Psychometric assessment of ICILS test items on Hong Kong and Korean students – A Rasch analysis

Oon Pey Tee; Nancy Law; Kim Soojin; Sungsook Kim; Tse See Ki

ABSTRACT

The IEA International Computer and Information Literacy Study (ICILS) 2013 is the first international comparative study aims to examine the outcome of Grade 8 students' computer and information literacy (CIL) across 19 countries. The measurement of CIL focuses on students' ability to use computer within and outside of school in the information age. Students are asked to complete two test modules, out of a total of four, which are conducted exclusively on computers. The present study reports psychometric properties of the test modules based on the results garnered from the Field Trial which was in preparation for the main study. A total sample of 804 students from Hong Kong and South Korea participated in the Field Trial. Rasch-anchored analysis was employed to analyze the data. The items for the four test modules demonstrated adequate psychometric properties – they target the students from Hong Kong and Korea well. However, significant Differential Item Functioning (DIF) was detected in some items from the 4 test modules. Future work on analysis on items reporting significant DIF will be carried out.

Key words: ICILS, computer and information literacy, psychometric assessment, Rasch analysis

BACKGROUND

The International Computer and Information Literacy Study (ICILS) is an international comparative study aims to assess students' computer and information literacy (CIL) level. It examines the extent to which students are able to use information and communication technology (ICT) in the digital age. Ultimately, Grade 8 students' CIL across 19 countries will be compared against each other and the international means. The results from ICILS will shed light to policy makers and stakeholders in improving students' CIL in the digital age.

There are two approaches in measuring CIL of students. One approach is to measure learning area-specific outcomes use of computer, for example, the use of ICT to solve science problems. The second approach is to measure ICT achievement as a discrete learning area without anchoring at any subjects. The former assumes that ICT achievement is inseparable from subject-based achievement while the latter assumes that ICT achievement is not limited by any disciplines and comprises a set of skills that are transferable to new contexts. ICILS adopts the second approach.

The ICILS instrument intended to measure students' CIL encompassing items that are delivered in 30-minute test module. There are four modules in total and each student is requested to complete two modules in one test session.

PURPOSE

The purpose of this paper is to ascertain the psychometric properties of the items for the four modules in respect of the following Research Questions (RQ):
RQ1: Do the items target the Hong Kong and Korean students well?
RQ2: Do the items show the property of measurement invariance across relevant subsamples – students from Hong Kong and Korea?

SIGNIFICANCE OF RESEARCH

Only three Asian countries out of the 19 countries participated in the ICILS study. It would thus be of interest to examine whether the items demonstrate non-bias and invariance properties on Asians students. The results provide insights from an Asian perspective.

THEORETICAL FRAMEWORK

The Rasch model has been gaining its popularity in examining item quality for international comparative test items (eg. Glynn, 2012). The Third International Mathematics and Science Study (TIMSS), conducted by the IEA, is the first international comparative large scale assessments that fully applied the Rasch model (Wendt, Bos, & Goy, 2011). It is a probabilistic model that converts ordinal scores into interval measures (Rasch, 1960) – even the simplest statistics, such as mean and standard deviation, demand this linearity of scores (Wright & Masters, 1982). Observed data that satisfied the model promises interval level measurement and this has made it prominent in many international comparative large scale assessments of educational achievement (Wendt, Bos, & Goy, 2011). Attributes that are particularly important to cross-national comparisons are whether the test items targeting students from different countries well, and whether the test items demonstrate similar item difficulties across countries.

For situating our study as mentioned in the foregoing, we review a key area of the scholarship: the treatment of test items using of Rasch framework. We examine the extent to which the ICILS test items target the Hong Kong and Korean students well. If this attribute is recognized from the data, the item difficulty and student ability estimates should spread evenly along a standardized linear scale (Bond & Fox, 2007; Oon & Subramaniam, 2011a; 2011b; Oon & Subramaniam, 2013). Besides, the invariance property of the items has been examined too. If the invariance property is evident, the item estimates should remain stable, showing non-significant differential item functioning

(DIF), across the relevant sub-samples (Bond & Fox, 2007; Oon & Subramaniam, 2011a; 2011b; Oon & Subramaniam, 2013).

