carefulness. It is then obviously insufficient to evaluate students who work slowly but accurately and students who work rapidly but inaccurately on the same scale. Verhelst, Verstrale, and Jansen (1992) indicated that ability parameters were estimated incorrectly when speed and accuracy of response were not differentiated. Similarly, less motivated students will be more likely to omit items that are unfamiliar and appear to be cumbersome, although the items might have been easier than they had appeared to be initially. In these cases, the omissions cannot be explained by their lack of ability. Therefore, it would be more natural to distinguish the slowness and omission tendencies from the substantial knowledge and skills that are required to solve PISA items. Of course, students must reach an item and then decide to complete it before giving an answer. Thus, we can recognize different dimensions among response categories within an item as they are at different stages. Previous studies on the scaling of PISA have not mentioned the mechanism behind missing responses (e.g. Goldstein, Bonnet, & Rocher, 2007; Li, Oranje, & Jiang, 2009), while the scores and rankings are sensitive to the choice of statistical models (Brown, Micklewright, Schnepf, & Waldmann, 2007). Therefore, the main question of this study was to ascertain how Japanese students would be evaluated if the response categories were scaled to reflect different stages in a sequential process of solving reading items.

1．3 Modeling cognitive process by response trees

Let us consider a response tree that expresses the sequential process of solving an item. As shown in Figure 1, the tree consists of three nodes that provide two branches respectively. The top node pertains to whether a respondent reached the item or not. If the respondent reaches it, then he or she decides whether to omit it at the second node that reflects the omission tendency. Answers are reported only if the respondent did not give up on solving the item. The third node pertains to the ability to give a correct answer (or answers) to the item. Drawing such trees helps us to understand what kinds of traits exist and how these traits produce the response categories through cognitive processing. This response tree can also be applied to modeling the cognitive processing in PISA.

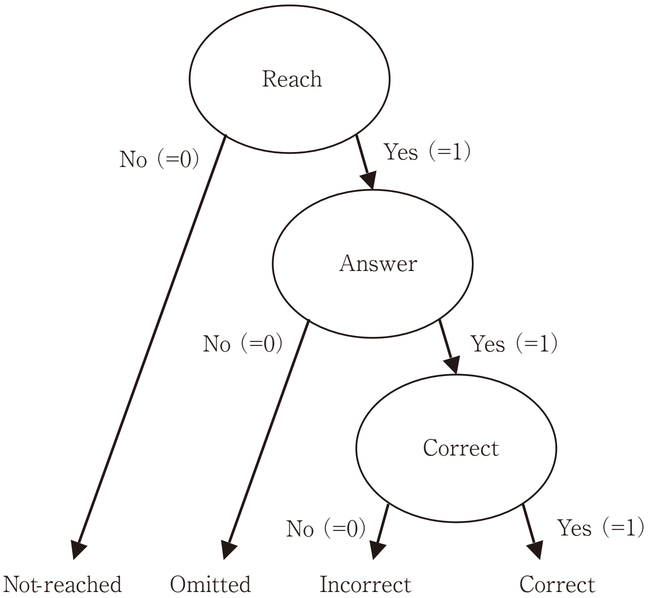

Figure 1. The response tree for the four categories. Ellipses are the assumed internal nodes. The branches labeled 0 or 1 represent the binary responses specific to the nodes.

De Boeck and Partchev (2012) proposed tree-based item response (IRTree) models for describing cognitive processes that are expressed in response trees. The models assume that a product of the probabilities that are specific to the corresponding nodes determine the probability of each response category. According to the formulation, the probabilities of the four response categories in Figure 1 are given by

$$\pi\,(Y_{pi}=0)=\pi\,(Y^{*}_{pi1}=0), \qquad (1.3.1a)$$

$$\pi\,(Y_{pi}=1)=\pi\,(Y^{*}_{pi1}=1)\,\pi\,(Y^{*}_{pi2}=0), \qquad (1.3.1b)$$

$$\pi\,(Y_{pi}=2)=\pi\,(Y^{*}_{pi1}=1)\,\pi\,(Y^{*}_{pi2}=1)\,\pi\,(Y^{*}_{pi3}=0), \qquad (1.3.1c)$$

$$\pi\,(Y_{pi}=3)=\pi\,(Y^{*}_{pi1}=1)\,\pi\,(Y^{*}_{pi2}=1)\,\pi\,(Y^{*}_{pi3}=1), \qquad (1.3.1d)$$

where Y_{pi} is the response to the ith item of the pth person, and Y^{*}_{pir} is the binary response to the rth node. De Boeck and Partchev (2012) assumed that the logit of each node-specific probability is a linear function of the properties of respondent and node, i.e., $\mathrm{logit}(\pi\,(Y^{*}_{pir}=1))=\theta_{pr}+\beta_{ir}$, with θ_{pr} as the propensity of pth person for taking 1 at the rth node, and with β_{ir} as the intensity of taking 1 (i.e. minus difficulty or easiness) at the rth node in the ith item.

