Comparison of Item and Test Characteristics between the U.S. and Japanese Versions of the Minnesota Test of Critical Thinking-II

Kentaro Kato[a*]
[a]Benesse Corporation,
1-34 Ochiai, Tama, Tokyo, Japan
*Corresponding Author: nekotak@mail.benesse.co.jp

ABSTRACT

Critical thinking is now widely recognized as one of the key competencies in work and life situations in our current and future society. Much work has been done in terms of conceptualization and development of critical thinking, and its assessment has also drawn growing attentions in the past two decades. Among many existing assessments is the Minnesota Test of Critical Thinking-II (MTCT-II). Its assessment framework builds upon the definition and taxonomy of critical thinking skills provided by the American Philosophical Association, and considerable validity evidence has been provided. Considering the relative scarcity of critical thinking assessments in Japan, we developed a Japanese version of MTCT-II in this study. We then compared its performance with the original U.S. version using two datasets from college students in the U.S. and Japan, respectively. Comparisons were made in terms of (a) dimensionality, (b) item characteristics (discrimination and difficulty), and (c) test characteristics (reliability and score distributions). Dimensionality analysis revealed that both versions were roughly unidimensional. Discrimination and difficulty of individual items were very similar with a few exceptions, and so were the overall test characteristics. The results indicate that the particular aspect of critical thinking ability measured by MTCT-II is applicable to Japanese students as well, which implies the possibility that MTCT-II is used for people with different cultural backgrounds.

Keyword: Critical Thinking, Generic Skills, Item Analysis, Test Development, International Comparison

1. Introduction

In the past two decades, there have been growing concerns with generic skills, which include skills related to higher-order reasoning, problem-solving, communication, teamwork, and so on and are generally applicable and crucial to a wide variety of work and life situations in the current and future society (e.g., Australian Education Council Mayer Committee, 1992; Binkley, Erstad, Herman, Raizen, Ripley, &

Rumble, 2010). Although there are several frameworks for generic skills (e.g., National Centre for Vocational Education Research, 2003; O'Neil, Allred, & Baker, 1997; Binkley et al., 2010), critical thinking is one of the key competencies that constitute generic skills.

Teaching and learning of critical thinking has been one of the major topics in education and learning psychology even before the recent concern with generic skills, and also important is its assessment. There are standardized tests such as the Ennis-Weir critical thinking essay test (Ennis & Weir, 1985), the California Critical Thinking Skills Assessment (Facione, 1990), the Watson-Glaser critical thinking appraisal (Watson & Glaser, 1994), and the Collegiate Learning Assessment (CLA; Council for Aid to Education, n.d.). Also, several large-scale assessments in both national and international contexts (e.g., Assessment of Higher Education Learning Outcomes; AHELO) have adopted critical thinking in their assessment frameworks (Kusumi, Koyasu, & Michita, 2011, chap. 1).

Among these attempts, Edman, Robey, and Bart (2002) developed the Minnesota Test of Critical Thinking-II (MTCT-II), which was intended to measure critical thinking skills and the willingness to critically evaluate arguments that are congruent with one's own goals and beliefs. MTCT-II builds upon a taxonomy of critical thinking skills that was derived from the American Philosophical Association's (APA) definition of critical thinking, which resulted from a two-year comprehensive study conducted by a panel of critical thinking theorists and researchers (the Delphi study; American Philosophical Association, 1990). The panel defined the critical thinking as "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based" (p. 3). As a result, the taxonomy included the following six critical thinking skills, each of which further consisted of several subskills: Interpretation, Analysis, Evaluation, Inference, Explanation, and Self-Regulation.

Each item in MTCT-II was aimed at measuring one of the above six critical thinking skills. Although the six-skill structure that they intended was not confirmed, Edman et al. (2002) reported that MTCT-II, as a whole, had relatively high reliability as well as acceptable concurrent validity with other relevant measures. Thus, MTCT-II could be a useful tool to measure critical thinking ability in a relatively simple and objective manner.

The main purpose of the present study is to develop a Japanese version of MTCT-II and evaluate and compare its psychometric characteristics with the original version. The literature indicates that researchers and organizations in the U.S. and Europe have been leading the assessment of critical thinking. In contrast, the research literature in

critical thinking, especially for its assessment, is relatively scarce in Japan with only a few examples (e.g., Hirayama, Tanaka, Kawasaki, & Kusimi, 2010; Kuhara, Inoue, & Hatano, 1983) as Kusumi, Koyasu, & Michita (2011) state, while the importance of critical thinking is being more recognized in the context of higher education in Japan as well. Given this situation, MTCT-II could be an important addition to the inventory of critical thinking assessments in Japan if it shows evidence for reliability and validity. This also increases the generalizability of the test (or the construct) to people in different cultural backgrounds. Accordingly, this study compares item and test characteristics of MTCT-II between the U.S. and Japan.

2. Method

2.1 Participants

Two sets of data, which are from Japan and the U.S., respectively, were used in this study. The Japanese dataset consisted of responses of 200 Japanese undergraduate students (104 males and 96 females, mean age 19.73) from a variety of universities and in various majors. They were recruited for incentive and participated in a one-day testing session in which they took the Japanese version of MTCT-II as well as other tests. They were given 75 minutes to complete MTCT-II.

The U.S. dataset was obtained directly from one of the authors of Edman et al.'s (2002), and consisted of responses of 210 examinees. The data might not be exactly the same as described by Edman et al. (2002), because the numbers of examinees did not match (they had 232 examinees in total). Edman et al. (2002) reported that their data included graduate students as well as undergraduates, who were recruited mainly from educational psychology courses. Edman et al. (2002) reported that the mean age of their examinees was 21.81, which was higher than the Japanese mean probably due to the inclusion of college graduates.

2.2 Instruments

The original MTCT-II (the U.S. version) presents six controversial conversations between two persons. Each controversy is about a certain topic, and these topics are "Logging in National Forests" (#1), "School Vouchers" (#2), "Legalizing Drugs" (#3), "The Death Penalty" (#4), "Grade Retention and Promotion" (#5), and "State Sponsored Lottery" (#6). These topics were chosen to reflect general interest (Edman et al., 2002). In each controversy, the two persons show opposite views about the topic. Examinees are asked to read them and respond to 10 multiple-choice and one free-response items for each controversy (i.e., there are 60 multiple-choice and 6 free-response items in total). In each controversy, each two out of the 10 multiple-choice items represent one of the five APA critical thinking skills:

Interpretation, Analysis, Evaluation, Inference, and Self-Regulation; one free-response item corresponds to Explanation. All multiple-choice items have four response options, except for two items which have five statements for each of which examinees judge whether it is a reason or a conclusion (all correct judgments to the five statements are coded as correct as an item response).