METHODS AND STATISTICAL TECHNIQUES

The test instruments which are designed to measure CIL encompassed questions and tasks that are delivered in four 30-minute modules. Each student answered two test modules which were randomly assigned to them.

Each module consisted three types of questions, they are:

(a) Information-based response tasks: It is typically a non-interactive representation of a computer-based problem that makes use of technology simply to record students' responses. The response formats include the multiple choices, the constructed response as well as the drag-and –drop questions.
(b) Skills tasks: It requires students to use interactive simulations of software or universal applications to complete the tasks. It consists either a linear (eg. Open a file) or non-linear skills tasks (eg. Select an image then copy and paste the image).
(c) Authoring tasks: It requires students to modify and create information using authentic software applications. Students' work will be automatically saved for subsequent assessments which will be scored by scorers according to a prescribed set of rubrics.

As mentioned, some questions are automatically scored while the others were scored by trained scorers, recruited by the ICILS National Research Centre in Hong Kong and Korea, based on a prescribed set of rubrics. For the latter, some items were scored from 0 to 3 while some from 0 to 4, with 0 indicating no credit. The higher the score indicating the more satisfaction is the answer. The converse holds true for the lower score. The '*Not reached*' and '*Not administered / missing by design*' responses were coded as missing values while '*Presented but not answered*' responses were coded as zero.

Rasch framework, anchored on Master's Partial Credit Model (PCM), was employed to analyze the data using WINSTEPS software version 3.68.1 (Linacre, 2009). A PCM was used as each item type, as mentioned above, contains different thresholds as the result of different rating scale (Bond & Fox, 2007).

DATA SOURCES

The data for the present study was garnered from the ICILS National Research Centre from Hong Kong and South Korea. The data was collected from the Field Trial which was in preparation for the main study.

A total of 300 students from 15 schools and 504 students from 26 schools of 8th graders from Hong Kong and South Korea, respectively, participated in the Field Trial.

RESULTS AND DISCUSSION

Test 1 (After School Exercise) consists of 23 items, Test 2 (Band Competition) 21 items, Test 3 (How people breathe) 17 items, and Test 4 (School trip) 16 items, respectively. The data for these four modules was subjected to Rasch analysis.

(a) Maps of Persons and Items

Person-Item map visually displays relationships between student abilities and test item difficulties in a linear scale in unit *logit*. Person estimates are distributed on the left side and item estimates on the right based on their ability and difficulty estimates, respectively. This can therefore determine if the item difficulties were appropriate for the targeted students (Bond & Fox, 2007; Oon & Subramaniam, 2011a; 2011b; Oon & Subramaniam, 2013).

In the map, "M" represents mean for person and item, "S" represents one sample standard deviation away from the mean, and "T" represents two sample standard deviations away from the mean.

Students located closer to the lower end performed less well compared against those located at the upper end. Items at the bottom are less challenging while items at the top are more challenging to be answered correctly by students.

An instrument that is well-targeted at the intended sample will show that the cluster of person is located opposite to the cluster of item (Bond & Fox, 2007; Oon & Subramaniam, 2011a; 2011b; Oon & Subramaniam, 2013).

For the data in the present study, the Person-Item map for each test module is presented separately, as follows.

Figure 1 displays the spread of items and students for Module 1 (After School Exercise).