Table 1 is a mapping matrix for this response tree where each row stores the hypothesized binary responses for the internal nodes that correspond to an observed response category. The matrix shows that every response category can be expressed in a combination of the responses to the nodes. The probabilities of the response categories can be integrated by referring to the mapping matrix. In the general case where the response tree has R nodes, the probability of the mth response category is

$$\pi\,(Y_{pi}=m)=\Pi^{R}_{r=1}(\pi(Y^{*}_{pir}=1)^{t_{mr}}\,(1-\pi(Y^{*}_{pir}=1)))^{d_{mr}}, \qquad (1.3.2)$$

where t_{mr} is the (m, r)th entry of the mapping matrix, and d_{mr} takes 0 if t_{mr} is missing and 1 otherwise. The feature of IRTree models is that they can specify the latent traits at different stages in the sequential process, so they can assume different traits for response categories in an item. IRTree models are the generalized linear mixed models (GLMM) where the outcome variables are the binary responses to the nodes that are identified by the mapping matrix (De Boeck & Partchev, 2012).

Table 1
A mapping matrix for converting response categories into node-specific responses

	Node 1 (Y_1^*)	Node 2 (Y_2^*)	Node 3 (Y_3^*)
Not-reached ($Y=0$)	0	-	-
Omitted ($Y=1$)	1	0	-
Incorrect ($Y=2$)	1	1	0
Correct ($Y=3$)	1	1	1

As an application of IRTree models, Partchev and De Boeck (2012) examined whether the rapid responses and slow responses reflected different aspects of intelligence by using the data of two intelligence tests. They found that response trees with constraints of one-dimensionality were poorly fitted to the data compared with trees that had no such constraints, although the two aspects were highly correlated with each other.

The purpose of this study was to model the stages of cognitive processing by applying an IRTree model to the reading data of PISA 2009. In particular, this study will focus on the effects of properties of the reading items (i.e., item format, text format, and requisite aspects of cognition) at each stage of the process. Previous studies also have argued that these properties have affected the ability of Japanese students to solve reading items but they have not specified the sequential process (Arimoto, 2006; 2008). However, the degrees and directions of the effects can change from stage to stage. Such interactions cannot be grasped by the methods that have been adopted so far. The tree-based approach can identify the node-specific effects of the item properties, so it will provide practical clues for understanding Japanese omission tendencies in PISA. In this study, Japanese achievements will be discussed though a comparison with Finnish students, because Finland has always shown top-level scores and equality among both schools and students in the PISA. Finnish students also omitted fewer items than Japanese students on average. Previous research has often discussed Japanese policy in school education by referring to Finland as an exemplar (Takayama, 2009).

2 Methods

2. 1 Data preparation

The cognitive data of Japan and Finland were downloaded from the database of PISA 2009[1]. Japanese and Finnish data consisted of 6,088 and 5,810 students sampled from 186 and 203 schools, respectively. Each country administered 101 reading items in common, although PISA 2009 originally prepared 131 items for the reading test. Students solved about 30 reading items on average, because PISA divides the cognitive items into several booklets based on the balanced incomplete block design (OECD, 2012). Six students and three students were excluded from the Japanese and Finnish data, respectively, because no reading items were administered.

According to the codebook that was downloaded from the database, the original responses were recoded into four categories: Not-reached (= 0), Omitted (= 1), Incorrect (= 2), and Correct (= 3). This study gave full credit only for correct responses, although PISA originally permitted partial credits for some open-ended questions. This was because IRTree models assume a binary response tree. By referring to the mapping matrix shown in Table 1, the recoded data were formatted into a long form, where each row of the data matrix pertains to the event of a response to a node. This was because the lme4 package in R used for fitting IRTree models required this formation (De Boeck & Partchev, 2012).