The Japanese version of MTCT-II was a direct translation of the original test. Three Japanese persons (including the author) were involved in the translation process. All controversies and pertaining items (including item stems and response options) were translated into Japanese so that they maintained the original meanings and logical structures while they sounded as natural as spoken Japanese as possible. Several terms in the text were considered unfamiliar to Japanese students, and these terms were annotated.

Scoring of student responses was made in the same manner as the original version, but the free-response part (i.e., the Explanation items) was not considered in the current study. There is one correct response for each of the 60 multiple-choice items, so the total (number-correct) score ranges from 0 to 60. The total score can be decomposed to five skill subscores. Overall, there are 12 items for each of the five skills, and all five skill subscores ranged from 0 to 12. Scores by controversy can also be considered; each of the six controversy subscores ranges from 0 to 10.

For the Japanese data, several examinee characteristics were available; they include age, college grade level (freshman, sophomore, junior, or senior), gender, academic major (humanity or science), and achievement level (given as the "difficulty of entrance" to a particular college department which each student attends; indicated by a deviation score with mean 50 and standard deviation 10). In addition, examinees' prior opinions about the topics discussed in the controversies were recorded before they read the controversies. Each prior opinion was provided on a 4-point scale; "1" indicates that the examinee strongly agrees one of the two opposite view, "4" indicates that the examinee strongly agrees the other view, and "2" and "3" indicates agreement to a lesser degree.

2.3 Statistical Analysis

The main purpose of the present study was to compare the U.S. and Japanese versions in terms of test and item characteristics. First, factor analysis was conducted to examine the dimensionality and equivalence of correlational structures between the U.S. and Japanese versions. Second, test scores (total scores, skill subscores, and controversy subscores) were computed and compared. Third, item statistics (difficulty and discrimination) were computed. Based on the item analysis, several items in the Japanese version that behaved differently from the U.S. version were identified. After

screening out several items that were potentially functioning differently between the U.S. and Japan, test scores were computed again for comparison. Finally, test scores from the Japanese data were correlated with examinee characteristics and prior opinions for validation. The statistical package R (R Development Core Team, 2012) was used for all analyses.

3. Results

3.1 Test Score Analysis

Exploratory factor analysis was conducted to examine the dimensionality of MTCT-II, although a caution should be taken because the sample size is not large enough to obtain stable results. Eigenvalues indicated that the first factor explained 16% and 12% of the total variation for the U.S. and Japanese data, respectively. The contribution of subsequent factors was relatively small (cf., six factors accounted for approximately 30% in total). Thus, it was concluded that the MTCT-II items were roughly unidimensional. Although the number of factors was increased up to six, no meaningful factor loading patterns were found for both U.S. and Japanese data. In the following analyses, the total score was used to represent the critical thinking ability measured by MTCT-II unless noted otherwise.

Table 1 shows means, standard deviations, Cronbach's alpha coefficients, and correlations for subscale and total scores. The Japanese students scored slightly lower than the U.S. students in terms of the total score (31.11 vs. 33.39). The standard deviation was also smaller for the Japanese students than for the U.S. students (7.86 vs. 10.68). Figure 1 depicts the total score distributions for the U.S. and Japan. The U.S. data have wider tails and are more skewed to the left than the Japanese data.

Skill subscores for the Evaluation, Inference, and Self-Regulation followed the same tendency as the total score, while the mean was almost the same between the U.S. and Japan for the Analysis and Interpretation subscales. The fact that the U.S. data showed higher means could be attributed to the inclusion of graduate students, but other factors such as familiarity to the controversy topics might also affect the results.

Reliability for the U.S. version was almost the same as those reported by Edman et al. (2002). Reliability of the total score was .90, which indicated the test as a whole was highly internally consistent, and subscale reliability ranged from .50 to .71. There were slight differences between Edman et al.'s (2002) result and the current one, but these were probably due to the exclusion of the open-ended items from the calculation of total scores. The Japanese version produced lower reliability for all subscales and the entire test; reliability for the total score was .82, and subscale reliability ranged from .28 to .59. Reliability for Evaluation was low for the U.S. data, but it was even lower (.28) for the Japanese data. However, the magnitude of reliability across the

subscales showed the same pattern as in the U.S. version. (As we will see later, the lower reliabilities of the Japanese version are due to several items that showed negative discrimination.)

Correlations between subscale scores were very high (ranging from .54 to .75 for the U.S. data and .21 to .51 for the Japanese data). The U.S. version showed higher correlations than the Japanese version. Given these high correlations and the result of factor analysis, it may not make practical sense to separate the entire scale into subscales.

Table 1. Summary of the total and subscale scores.

	IP	AN	EV	IF	SR	Total
Number of Items	12	12	12	12	12	60
U.S. (*N* = 210)						
Mean	7.37	6.74	6.38	6.29	6.61	33.39
SD	2.60	2.63	2.17	2.65	2.81	10.68
Reliability	0.70	0.71	0.51	0.67	0.70	0.90
Correlation						
IP		0.75	0.56	0.66	0.63	0.87
AN			0.54	0.64	0.59	0.85
EV				0.59	0.56	0.77
IF					0.57	0.84
SR						0.82
Japan (*N* = 200)						
Mean	7.48	6.74	5.64	5.78	5.48	31.11
SD	2.20	2.10	1.79	2.28	2.42	7.86
Reliability	0.57	0.58	0.28	0.54	0.59	0.82
Correlation						
IP		0.46	0.21	0.50	0.39	0.72
AN			0.36	0.51	0.43	0.75
EV				0.39	0.35	0.60
IF					0.47	0.80
SR						0.75

Note. IP = Interpretation, AN = Analysis, EV = Evaluation, IF = Inference, SR = Self-Regulation.

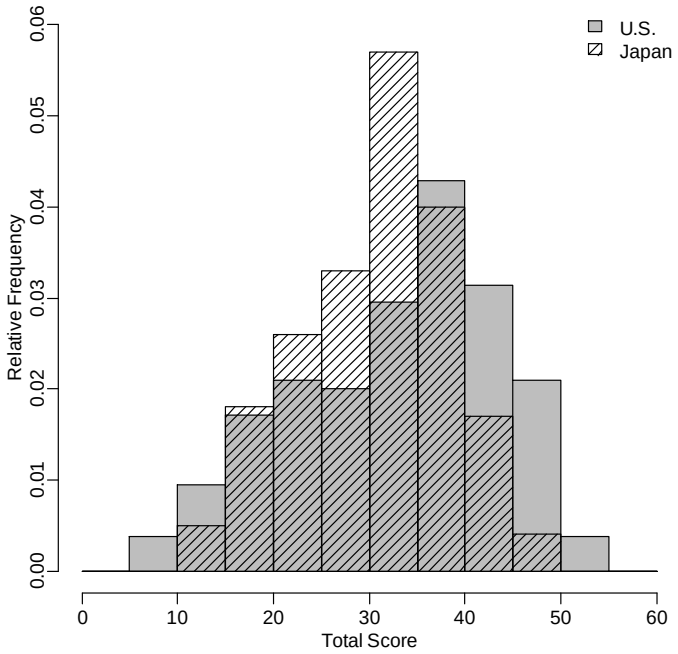

Figure 1. Total score distributions for the U.S. (shown in gray bars) and Japan (shown in shaded bars). The vertical axis represents relative frequency.