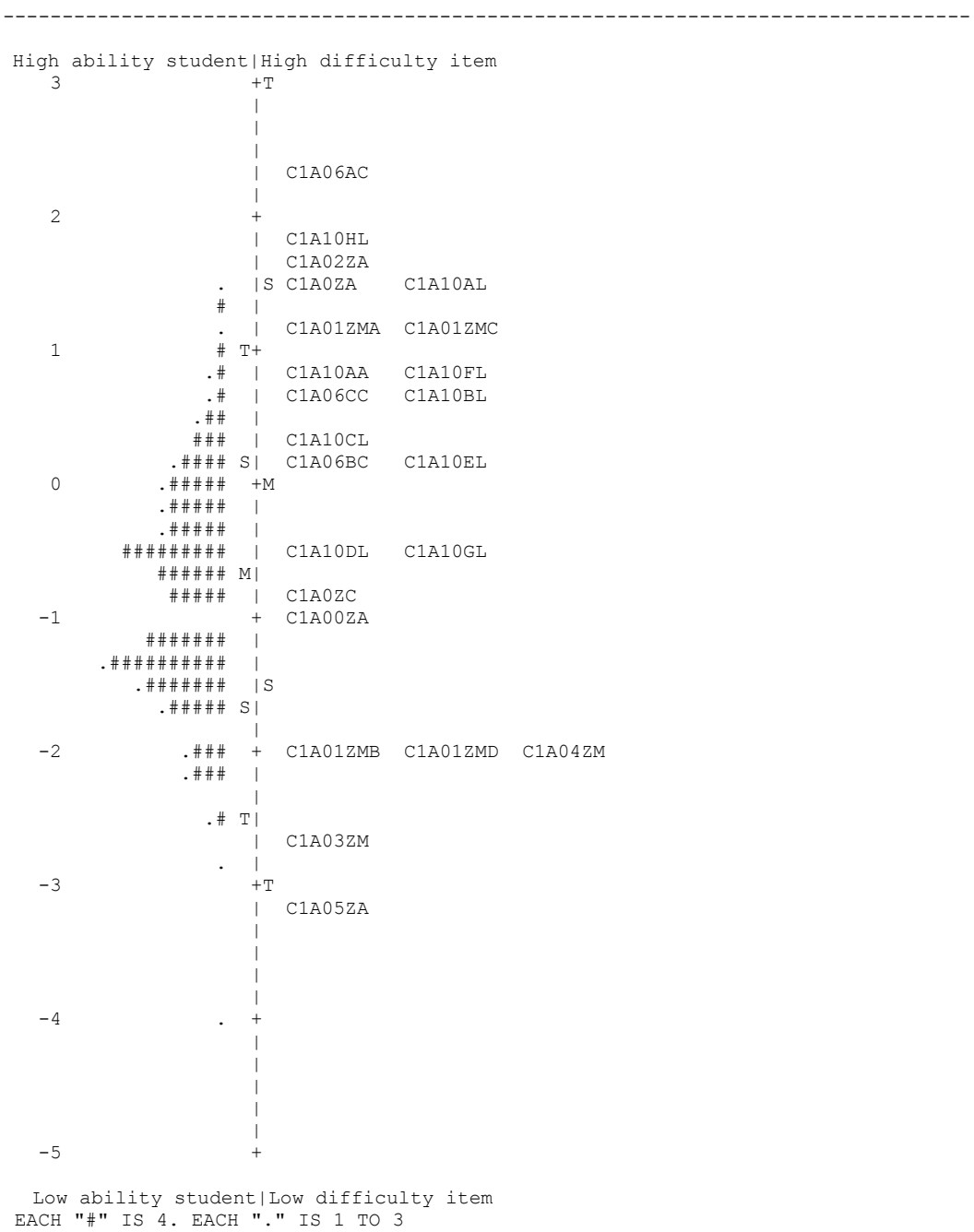


Figure 1: Person-Item map of Module 1(After School Exercise)

The student ability estimates ranged from -4.03 to 1.58 while item difficulty estimates ranged from -3.24 to 2.30. The person mean estimate is -.77 and item mean is .00, which has been set arbitrarily. The item difficulty and student ability estimates spread evenly along the standardized linear scale with good overlap between them. A minuscule of students with lower CIL level does not sufficiently examined by the items in this test.

Figure 2 displays the spread of items and students for Module 2 (Band Competition).