Subsequently, properties of the reading items were extracted from the technical report (OECD, 2012): item format (closed-ended question or open-ended question), text format (*continuous or non-continuous*), and requisite aspects of cognition (*access and retrieve, integrate and interpret*, or *reflect and evaluate*). To identify the item properties, the following four dummy variables were added to the data matrix:

- ITM, the item format (open-ended = 1 and closed-ended = 0).
- TXT, the text format (non-continuous = 1 and continuous = 0).
- INT, the second cognitive aspect in contrast to the first aspect, (integration/interpretation = 1 and otherwise = 0).
- REF, the third cognitive aspect in contrast to the first aspect, (reflection/evaluation = 1 and otherwise = 0).

All of the four dummy variables take zero for closed-ended items with continuous texts that require students to access and retrieve appropriate information in the texts. Thus, they are the reference items when interpreting the parameters described below. Finally, dummy variables named JPN (Japan = 1 and Finland = 0) and FIN (Japan = 0 and Finland = 1) were added to the data for specifying the countries.

2. 2 Model formulation

The original IRTree models do not include covariates and do not assume multilevel data structure. For investigating the effects of the item properties in PISA, this study extended the IRTree models in the following two points: First, the item properties and the country variable that were coded above were included as linear predictors. Second, the model assumed that the node-specific latent traits vary randomly between schools as well as between persons in order to handle the multilevel data[2]. Then, the logit of each node-specific probability is:

$$\begin{aligned}\mathrm{logit}(\pi(Y^{*}_{pir}=1)) = {} & \theta_{0r}+\theta_{gpr}+\theta_{1r}\times JPN_{gpr}+\beta_{1r}\times ITM_{i}+\beta_{2r}\times TXT_{i}+\beta_{3r}\times INT_{i}+\beta_{4r}\times REF_{i}\\ & +JPN_{gpr}\times(\lambda_{1r}\times ITM_{i}+\lambda_{2r}\times TXT_{i}+\lambda_{3r}\times INT_{i}+\lambda_{4r}\times REF_{i})+e_{gpir}, \qquad (2.2.1)\end{aligned}$$

where e_{gpir} is the residual error that absorbs an imperfect prediction and is randomly distributed with a node-specific variance σ^2_r. It was assumed that the vectors of the node-specific latent traits $\boldsymbol{\theta}_{gp}(=\{\theta_{gpr}\})$ were distributed normally over students around a group mean $\boldsymbol{\theta}_g$ with variances $\boldsymbol{\Phi}_1$ for Japan and $\boldsymbol{\Phi}_2$ for Finland. The group means are distributed normally among schools around an overall mean of $\mathbf{0}$ with variances $\boldsymbol{\Psi}_1$ for Japan and $\boldsymbol{\Psi}_2$ for Finland. Because the grand means are fixed to zero, the non-zero intercepts θ_{or} are included in the linear function. According to the method used to code the response categories, these fixed intercepts are the mean of the node-specific traits of Finnish students when they solve the reference items, while the fixed slopes θ_{1r} are the mean increments of the traits for Japanese students against Finnish students. The fixed parameters βs are the node-specific main effects of the item properties for Finnish students, while the interaction terms λs are increments of the effects for Japanese students against Finnish students.

To test the multidimensionality of the response categories, the goodness of fit was compared with a one-dimensional model by using the likelihood ratio test. The one-dimensionality was modeled by constraining the node-specific traits to have a common variance-covariance matrix for each country.

2. 3 Parameter estimation

The extended IRTree model was fitted to the data by using the glmer function in the lme4 package built in R of the version 2.15.0 (R Development Core Team, 2012). The glmer function can handle GLMM by the penalized iteratively reweighted least squares (PIRLS) algorithm and the Laplace approximation to the likelihood (Doran, Bates, Bliese, & Dowling, 2007). De Boeck and Partchev (2012) illustrated the way to fit completely descriptive IRTree models by using the glmer function. In this analysis, to model the effects of the item properties in a multilevel fashion, the corresponding linear predictors and error terms were described in the glmer code. Aiming at fast convergence, starting values of the fixed effects were prepared by using the glm function, which fits the logistic regression models that do not consider the random fluctuations of the regression coefficients. The model comparison was conducted by using the anova function. The R code for performing this analysis is shown in the Appendix.