In addition to the subscale scores, scores by controversy were also examined (Table 2). Again, the U.S. data yielded higher means, standard deviations, reliabilities, and intercorrelations than the Japanese data. Controversy #6 was most difficult among all controversies for both groups of students, but it was especially difficult for the Japanese students with the average score being 3.78.

Testlet effects were also examined for the controversy scores. A testlet effect is present if each testlet (i.e., items in each controversy) measures something specific to that testlet in addition to the target construct. A bootstrap method proposed by Zenisky, Hambleton, and Sireci (2002) was applied to both U.S. and Japanese datasets. For the U.S. data, the observed testlet reliability was .87 (p = 0). In the Japanese data, the observed testlet reliability was .79 (p = .005). Both of these estimates were significantly higher than the corresponding means of "randomly constructed" testlet reliabilities. Thus, there was an indication of testlet effects for both datasets; it is likely that some examinees had advantages (or disadvantages) over other examinees on particular controversies beyond their critical thinking ability.

3.2 Item Analysis

Figures 2 and 3 show difficulty (proportion correct) and discrimination (point biserial correlations) for each item for the U.S. and Japanese versions, respectively (detailed item statistics are shown in Tables A1 and A2 in the Appendix). In these figures, items

Table 2. Summary of controversy scores.

	Controversy					
	1	2	3	4	5	6
Number of Items	10	10	10	10	10	10
U.S. (*N* = 210)						
Mean	6.15	5.33	6.22	5.97	5.22	4.50
SD	2.03	2.30	2.43	2.33	2.13	2.42
Reliability	0.55	0.61	0.70	0.66	0.56	0.71
Correlation						
1		0.51	0.53	0.61	0.38	0.47
2			0.51	0.55	0.47	0.52
3				0.65	0.45	0.60
4					0.53	0.62
5						0.57
Japan (*N* = 200)						
Mean	6.07	4.92	4.97	5.88	5.51	3.78
SD	1.80	1.79	1.73	1.94	2.04	1.91
Reliability	0.41	0.37	0.41	0.50	0.57	0.52
Correlation						
1		0.36	0.41	0.31	0.30	0.38
2			0.37	0.37	0.31	0.36
3				0.42	0.43	0.43
4					0.46	0.45
5						0.47

in controversy #1 are numbered as 1 through 10, items in controversy #2 are numbered as 11 through 20, and so forth. The average difficulty over all items was .56 and .52 for the U.S and Japan, respectively. Thus, the test was slightly more difficult for Japanese students. Overall, item difficulties show very similar patterns between the U.S. and Japan except for a few cases. Items 13, 25, 26, 29, and 44 showed relatively large discrepancies between the U.S. and Japan. All of these items but item 44 were much more difficult for Japanese students than for U.S. students. In opposite, nearly 80% of the Japanese students were correct on item 44 while only slightly less than 50% of the U.S. students were.

The average item discrimination over all items was .39 and .29 for the U.S and Japan, respectively. The U.S. version generally yielded higher discriminations than the Japanese version. The overall pattern agrees quite well except for several items which indicated large discrepancies with respect to item discrimination. Especially, items 12,

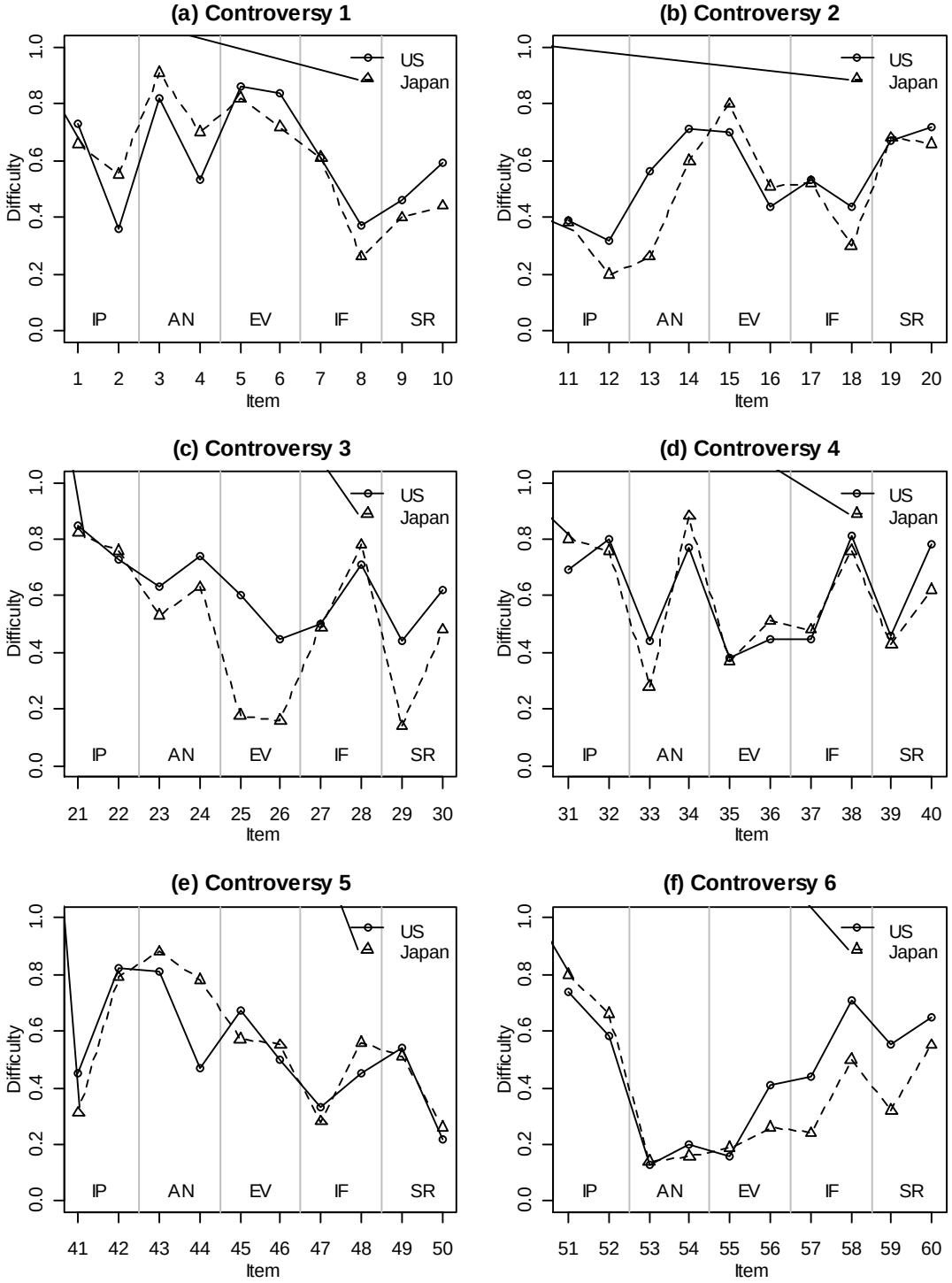

Figure 2. Item difficulty by controversy and skill. IP = Interpretation, AN = Analysis, EV = Evaluation, IF = Inference, and SR = Self-Regulation.