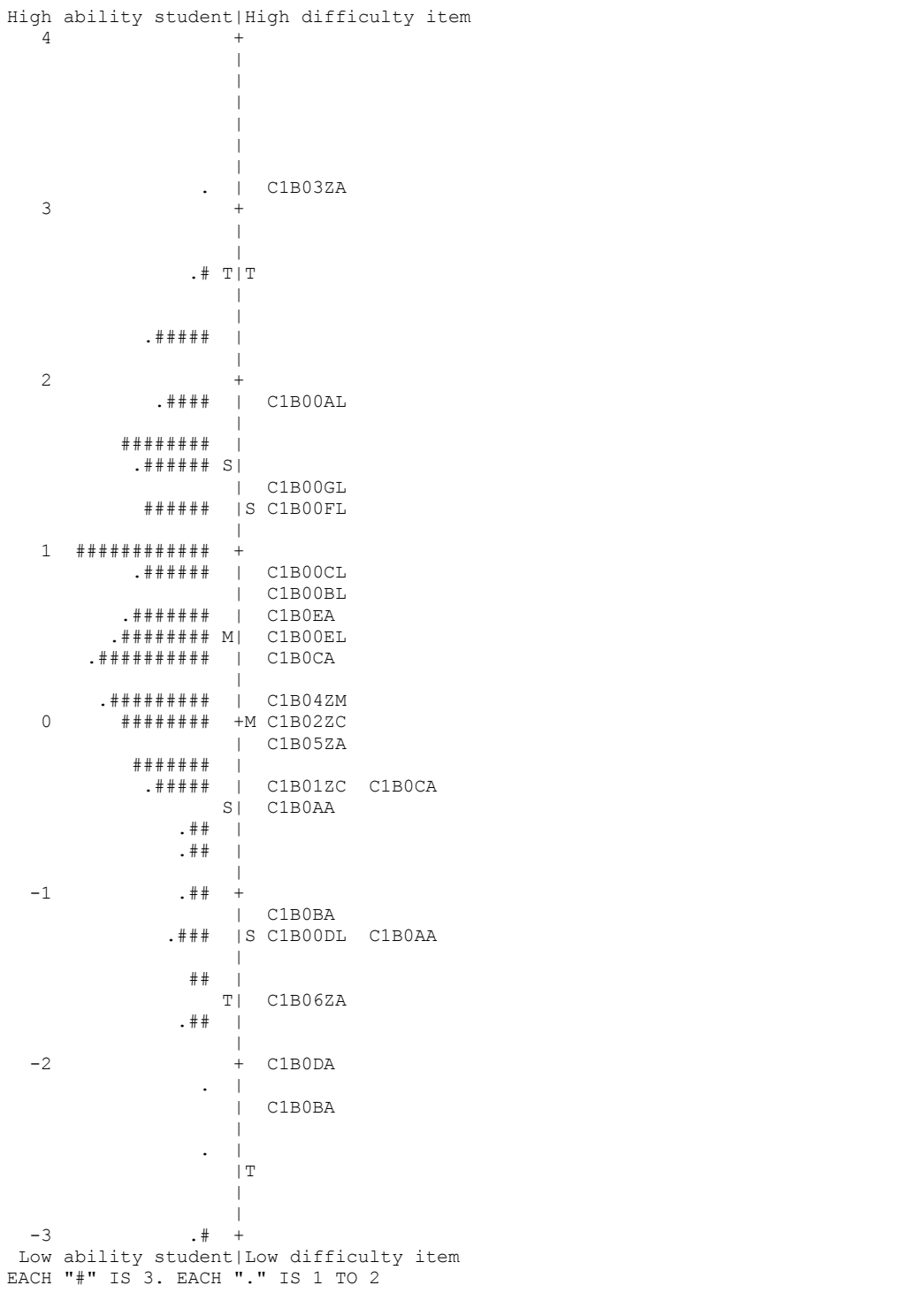


Figure 2: Item-Student map of Module 2(Band Competition)

The student ability estimates ranged from -2.98 to 3.14 while item difficulty estimates ranged from -2.26 to 3.10. The person mean is .48 while the item mean is .00 with no wide divergence. Good overlapping is evident in light to the item and person distributions. The items are well-targeted at the sampled students although some of the sampled students' ability estimates stayed below the estimates to the least difficult items.

Figure 3 displays the spread of items and students for Module 3 (How People Breathe).