3 Results

The likelihood ratio test showed that the multidimensional model (AIC = 447957, log-likelihood = −223918) had a better goodness of fit than the one-dimensional model (AIC = 491559, log-likelihood = −245744) (χ^2_{25} = 43652, p <. 001). Therefore, the remaining description will be restricted to the multidimensional model.

Table 2 describes the estimates of the fixed coefficients for each node with their two-tailed p values and the 95% confidence intervals. The significant main effects and interactions appeared only at the second and third nodes while no predictors significantly affected the first node. The text format had neither significant main effects nor interaction effects with the country dummy at any of the nodes.

At the second node where students decide whether to answer an item or not, the main effects of the open-ended format and the reflection/evaluation aspect were significant and negative. This means that Finnish students tended to omit these items more than reference items. The interactions between the country variable and the three item properties of open-ended format, integration/interpretation, and reflection/evaluation were significant and negative. Thus, Japanese students were more likely to omit these items than the reference items when compared with Finnish students.

At the third node that pertains to whether the answers given are correct or incorrect, the main effects of integration/interpretation and reflection/evaluation were first of all significant and negative. Finnish students had more difficulties in solving these types of items than reference items. Second, the main effect of the country dummy variable was significantly negative. Thus, the reference items were, on average, more difficult for Japanese students than for Finnish students. As with the second node, the interactions of the country dummy variable and the three item properties of open-ended items, integration/interpretation, and reflection/evaluation were significant. Against the second node, however, these effects were all positive. Therefore, the increments of difficulty associated with these items were smaller for Japanese students than for Finnish students. In other words, Japanese students had less difficulty in solving these items than they did reference items when compared with Finnish students.

The variance components of random effects and correlations are shown in Table 3. The school-level variances of the node-specific traits were several times larger in Japan than in Finland. Specifically, Japanese variances of the second and third nodes were about 6.7 and 5.0 times larger than Finnish variances, respectively. The student-level variances of the first and second nodes were almost equivalent between the two countries, excluding that the Finnish value was about 1.5 times larger than the Japanese value at the third node. There were no salient differences in the residual variances between the two countries.

The school-level correlations between the node-specific traits were larger in Japan than in Finland and especially the second and third nodes were strongly correlated in Japan, although the other values were very small. At the student-level, the second node was positively correlated with both the first and third nodes, although the first and third nodes themselves were almost uncorrelated. The correlations were slightly larger in Japan than in Finland, although there were no salient differences in the pattern. The residuals were positively correlated only between the first and second nodes, while the other values were almost zero.

Table 2
Estimates of the fixed effects in the IRTree model

	Coefficient	*SE*	*z*	*p* value	95%CI
node1:					
Item Format	0.35	0.25	1.40	.16	[-0.14, 0.84]
Text Format	0.40	0.24	1.70	.09	[-0.06, 0.87]
Integrate/Interpret	0.22	0.30	0.73	.47	[-0.38, 0.82]
Reflect/Evaluate	0.38	0.33	1.14	.25	[-0.27, 1.03]
Country	0.17	0.15	1.08	.28	[-0.14, 0.47]
Country * Item Format	-0.12	0.09	-1.32	.19	[-0.30, 0.06]
Country * Text Format	-0.02	0.09	-0.24	.81	[-0.20, 0.16]
Country * Integrate/Interpret	0.09	0.11	0.80	.42	[-0.13, 0.31]
Country * Reflect/Evaluate	0.17	0.12	1.35	.18	[-0.08, 0.41]
node2:					
Item Format	-1.71 ***	0.21	-8.07	.00	[-2.12, -1.29]
Text Format	0.35	0.20	1.74	.08	[-0.04, 0.73]
Integrate/Interpret	-0.38	0.26	-1.48	.14	[-0.88, 0.12]
Reflect/Evaluate	-0.75 **	0.28	-2.72	.01	[-1.30, -0.21]
Country	0.12	0.11	1.01	.31	[-0.11, 0.34]
Country * Item Format	-0.93 ***	0.06	-16.19	.00	[-1.04, -0.82]
Country * Text Format	0.07	0.04	1.76	.08	[-0.01, 0.15]
Country * Integrate/Interpret	-0.19 ***	0.06	-3.36	.00	[-0.30, -0.08]
Country * Reflect/Evaluate	-0.44 ***	0.05	-8.10	.00	[-0.54, -0.33]
node3:					
Item Format	-0.04	0.24	-0.16	.88	[-0.50, 0.43]
Text Format	-0.08	0.22	-0.35	.73	[-0.51, 0.36]
Integrate/Interpret	-0.59 *	0.29	-2.05	.04	[-1.15, -0.03]
Reflect/Evaluate	-0.66 *	0.31	-2.11	.04	[-1.28, -0.05]
Country	-0.41 ***	0.06	-7.02	.00	[-0.52, -0.29]
Country * Item Format	0.55 ***	0.02	26.56	.00	[0.51, 0.60]
Country * Text Format	-0.02	0.02	-0.81	.42	[-0.05, 0.02]
Country * Integrate/Interpret	0.14 ***	0.03	4.97	.00	[0.08, 0.19]
Country * Reflect/Evaluate	0.15 ***	0.03	5.11	.00	[0.09, 0.21]