18, 26, 29, and 33 in the Japanese version had negative or zero discrimination. Among these items, items 12, 18, and 29 had another response option that behaved like a correct response, having a relatively large positive correlation with the total score (see Table A2). These distractors might inadvertently attract Japanese students with higher

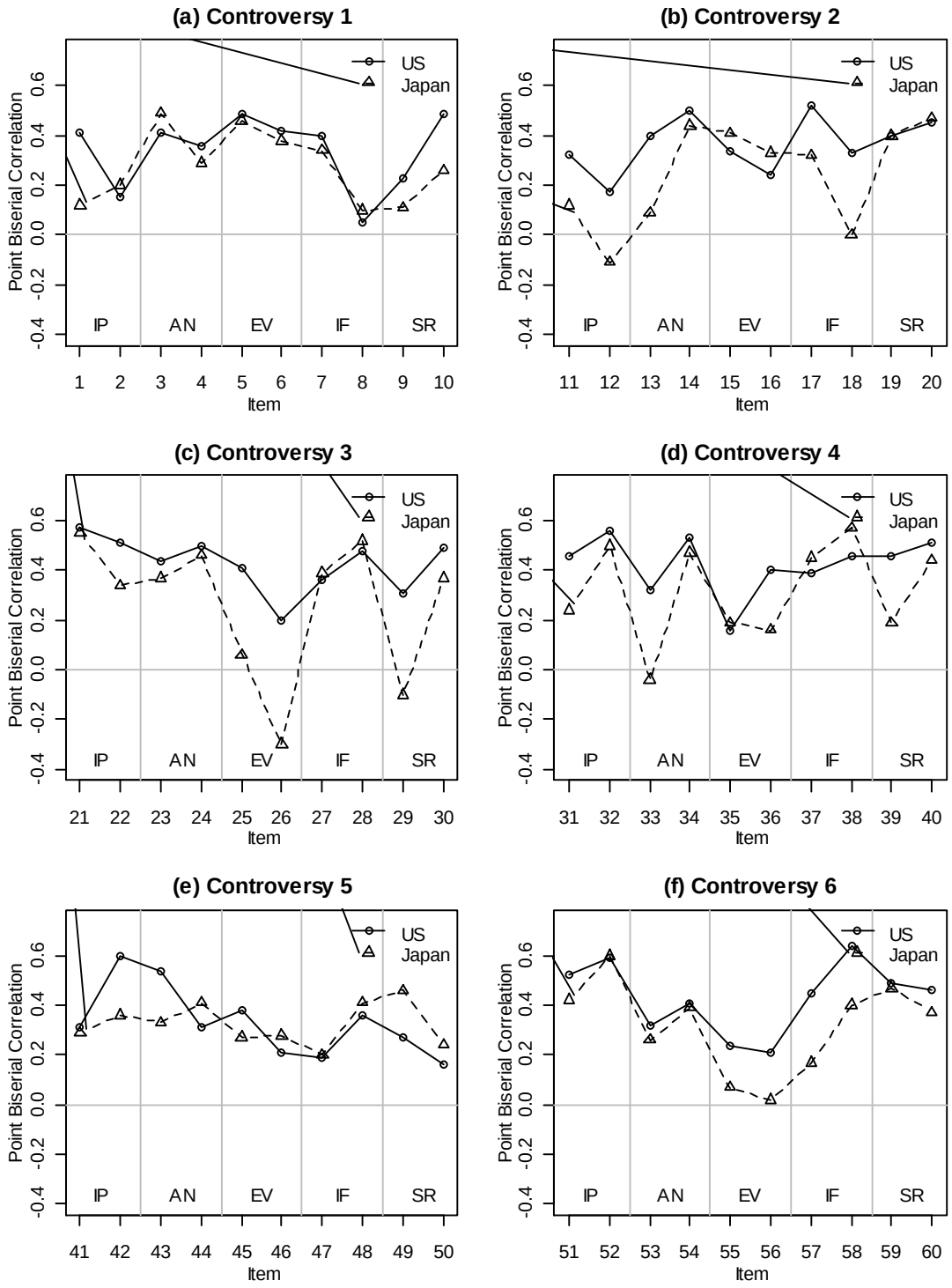

Figure 3. Item discrimination by controversy and skill. IP = Interpretation, AN = Analysis, EV = Evaluation, IF = Inference, and SR = Self-Regulation.

critical thinking ability. In spite of these discrepancies, more than one third of the items showed very similar characteristics in both versions. As a result, the correlation of item difficulty between the U.S. and Japan was .81 over the 60 items, and that of item discrimination was .60. Notably, items in controversy #3, especially items 25, 26,

Table 3. Summary of prior opinions and corresponding controversy scores (Japanese data only)

Contro-Versy	Proportion[a]				Mean (SD) Controversy Score[a]				ANOVA[b]	
	1	2	3	4	1	2	3	4	*F*	*P*
1	.08	.29	.49	.15	5.75 (2.05)	5.98 (2.01)	6.18 (1.72)	6.00 (1.54)	0.35	0.79
2	.22	.26	.31	.22	4.79 (1.73)	4.75 (1.79)	4.85 (1.86)	5.35 (1.76)	1.08	0.36
3	.06	.09	.24	.62	5.36 (1.63)	4.56 (1.62)	5.50 (1.68)	4.78 (1.74)	2.59	0.05
4	.16	.12	.32	.41	5.84 (1.51)	6.22 (1.48)	5.89 (2.18)	5.78 (2.02)	0.31	0.82
5	.13	.08	.29	.51	4.77 (1.86)	4.50 (1.93)	5.65 (1.99)	5.77 (2.07)	3.18	0.03
6	.29	.33	.18	.21	3.88 (1.83)	3.38 (1.93)	4.17 (1.90)	3.93 (1.94)	1.57	0.20

[a] Opinion ratings were made on the four-point scale with 1 indicating strong agreement with Statement A and 4 indicating strong agreement with Statement B.
[b] The degrees of freedom were (3,196) for all comparisons.

and 29, showed discrepancies larger than those in other controversies in both difficulty and discrimination.