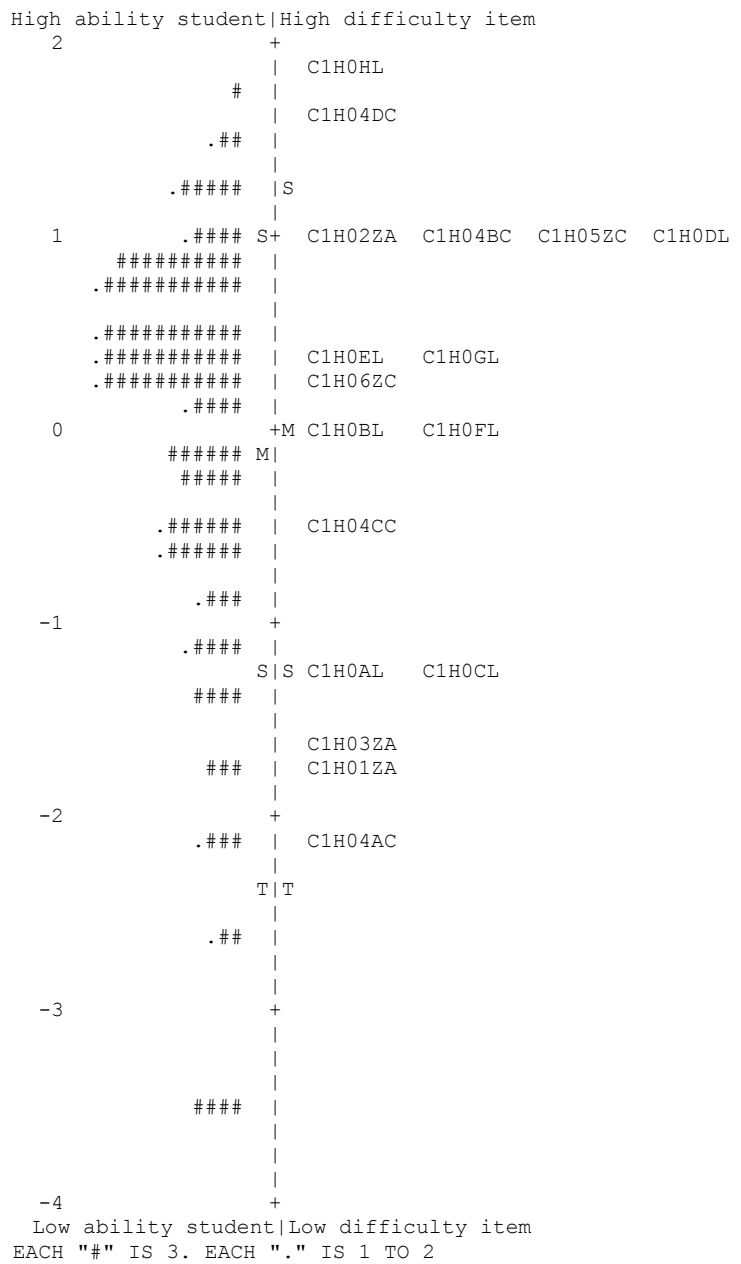
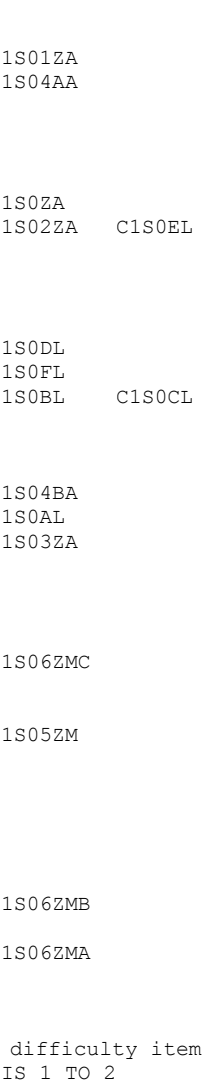


Figure 3: Item-Student map of Module 3 (How People Breathe)

The student ability estimates ranged from -3.47 to 1.72 while item difficulty estimates ranged from -2.12 to 1.89. The person mean estimate (-.14) is centered around item mean estimate (.00), with no wide divergence. The items and persons clustered opposite to each other and spread evenly along the standardized scale. These indicate good targeting of items to the sampled students. However, about 2.5% of the sampled students reported ability estimates lower than the least difficult item – this indicates that none of the items in this test module assess sufficiently this portion of students who demonstrated lower CIL level.