Note. *N* = 11889 (in 389 schools), AIC = 447957, log-likelihood = -223918, CI = confidence interval. Intercepts are omitted for simplicity.
*p < .05. **p < .01. ***p < .001.

Table 3
Estimated variances, standard deviations, and correlations of the random effects in the IRTree model

	Variance		*SD*		Correlation		
School-level:	Japan	Finland	Japan	Finland	node1	node2	node3
node1	0.43	0.23	0.66	0.48		.05	-.04
node2	0.94	0.14	0.97	0.37	.10		.15
node3	0.35	0.07	0.59	0.26	.11	.74	
Student-level:							
node1	4.90	4.14	2.21	2.03		.45	.09
node2	3.27	3.32	1.81	1.82	.36		.54
node3	0.65	0.95	0.81	0.97	.03	.42	
Residual:							
node1	1.19		1.09				
node2	0.87		0.93		.36		
node3	1.14		1.07		.06	.07	

Note. *N* = 11889 (in 389 schools), AIC = 447957, log-likelihood = -223918. The lower and upper triangular parts of the correlation matrices of the random effects are for Japan and Finland, respectively.

4 Discussion

According to the results, the likelihood of items not being reached did not depend on the item properties and countries. Perhaps this is because there were no salient differences in the speed of solving items between Japanese students and Finnish students and because the locations of the item clusters were randomly distributed in the booklets of the PISA. Thus, the remaining description will be restricted to the omission tendency and the reading ability of students.

From the cognitive framework, it is not possible to interpret or integrate information without having first retrieved it and it is not possible to reflect on or evaluate information without having made some sort of interpretation (OECD, 2009). Therefore, integration/interpretation items and reflection/evaluation items require of students more higher processing than to simply access/retrieve words or sentences in texts. Similarly, item formats also vary in their typical cognitive demands: closed-ended questions often elicit low-level cognitive processing whereas open-ended questions more often evoke complex thinking (Martinez, 1999). In particular, low-skilled students tend to do better on the closed-ended items than on open-ended items (Lafontaine & Monseur, 2009). The results showed that both Japanese and Finnish students commonly tended to omit or otherwise miss reading items that required them to construct their own opinions logically.

By modeling the sequential process behind the response categories, however, differences between Japanese and Finnish students became clearer. The influences of the item properties on the omission tendency were different between Japan and Finland. That is, compared with Finnish students, Japanese students were more likely to omit items that required them to answer in sentences and items that required them to integrate and interpret or reflect and evaluate written texts. This tendency in Japanese students coincides with that identified in previous studies, which have ascribed the high rates of omission to a lack of reading ability (e.g., Arimoto, 2006; 2008). That is, they have argued that Japanese students had no choice but to omit the items because they had difficulty in combining appropriate information and in logically expressing their own opinions. If this had been true, the difficulty of such advanced items also should have increased from that of the most basic items in parallel with the increments of the omission tendency. Nevertheless, the increments of the item difficulty owing to such advanced processing were smaller for Japanese students compared with Finnish students. Of course, this does not mean that Japanese students could more easily solve these items than Finnish students, but only that they tended to omit more advanced items than predicted from their likelihood of giving correct answers. This implies that Japanese students decided whether to answer or omit items based not only on the actual difficulties but also on superficial impressions of the difficulties before tackling them. In contrast to Japanese students, Finnish students relatively more often omitted advanced items because they experienced them as more difficult than the basic items.