Given the above results, controversies #2 and #3 were excluded and the test and item statistics were re-examined. However, substantial changes were not observed in terms of item discriminations and reliabilities.

3.3 Relation to Examinee Characteristics and Prior Opinions

Additional information (i.e., examinee characteristics and prior opinions about the controversy topics) was available in the Japanese data. Table 3 shows the summary of prior opinions about the controversy topics and controversy score comparisons by prior opinion. Response proportions of opinion questions (columns 2 through 5) varied across controversies. Opinions about controversies #2 and #6 were distributed more uniformly than other controversies, while those about controversies #3 and #5 were highly disproportionate. Columns 6 though 9 in Table 3 show the means and standard deviations of controversy scores. Controversy scores were subjected to ANOVA in order to see whether they were affected by prior opinions. Difference by prior opinions was found in controversy #5 ($F = 3.18$, $p = .03$), in which examinees

Table 4. Means and standard deviations of the total score by examinee characteristics.

Grade	Fresh-man	Sopho-More	Junior	Senior			
Mean	29.52	31.28	30.21	34.56			
SD	7.49	8.96	7.91	5.94			
N	81	54	24	41			
Age	18	19	20	21	22	23	24
Mean	28.62	30.62	31.47	33.10	33.87	27.80	39.50
SD	7.72	8.08	6.43	7.84	7.79	10.85	2.12
N	39	68	32	39	15	5	2
Gender	Male	Female					
Mean	30.86	31.39					
SD	8.22	7.50					
N	104	96					
Major	Humanity	Science					
Mean	31.12	31.05					
SD	8.03	7.21					
N	162	38					

Table 5. Score comparison by examinee characteristics

	F	*p*	R^2
Grade	9.666	.002	.047
Age	7.127	.008	.035
Gender	0.226	.635	.001
Major	0.003	.960	<.001
Achievement	18.202	<.001	.084

who agreed with Statement B tended to score higher. This may threaten the neutrality of the controversy #5 and the corresponding test items. No significant differences were found in the other controversies.

Finally, the total score was analyzed by examinee characteristics for the Japanese data (see Table 4 and Figure 4). Effects of these characteristics on the test score were examined by linear regression analysis. Table 5 shows the results of *F*-tests; there was no significant difference for gender and academic major (humanity vs. science). However, females and science-major examinees tended to have smaller variability than their respective counterparts. Age, grade, and achievement level had significant positive correlations with the total score (.18, .22, and .29, respectively), although the

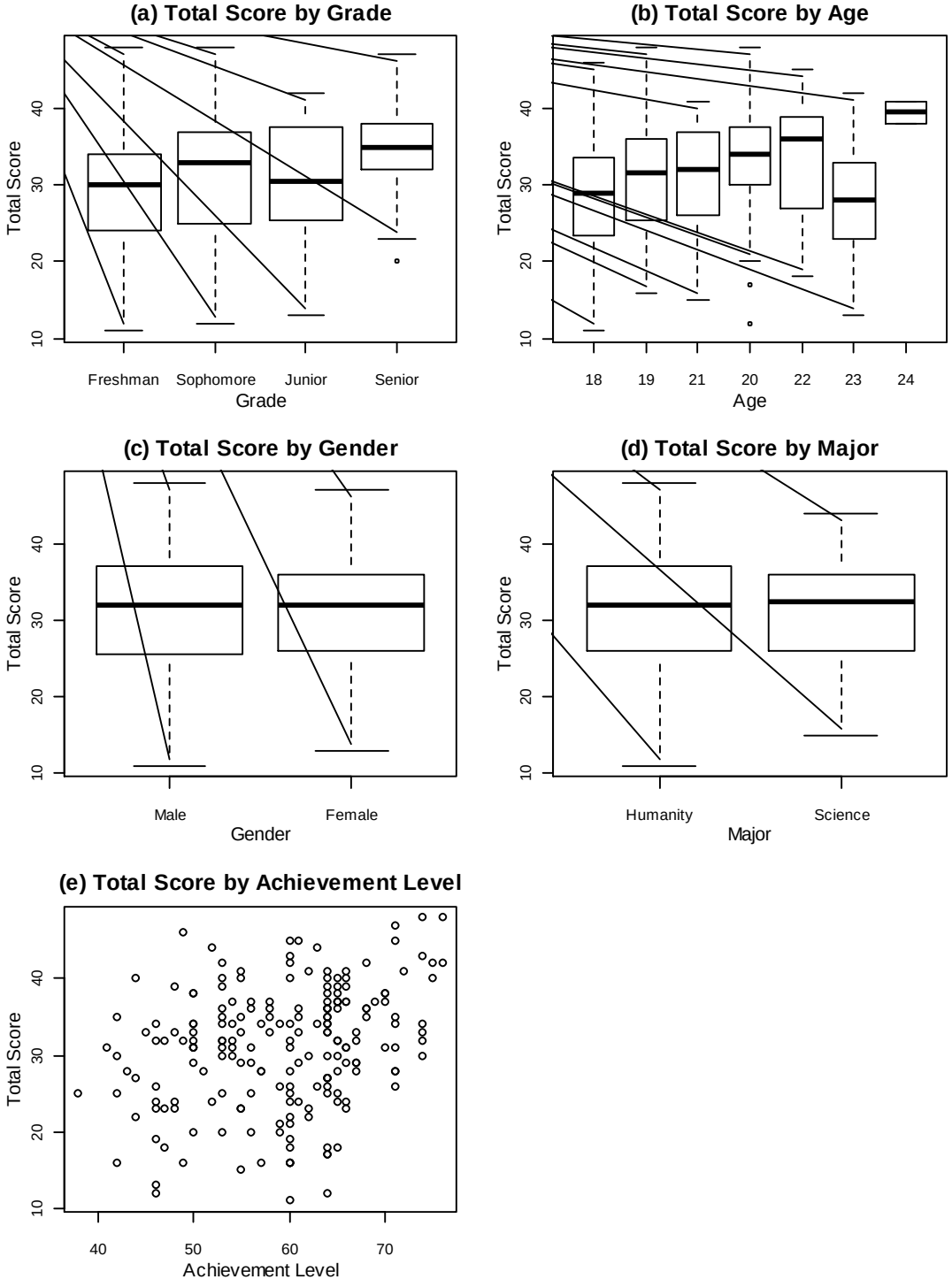

Figure 4. Plots of the total score by examinee characteristics.

magnitude of these correlations was small. The corresponding R^2s are shown in the last column of Table 5.

4. Conclusions

The U.S. and Japanese versions of MTCT-II both had high reliability with respect to the total score. The results of factor analysis suggested that the MTCT-II items were roughly unidimensional. Thus, both versions of the test would serve as an appropriate measure of critical thinking if their validity is further established. There was an indication of testlet effects for both U.S. and Japanese versions; we should keep it in mind that examinee scores likely depend not only on their critical thinking ability but also on knowledge or familiarity to particular topics chosen for the controversies.