Figure 4 displays the spread of items and students for Module 4 (School Trip).



The student ability estimates ranged from -3.25 to 1.84 while item difficulty estimates ranged from -3.55 to 2.68. The person mean estimates (.09) and item mean estimate (.00) centered around each other with no divergence. All of the sampled students are located opposite to the cluster of items. The very good overlap of persons and items indicates that the all items in this test sufficiently measure students with different CIL levels.

All items in the 4 test modules are well-targeted at the Hong Kong and Korean students. Items from Test 4 (School Trip) showed the best targeting compared to the other 3 test modules. Also, the modules included both difficult and easy items, as evident in the spread of items along the standardized linear scale.

(b) Differential Item Functioning (DIF)

If the invariance property holds, item estimates should remain stable across the relevant subsamples (Bond & Fox, 2007; Oon & Subramaniam, 2011a; 2011b; Oon & Subramaniam, 2013). To investigate whether the items use to measure CIL function in

the same way for the subsamples, Rasch measures for Hong Kong and Korean students were calculated.

The DIF contrast is the difference in measure for an item between Hong Kong and Korean students. It should be at least .50 *logits* for DIF to be noticeable (Linacre, 2009). If DIF is evident, the invariance property is concluded to be absent.

Eleven items from Test 1 (C1A01ZMC, C1A02ZA, C1A06AC, C1A06BC, C1A0ZC, C1A0ZA, C1A00ZA, C1A10AL, C1A10CL, C1A10FL, C1A10HL), 10 items from test 2 (C1B03ZA, C1B06ZA, C1B0AA, C1B0AA, C1B0BA, C1B0EA, C1B00BL, C1B00CL, C1B00DL, C1B00FL, C1B00GL), 11 items from Test 3 (C1H02ZA, C1H03ZA, C1H04AC, C1H04BC, C1H04CC, C1H04DC, C1H0DL, C1H0EL, C1H0FL, C1H0GL, C1H0HL) and 10 items from Test 4 (C1S01ZA, C1S03ZA, C1S04AA, C1S04BA, C1S05ZM, C1S06ZMA, C1S06ZMB, C1S06ZMC, C1S0ZA, C1S0EL) report significant DIF between Hong Kong and Korean students (Table 1).

Table 1 Differential Item Functioning for Hong Kong and Korean students

Item	**Hong Kong**	**Korea**	**DIF Contrast**
Test 1: After School Exercise			
C1A01ZMA	1.08	1.32	-0.24
C1A01ZMB	-2.00	-1.78	-0.22
C1A01ZMC	1.73	1.03	0.70
C1A01ZMD	-1.95	-1.70	-0.25
C1A02ZA	0.26	3.78	-3.52
C1A03ZM	-2.59	-2.55	-0.04
C1A04ZM	-1.38	-1.87	0.49
C1A05ZA	-2.87	-3.19	0.32
C1A06AC	3.63	2.05	1.58
C1A06BC	-0.02	0.48	-0.50
C1A06CC	1.20	0.75	0.45
C1A0ZC	0.32	-0.98	1.30
C1A0ZA	0.07	3.23	-3.16
C1A00ZA	-1.22	-0.55	-0.67
C1A10AA	0.80	1.17	-0.37
C1A10AL	1.91	2.50	-0.59
C1A10BL	1.40	1.69	-0.29
C1A10CL	1.03	1.54	-0.51
C1A10DL	0.62	0.86	-0.24
C1A10EL	0.55	0.20	0.35
C1A10FL	2.79	1.39	1.40
C1A10GL	1.14	0.69	0.45
C1A10HL	2.99	1.64	1.35
Test 2: Band Competition			
C1B01ZC	-1.36	-1.36	0