The PISA report indicated that the achievement gaps between schools were comparatively large in Japan (OECD, 2010). The tree-based analysis also showed that in Japan not only the variances of but also the correlations between the omission tendency and the reading ability were especially high at the school-level. This implies that student performances on the test were largely dependent on their engagement with reading in their classroom. The inequality among Japanese schools occurs because Japanese students take the PISA after they are assigned to high schools mostly based on scholastic tests. Thus, teachers in some lower-ranked schools will have to consume more time instructing students in basic reading skills than in encouraging them to grasp contents and evaluate them critically.

If only their reading skills had been affected by their engagement on the higher-level reading activities, the unbalanced increments in the omission tendency and the difficulty of items would not have occurred. Therefore, the lack of such engagement will also cause lower motivation and negative attitudes to solving advanced items. It is possible that this was caused by self-handicapping. In general, when people are confronted with an achievement situation in which they expect to fail, they take steps to protect their self-esteem by withdrawing effort, thereby creating an explanation other than lack of ability for the failure (i.e., lack of effort) (Urdan & Midgley, 2001). Thus, Japanese students may have avoided putting forth their best efforts to tackle unfamiliar tasks that did not assure them of success. In that case, Japanese PISA scores will be improved by teaching them skills for expressing their opinions logically based on the interpretations of written texts.

Another possibility is that Japanese students automatically refrained from expressing their own views about texts because modesty as a norm is prevalent in East Asian cultures. It is known that Japanese people express lower self-esteem and self-evaluations of their performance than Western people (Heine, Lehman, Markus, & Kitayama, 1999). The same tendency was indicated in the results of the TIMSS (Trends in International Mathematics and Science Study), a worldwide educational survey other than the PISA (Shen, 2002; Shen & Pedulla, 2000). People with low self-esteem tend to be less assertive (Leary, Schreindorfer, & Haupt, 1995), and in practice Japanese people were less assertive compared with Western (including Finnish) people in a cross-cultural survey (House, Hanges, Javidan, Dorfman, & Gupta, 2004). If Japanese students omitted items because of their social customs, their reading scores will not be improved by skill-oriented instruction. Instead, it will be more effective to motivate them to tackle such unfamiliar tasks by instructing them in the practical significance of making assertions based on critical considerations.

There are three limitations of this study. First, this study examined a tree-based model that only contained item properties as covariates in the analysis. In PISA, however, students and teachers participated in questionnaires that covered their background information and attitudes toward schools, learning, and instruction. Including such psychosocial covariates at student and school-levels will contribute to practical solutions for improving their reading activities. Second, the analysis regarded all partial credits as completely incorrect answers because IRTree models assume binary response trees where all nodes provide two branches. The models should be extended to allow partial credits by letting there be more than two branches at each node. Finally, the Japanese government revised the national standard curriculum for public schools in 2008. The revised curriculum, enacted in 2012, aims to develop the linguistic abilities of students through verbal activities such as record-keeping, explanation, critique, dissertation, and learning to debate in various subjects (MEXT, 2008). It is expected that the effectiveness of this policy will be examined in the future by tracking Japanese data from the PISA over time.

Footnotes

1) http://pisa2009.acer.edu.au/downloads.php

2) The assumptions of normality are imposed in the original PISA scaling.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Arimoto, T. (2006). PISA-chosa ni okeru Nihon no kadai. [What's wrong with Japanese students in PISA?] *Sokutei Report* (*School of Education, The University of Tokyo and Benesse*), *4*, 54–66. Retrieved on June 4, 2012 from http://www.p.u-tokyo.ac.jp/sokutei/pdf/200608/2006report5.pdf

Arimoto, H. (2008). Nihon no koukousei no PISA dokkairyoku to kagakuteki literacy no kadai. [Problems of Japanese students' reading literacy and scientific literacy.] *Kagaku Kyoiku Kenkyu, 32*, 245–250.

Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007). International surveys of educational assessment: How robust are the findings? *Journal of the Royal Statistical Society. Series A* (*Statistics in Society*), *170*, 623–646.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, 48*, 1–28.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software, 20*, 1–18.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel statistical equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics, 32*, 252–286.

Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review, 106*, 766–794.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies.* Thousand Oaks, CA: Sage.

Lafontaine, D., & Monseur, C. (2009). Gender gap in comparative studies of reading comprehension: To what extent do the test characteristics make a difference? *European Educational Research Journal, 8*, 69–79.

Leary, M. R., Schreindorfer, L. S., & Haupt, A. L. (1995). The role of low self-esteem in emotional and behavioral problems: Why is low self-esteem dysfunctional? *Journal of Social and Clinical Psychology, 14*, 297–314.