The Japanese version of MTCT-II was characterized by lower reliability and discrimination than the U.S. version. Also, Japanese students tended to score lower than the U.S. students, and the dispersion of scores was smaller. These discrepancies can be attributed to several factors.

First, the U.S. data likely included a group of graduate students, who might score higher than college students. This might increase the true score variability of the U.S. data, leading to the higher reliability estimates. It also explains the higher means and standard deviations of the U.S. data. Second, most Japanese students are, in general, not very familiar with the topics discussed in the MTCT-II controversies. This unfamiliarity might increase their cognitive workload and introduce additional construct-irrelevant variation to their test scores. Third, several items in the Japanese version showed near-zero or negative discriminations. These items had adverse effects on test reliability. Finally, even though the Japanese translation was made very carefully, it might affect subtle aspects of the controversies and items which are relevant to invoking students' critical thinking (in other words, some Japanese students might take the test as if it had been an usual reading comprehension test).

The above conjectures are not the only possibilities. Revealing what really made these differences will need detailed analysis of item contents and how examinees approached these items. Especially, what happened in the items that showed negative discriminations in the Japanese version should be examined in more detail in the future. No common features are obvious in these items, but the difference could be attributed to some cultural difference in thinking style (one item requires knowledge which is supposedly common to U.S. people for a correct response).

In spite of these differences, overall patterns of discrimination and difficulty across individual items were, except for a few cases, very similar between the U.S. and Japan; most of the MTCT-II items worked in almost the same manner in both countries. This implies that the logical structures presented in the controversies and the corresponding reasoning questions are fairly equally applicable and generalizable to examinees with different cultural background.

Finally, effects of prior opinions and examinee characteristics on the MTCT-II score

were examined for the Japanese version. Prior opinions did not affect the test score except for one controversy; this supports the result by Edman et al. (2002). Thus, the test is impartial for most parts, but further refinement may be necessary to avoid biases due to examinees' prior opinions. Gender and academic major had no influence on the test score, and this indicates that the test is almost unbiased in terms of these factors (although the score variances differed by both gender and major). Grade, age, and achievement level had positive correlations with the test score. This may provide partial support for the validity of MTCT-II, but the magnitude of the correlations was small.

REFERENCES

American Philosophical Association. 1990. *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. The Delphi Report: Research findings and recommendations prepared for the committee on pre-college philosophy.* (ERIC Document Reproduction Service, No. ED315-423)

Australian Education Council Mayer Committee. 1992. Key competencies. *Report of the Committee to Advise the Australian Education Council and Ministers of Vocational Education, Employment, and Training on Employment-Related Key Competencies for Postcompulsory Education and Training*. Australian Education Council and Ministers of Vocational Education, Employment, and Training. Canberra, Australia.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., & Rumble, M. 2010.*Draft white paper 1: Defining 21st century skills.* Retrieved January 5, 2013, from http://atc21s.org/wp-content/uploads/2011/11/1-Defining-21st-Century-Skills.pdf.

Council for Aid to Education. n.d. Collegiate Learning Assessment. http://www.collegiatelearningassessment.org/. Accessed January 5, 2013.

Edman, L. R. O., Robey, J., & Bart, W. M. 2002. *Critical thinking, belief bias, epistemological assumptions, and the Minnesota Test of Critical Thinking.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Ennis, R. H., & Weir, E. 1985. *The Ennis-Weir critical thinking essay test.* Pacific Grove, CA: Midwest.

Facione, P. A. 1990. *The California Critical Thinking Skills Test (CCTST): Forms A and B; and the CCTST test manual.* Millbrae, CA: California Academic Press.

Hirayama, R., Tanaka, Y., Kawasaki, M., & Kusumi, T. 2010. Development and evaluation of a critical thinking ability scale from Cornell Critical Thinking

Test Level Z. *Japan Journal of Educational Technology*, 33, 441-448. [in Japanese]

Kuhara, K., Inoue, N., & Hatano, G. 1983. Construction and validation of a test for assessing critical thinking ability. *The Science of Reading*, 27, 131-142. [in Japanese]

Kusumi, T., Koyasu, M., & Michita, Y. 2011. *Developing critical thinking in higher education*. Tokyo: Yuhikaku. [in Japanese]

National Centre for Vocational Education Research. 2003. *Defining generic skills: At a glance*. Adelaide, Australia: National Centre for Vocational Education Research.

O'Neil, H. F., Allred, K., & Baker, E. 1997. Review of workforce readiness theoretical frameworks. In H. F. O'Neil (Ed.), *Workforce readiness: Competencies and assessments*. NJ: Lawrence Erlbaum.

R Development Core Team. 2012. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Watson, G. B., & Glaser, E. M. 1994. *Watson-Glaser critical thinking appraisal Form S manual*. San Antonio, TX: Harcourt Brace.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. 2002. Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, 39, 291-309.

ACKNOWLEDGEMENT

The author thanks William Bart, Laird Edman, and Jennifer Robey for kindly providing MTCT-II items and their U.S. data for this study. The author also thanks Reiko Nakata for the translation into Japanese, and other staff of the assessment development unit at Benesse Corporation for data collection and cleaning.