Item	Hong Kong	Korea	DIF Contrast
C1B02ZC	1.62	1.12	0.5
C1B03ZA	0.73	3.17	-2.44
C1B04ZM	-0.76	-0.89	0.13
C1B05ZA	-1.04	-1.18	0.14
C1B06ZA	-2.18	-2.96	0.78
C1B0AA	-1.89	-1.29	-0.6
C1B0BA	-3.11	-3.19	0.08
C1B0CA	-0.59	-0.66	0.07
C1B0AA	-1.63	-2.67	1.04
C1B0BA	-1.76	-2.33	0.57
C1B0CA	-1.31	-1.44	0.13
C1B0DA	-2.69	-3.01	0.32
C1B0EA	-0.78	-0.24	-0.54
C1B00AL	0.59	0.86	-0.27
C1B00BL	-0.92	-0.04	-0.88
C1B00CL	1.49	2.23	-0.74
C1B00DL	-1.85	-2.47	0.62
C1B00EL	1.42	1.86	-0.44
C1B00FL	1.93	2.5	-0.57
Test 3: How People Breathe			
C1B00GL	-0.63	0.79	-1.42
C1H01ZA	-1.96	-2.18	0.22
C1H02ZA	2.22	-0.17	2.39
C1H03ZA	-2.38	-1.71	-0.67
C1H04AC	-1.51	-3.15	1.64
C1H04BC	2.1	-0.08	2.18
C1H04CC	3.32	0.16	3.16
C1H04DC	2.63	3.33	-0.7
C1H05ZC	0.64	0.39	0.25
C1H06ZC	-0.28	-0.28	0
C1H0AL	-1.64	-1.64	0
C1H0BL	1.87	1.69	0.18
C1H0CL	0.73	0.92	-0.19
C1H0DL	1.8	3.12	-1.32
C1H0EL	1.64	2.24	-0.6
C1H0FL	-0.99	-0.26	-0.73
C1H0GL	-0.14	0.6	-0.74
C1H0HL	1.44	0.3	1.14
Test 4: School Trip			
C1S01ZA	0.8	3.26	-2.46
C1S02ZA	3.09	3.23	-0.14
C1S03ZA	-0.5	-1.87	1.37
C1S04AA	0.15	3.75	-3.6

Item	Hong Kong	Korea	DIF Contrast
C1S04BA	-1.49	-0.89	-0.6
C1S05ZM	-1.68	-3.31	1.63
C1S06ZMA	-3.81	-4.47	0.66
C1S06ZMB	-4.33	-3.54	-0.79
C1S06ZMC	-2.55	-1.94	-0.61
C1S0ZA	0.52	1.35	-0.83
C1S0AL	-1.18	-1.15	-0.03
C1S0BL	-0.21	-0.52	0.31
C1S0CL	1.93	1.93	0
C1S0DL	-0.06	0.08	-0.14
C1S0EL	3.93	2.84	1.09
C1S0FL	2.05	2.01	0.04

Note: A = After School Exercise; B = Band Competition; H = How People Breathe; S = School Trip

The magnitude of the DIF contrast is depicted in Figure 5 for each item. The maximum DIF contrast detected in the 4 test modules are 3.52, 2.44, 3.16, and 3.6 logits, respectively.