Li, D., Oranje, A., & Jiang, Y. (2009). Estimation of hierarchical latent regression models for large-scale assessments. *Journal of Educational and Behavioral Statistics, 34*, 433–463.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*, 207–218.

MEXT (2008). *The revision of the course of study for elementary and secondary schools.* Retrieved on June 4, 2013 from http://www.mext.go.jp/english/elsec/__icsFiles/afieldfile/2011/03/28/1303755_001.pdf

OECD (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science.* Retrieved on June 28, 2013, from http://www.oecd.org/pisa/pisaproducts/44455820.pdf

OECD (2010). *PISA 2009 results: What makes a school successful?- resources, policies, and practices (volume IV).* Retrieved on July 2, 2013 from http://www.oecd.org/pisa/pisaproducts/48852721.pdf

OECD (2012). *PISA 2009 technical report.* Paris, France: OECD Publishing.

Partchev, I. and De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence, 40*, 23–32.

R Development Core Team. (2012). R: A language and environment for statistical computing [Computer software]. Vienna, Austria. Retrieved on August 1, 2012 from http://www.R-project.org/ (ISBN 3-900051-07-0)

Rose, N., von Danier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT).* (ETS Research Report ETS RR-10-11). Princeton, NJ: Educational Testing Service.

Shen, C. (2002). Revisiting the relationship between students' achievement and their self-perceptions: A cross-national analysis based on TIMSS 1999 data. *Assessment in Education, 9*, 161–184.

Shen, C., & Pedulla, J. J. (2000). The relationship between students' achievement and their self-perception of competence and rigour of mathematics and science: A cross-national analysis. *Assessment in Education, 7*, 237–253.

Takayama, K. (2009). Politics of externalization in reflexive times: Reinventing Japanese education reform discourses through "Finnish PISA Success". *Comparative Education Review, 54*, 51–75.

Urdan, T., & Midgley, C. (2001). Academic self-handicapping: What we know, what more there is to learn. *Educational Psychology Review, 13*, 115–138.

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. H. (1992). A logistic model for time limit tests. (Measurement and Research Department Reports 92-1). Arnhem: Cito.

Appendix

The R code for fitting the proposed IRTree model to the data of PISA is as follows. The variables named NODE and VALUE are the identification factors for the nodes and the node-specific responses respectively. Similarly, SCHOOL, ID, and ITEM are the identification factors for schools, students, and items.

```
library(lme4)
data01 <- read.csv("data01.csv")
fit00 <- glm(VALUE~0+NODE+JPN:NODE+NODE:(ITM+TXT+INT+REF)+
                        JPN:NODE:(ITM+TXT+INT+REF),
                        family=binomial,verbose=TRUE,data=data01)

fit01 <- glmer(VALUE~0+NODE+JPN:NODE+NODE:(ITM+TXT+INT+REF)+
                        JPN:NODE:(ITM+TXT+INT+REF)+
                        (0+NODE|ITEM)+
                        (0+NODE:JPN|ID)+
                        (0+NODE:FIN|ID)+
                        (0+NODE:JPN|SCHOOL)+
                        (0+NODE:FIN|SCHOOL),
                        family=binomial,
                        start=list(fixef=coefficients(fit00)),
                        verbose=TRUE,data=data01)

upper <- fixef(fit01)+qnorm(0.975)*sqrt(diag(vcov(fit01)))
lower <- fixef(fit01)+qnorm(0.025)*sqrt(diag(vcov(fit01)))
cbind(coef(summary(fit01)),lower,upper)

fit02 <- glmer(VALUE~0+NODE+JPN:NODE+NODE:(ITM+TXT+INT+REF)+
                        JPN:NODE:(ITM+TXT+INT+REF)+
                        (1|ITEM)+
                        (0+JPN|ID)+
                        (0+FIN|ID)+
                        (0+JPN|SCHOOL)+
                        (0+FIN|SCHOOL),
                        family=binomial,
                        start=list(fixef=coefficients(fit00)),
                        verbose=TRUE,data=data01)

upper <- fixef(fit02)+qnorm(0.975)*sqrt(diag(vcov(fit02)))
lower <- fixef(fit02)+qnorm(0.025)*sqrt(diag(vcov(fit02)))
cbind(coef(summary(fit02)),lower,upper)

anova(fit01,fit02)
```