APPENDIX

Table A1. Result of the item analysis for the U.S. version of MTCT-II

			Response Proportion				Point Biserial Correlation			
Item	N_V	N_M	A	B	C	D	A	B	C	D
1	210	0	0.07	**0.73**	0.15	0.06	-0.12	**0.41**	-0.31	-0.18
2	210	0	0.06	0.55	**0.36**	0.03	-0.12	-0.01	**0.15**	-0.22
3	210	0	0.04	**0.82**	0.08	0.06	-0.19	**0.41**	-0.30	-0.16
4	210	0	**0.53**	0.40	0.03	0.04	**0.36**	-0.17	-0.26	-0.25
5	210	0	0.06	**0.86**	0.05	0.03	-0.32	**0.49**	-0.26	-0.22
6	209	1	0.04	**0.84**	0.09	0.03	-0.26	**0.42**	-0.18	-0.29
7	208	2	0.14	0.07	0.18	**0.61**	-0.27	-0.24	-0.10	**0.40**
8	209	1	0.04	0.55	**0.37**	0.03	-0.26	0.15	**0.05**	-0.26
9	209	1	**0.46**	0.09	0.05	0.40	**0.23**	-0.26	-0.32	0.06
10	208	2	0.11	0.11	**0.59**	0.20	-0.23	-0.32	**0.49**	-0.18
11	202	8	0.09	0.32	**0.39**	0.21	-0.14	-0.17	**0.32**	-0.09
12	202	8	**0.32**	0.49	0.12	0.07	**0.17**	0.09	-0.25	-0.15
13	202	8	**0.56**	0.20	0.08	0.16	**0.40**	-0.16	-0.23	-0.19
14	201	9	0.12	0.07	0.10	**0.71**	-0.19	-0.33	-0.27	**0.50**
15	202	8	0.11	**0.70**	0.11	0.08	-0.35	**0.34**	-0.09	-0.07
16	209	1	**0.44**	0.17	0.22	0.18	**0.24**	-0.23	-0.19	0.11
17	209	1	0.17	0.17	**0.53**	0.14	-0.31	-0.37	**0.52**	-0.02
18	208	2	0.09	**0.44**	0.08	0.39	-0.34	**0.33**	-0.24	-0.01
19	209	1	**0.67**	0.04	0.25	0.04	**0.40**	-0.30	-0.15	-0.31
20	209	1	0.12	0.06	**0.72**	0.10	-0.19	-0.29	**0.45**	-0.23
21	210	0	**0.85**	0.05	0.06	0.04	**0.57**	-0.38	-0.31	-0.24
22	207	3	0.12	0.09	**0.73**	0.06	-0.23	-0.39	**0.51**	-0.18
23	210	0	0.14	**0.63**	0.05	0.18	-0.26	**0.44**	-0.37	-0.11
24	208	2	0.12	0.11	**0.74**	0.03	-0.25	-0.31	**0.50**	-0.24
25	205	5	**0.60**	0.07	0.09	0.24	**0.41**	-0.19	-0.36	-0.12
26	209	1	0.33	**0.45**	0.14	0.08	-0.26	**0.20**	-0.03	0.14
27	208	2	0.11	0.11	**0.50**	0.29	-0.29	-0.18	**0.36**	-0.08
28	208	2	0.12	0.12	**0.71**	0.05	-0.20	-0.28	**0.48**	-0.30
29	208	2	0.07	0.16	**0.44**	0.33	-0.39	-0.09	**0.31**	-0.05
30	210	0	0.19	0.13	0.06	**0.62**	-0.19	-0.16	-0.45	**0.49**

Note. N_V = number of valid responses; N_M = number of missing responses. Numbers in bold indicate the correct response options. Items are numbered so that items 1 through 10 correspond to those in controversy #1, items 11 through 20 to those in controversy #2, and so forth. Items 53 and 54 were treated as dichotomous items.

Table A1 (cont.). Result of the item analysis for the U.S. version of MTCT-II

			Response Proportion				Point Biserial Correlation			
Item	N_V	N_M	A	B	C	D	A	B	C	D
31	209	1	0.19	**0.69**	0.05	0.07	-0.23	**0.46**	-0.11	-0.38
32	209	1	**0.80**	0.11	0.04	0.05	**0.56**	-0.35	-0.34	-0.22
33	207	3	0.25	0.21	**0.44**	0.10	-0.02	-0.27	**0.32**	-0.12
34	208	2	**0.77**	0.06	0.12	0.04	**0.53**	-0.21	-0.35	-0.26
35	205	5	0.15	0.38	0.09	**0.38**	-0.26	0.19	-0.27	**0.16**
36	209	1	0.34	0.09	**0.45**	0.11	-0.06	-0.39	**0.40**	-0.18
37	208	2	0.12	0.22	0.20	**0.45**	-0.28	-0.16	-0.08	**0.39**
38	209	1	**0.81**	0.13	0.05	0.01	**0.46**	-0.24	-0.35	-0.22
39	209	1	0.32	0.14	**0.46**	0.09	-0.14	-0.33	**0.46**	-0.17
40	202	8	0.07	**0.78**	0.06	0.09	-0.29	**0.51**	-0.38	-0.15
41	209	1	0.13	0.36	0.06	**0.45**	-0.22	-0.03	-0.27	**0.31**
42	209	1	**0.82**	0.05	0.05	0.07	**0.60**	-0.30	-0.41	-0.28
43	207	3	0.08	**0.81**	0.09	0.02	-0.31	**0.54**	-0.38	-0.12
44	209	1	0.27	0.13	**0.47**	0.12	-0.02	-0.32	**0.31**	-0.11
45	209	1	**0.67**	0.14	0.09	0.10	**0.38**	-0.19	-0.08	-0.29
46	209	1	0.15	0.13	0.22	**0.50**	-0.07	-0.02	-0.17	**0.21**
47	208	2	**0.33**	0.27	0.36	0.05	**0.19**	0.00	-0.10	-0.19
48	206	4	0.12	0.13	0.31	**0.45**	-0.21	-0.38	0.03	**0.36**
49	209	1	0.08	**0.54**	0.22	0.16	-0.23	**0.27**	-0.01	-0.19
50	209	1	0.20	0.30	0.28	**0.22**	-0.39	0.20	-0.01	**0.16**
51	205	5	0.11	0.10	**0.74**	0.05	-0.25	-0.24	**0.52**	-0.35
52	205	5	0.18	0.08	0.16	**0.58**	-0.16	-0.43	-0.30	**0.59**
53	210	0	0.87	**0.13**			-0.32	**0.32**		
54	210	0	0.80	**0.20**			-0.41	**0.41**		
55	206	4	**0.16**	0.25	0.15	0.44	**0.24**	-0.23	-0.31	0.24
56	207	3	0.31	0.21	**0.41**	0.07	0.09	-0.27	**0.21**	-0.12
57	207	3	0.32	0.13	0.11	**0.44**	-0.05	-0.29	-0.34	**0.45**
58	207	3	0.14	**0.71**	0.08	0.06	-0.40	**0.64**	-0.24	-0.36
59	205	5	0.07	0.24	0.13	**0.55**	-0.28	-0.14	-0.32	**0.49**
60	204	6	0.16	0.13	**0.65**	0.06	-0.17	-0.33	**0.46**	-0.19

Note. N_V = number of valid responses; N_M = number of missing responses. Numbers in bold indicate the correct response options. Items are numbered so that items 1 through 10 correspond to those in controversy #1, items 11 through 20 to those in controversy #2, and so forth. Items 53 and 54 were treated as dichotomous items.