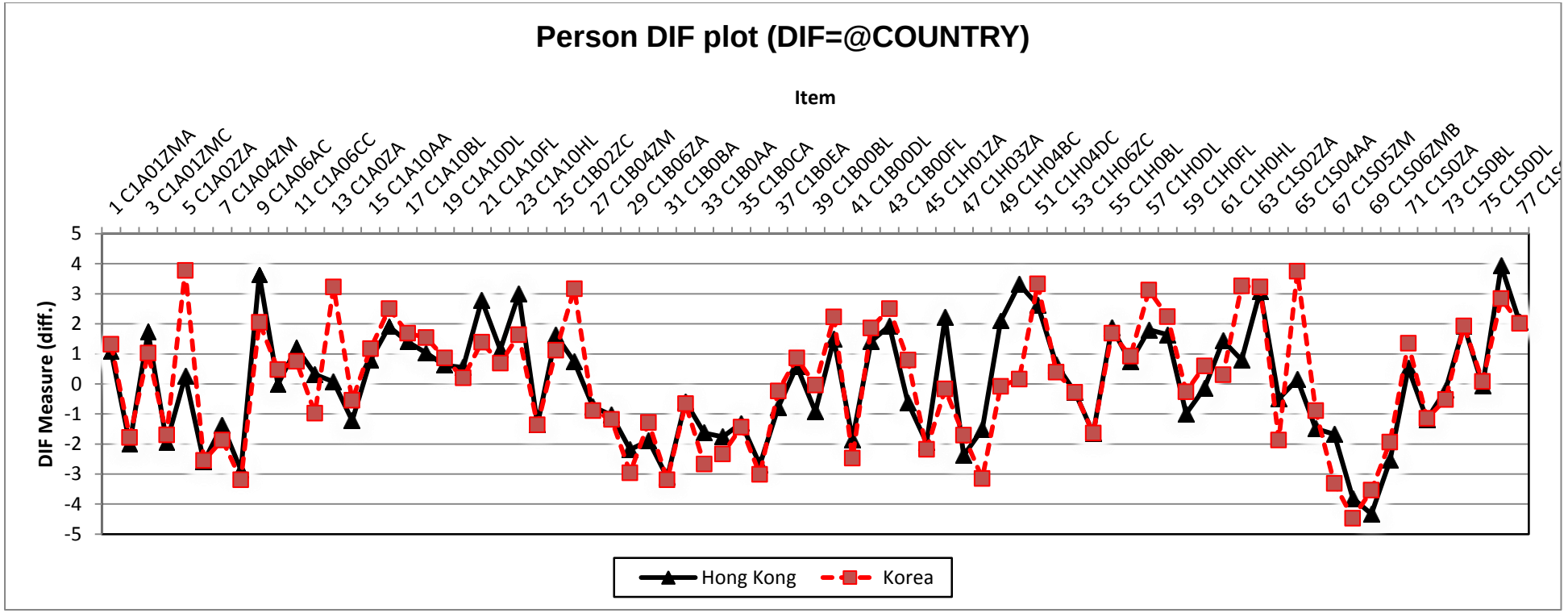


Figure 5: Plot of item estimates between Hong Kong and Korean students from Differential Item Functioning analysis

Note: A = After School Exercise; B = Band Competition; H = How People Breathe; S = School Trip

CONCLUSIONS

The findings from the present study show that the items from the 4 test modules target particularly well on Asian students with different CIL level, with Test 4 (School Trip emerged as the test showing best targeting on the sampled students. Also, results indicated that more easier items could be added in the 3 test modules (After School Exercise, Band Competition, and How People Breathe) in order to examine students with lower CIL level.

The results for the present study also reported that significant DIF plagued at least 40% of the items from each test module. The authors do not report the DIF analysis at item level in the present study as the main data collection is ongoing in many countries. Such analysis will inevitably disclose the individual item from which significant DIF is evident.

FUTURE WORK

Item analysis on items showing significant DIF will be scrutinized. Students' responses for the 4 test modules from a few countries from the west will be taken as reference in DIF analysis at item level.

REFERENCES

Bond, T.G., & Fox, C.M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glynn, S. M. (2012). International assessment: A Rasch model and teachers' evaluation of TIMSS science achievement items. *Journal of Research in Science Teaching, 49*(10), 1321 – 1344.

Linacre, J. M. (2009). WINSTEPS (Version 3.68.1) [Computer Software]. Chicago: Winsteps.com.

Rasch, G. (1960). *Probabilistic models for some intelligent and attainment tests.* Copenhagen: Dermarks Paedagogiske Institute.

Oon, P. T., & Subramaniam, R. (2011a). On the declining interest in physics among students – from the perspective of teachers. *International Journal of Science Education, 33*(5), 727 – 746.

Oon, P. T., & Subramaniam, R. (2011b). Rasch modeling of a scale that explores the take-up of Physics among school students from the perspective of teachers. In R. F. Cavanagh & R. F. Waugh (Eds.), *Applications of Rasch measurement in learning environments research.* Rotterdam: Sense Publishers (Netherlands), *pp.* 119 – 139.

Oon, P. T., & Subramaniam, R. (2013). Singapore school students' views about physics according to whether they intend to choose this subject as a tertiary field of study: A Rasch Analysis. *International Journal of Science Education*, *35*(1), 86 – 118.

Wendt, H., Bos, W., & Goy, M. (2011). On applications of Rasch models in international comparative large-scale assessments: A historical review. *Educational Research and Evaluation, 17*(6), 419 – 446.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: Mesa press.