Table A2. Result of the item analysis for the Japanese version of MTCT-II

			Response Proportion				Point Biserial Correlation			
Item	N_V	N_M	A	B	C	D	A	B	C	D
1	199	1	0.06	**0.66**	0.23	0.05	-0.04	**0.12**	-0.12	0.00
2	200	0	0.08	0.36	**0.55**	0.00	-0.20	-0.08	**0.20**	-0.14
3	200	0	0.02	**0.91**	0.06	0.02	-0.21	**0.49**	-0.37	-0.22
4	200	0	**0.70**	0.28	0.00	0.02	**0.29**	-0.16	-0.15	-0.32
5	200	0	0.05	**0.82**	0.06	0.06	-0.29	**0.46**	-0.23	-0.23
6	200	0	0.04	**0.72**	0.21	0.03	-0.37	**0.38**	-0.24	0.00
7	200	0	0.08	0.08	0.23	**0.61**	-0.28	-0.23	-0.06	**0.34**
8	200	0	0.04	0.68	**0.26**	0.02	-0.19	0.02	**0.10**	-0.15
9	200	0	**0.40**	0.08	0.04	0.48	**0.11**	-0.33	-0.36	0.20
10	200	0	0.09	0.12	**0.44**	0.36	-0.14	-0.23	**0.26**	-0.04
11	200	0	0.06	0.29	**0.38**	0.27	-0.26	-0.11	**0.12**	0.12
12	200	0	**0.20**	0.64	0.08	0.08	**-0.11**	0.30	-0.38	0.02
13	200	0	**0.26**	0.10	0.18	0.45	**0.09**	-0.07	0.04	-0.07
14	200	0	0.22	0.12	0.06	**0.60**	-0.20	-0.38	-0.04	**0.44**
15	200	0	0.07	**0.80**	0.06	0.06	-0.21	**0.41**	-0.06	-0.38
16	200	0	**0.51**	0.18	0.20	0.12	**0.33**	-0.30	-0.12	-0.01
17	200	0	0.21	0.12	**0.52**	0.14	-0.12	-0.28	**0.32**	-0.06
18	200	0	0.06	**0.30**	0.08	0.55	-0.27	**0.00**	-0.25	0.26
19	200	0	**0.68**	0.05	0.24	0.02	**0.40**	-0.32	-0.19	-0.20
20	200	0	0.14	0.06	**0.66**	0.13	-0.19	-0.29	**0.47**	-0.24
21	200	0	**0.82**	0.08	0.06	0.04	**0.55**	-0.44	-0.26	-0.15
22	200	0	0.08	0.12	**0.76**	0.05	-0.33	-0.08	**0.34**	-0.14
23	200	0	0.19	**0.53**	0.08	0.20	-0.02	**0.37**	-0.36	-0.19
24	200	0	0.07	0.26	**0.63**	0.04	-0.04	-0.36	**0.46**	-0.29
25	200	0	**0.18**	0.12	0.18	0.52	**0.06**	-0.16	-0.21	0.22
26	200	0	0.56	**0.16**	0.18	0.10	0.17	**-0.30**	0.03	0.05
27	200	0	0.05	0.18	**0.49**	0.28	-0.19	-0.26	**0.39**	-0.12
28	200	0	0.08	0.12	**0.78**	0.03	-0.33	-0.33	**0.52**	-0.13
29	200	0	0.11	0.32	**0.14**	0.44	-0.28	-0.02	**-0.10**	0.27
30	199	1	0.30	0.17	0.06	**0.48**	-0.21	-0.04	-0.32	**0.37**

Note. N_V = number of valid responses; N_M = number of missing responses. Numbers in bold indicate the correct response options. Items are numbered so that items 1 through 10 correspond to those in controversy #1, items 11 through 20 to those in controversy #2, and so forth. Items 53 and 54 were treated as dichotomous items.

Table A2 (cont.). Result of the item analysis for the Japanese version of MTCT-II

			Response Proportion				Point Biserial Correlation			
Item	N_V	N_M	A	B	C	D	A	B	C	D
31	200	0	0.12	**0.80**	0.05	0.04	-0.11	**0.24**	-0.23	-0.06
32	200	0	**0.76**	0.10	0.06	0.09	**0.50**	-0.26	-0.30	-0.24
33	199	1	0.39	0.17	**0.28**	0.16	0.18	-0.21	**-0.04**	0.02
34	200	0	**0.88**	0.06	0.06	0.02	**0.47**	-0.33	-0.26	-0.17
35	200	0	0.23	0.21	0.19	**0.37**	-0.04	0.01	-0.20	**0.19**
36	200	0	0.31	0.04	**0.51**	0.14	0.03	-0.23	**0.16**	-0.12
37	199	1	0.07	0.31	0.15	**0.48**	-0.24	-0.21	-0.18	**0.45**
38	200	0	**0.76**	0.13	0.06	0.04	**0.57**	-0.32	-0.33	-0.29
39	200	0	0.32	0.12	**0.43**	0.14	0.05	-0.17	**0.19**	-0.17
40	199	1	0.12	**0.62**	0.09	0.18	-0.24	**0.44**	-0.30	-0.14
41	200	0	0.30	0.33	0.06	**0.31**	-0.11	-0.06	-0.25	**0.29**
42	200	0	**0.79**	0.06	0.08	0.08	**0.36**	-0.20	-0.25	-0.13
43	200	0	0.04	**0.88**	0.04	0.04	-0.01	**0.33**	-0.25	-0.27
44	200	0	0.06	0.07	**0.78**	0.08	-0.24	-0.27	**0.41**	-0.16
45	200	0	**0.57**	0.22	0.14	0.07	**0.27**	-0.18	0.01	-0.26
46	200	0	0.08	0.08	0.30	**0.55**	-0.20	-0.29	-0.02	**0.28**
47	200	0	**0.28**	0.34	0.35	0.02	**0.20**	0.04	-0.14	-0.28
48	200	0	0.08	0.16	0.20	**0.56**	-0.25	-0.20	-0.16	**0.41**
49	199	1	0.15	**0.51**	0.18	0.17	-0.34	**0.46**	-0.10	-0.20
50	199	1	0.12	0.47	0.15	**0.26**	-0.24	0.04	-0.14	**0.24**
51	199	1	0.03	0.15	**0.80**	0.03	-0.19	-0.26	**0.42**	-0.27
52	199	1	0.12	0.09	0.13	**0.66**	-0.36	-0.33	-0.21	**0.60**
53	200	0	0.86	**0.14**			-0.26	**0.26**		
54	200	0	0.84	**0.16**			-0.39	**0.39**		
55	198	2	**0.19**	0.22	0.18	0.40	**0.07**	-0.13	-0.17	0.18
56	198	2	0.30	0.33	**0.26**	0.11	0.15	-0.06	**0.02**	-0.15
57	198	2	0.52	0.16	0.09	**0.24**	0.18	-0.24	-0.25	**0.17**
58	197	3	0.23	**0.50**	0.19	0.09	-0.15	**0.40**	-0.19	-0.22
59	197	3	0.13	0.42	0.13	**0.32**	-0.34	-0.07	-0.22	**0.47**
60	197	3	0.22	0.09	**0.55**	0.14	-0.07	-0.29	**0.37**	-0.21

Note. N_V = number of valid responses; N_M = number of missing responses. Numbers in bold indicate the correct response options. Items are numbered so that items 1 through 10 correspond to those in controversy #1, items 11 through 20 to those in controversy #2, and so forth. Items 53 and 54 were treated as dichotomous items.