Title	Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis
Author(s)	Shiau-Wei Chan, Chee-Kit Looi and Bambang Sumintono
Source	*Journal of Computers in Education,* (2020)
Published by	Springer

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2020). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. Journal of Computers in Education. Advance online publication. https://doi.org/10.1007/s40692-020-00177-2

Assessing Computational Thinking Abilities among Singapore Secondary Students: A Rasch Model Measurement Analysis

Shiau-Wei Chan, Chee-Kit Looi, Bambang Sumintono

ABSTRACT

In recent years, computational thinking (CT) skills have been globally recognized as a 21st-century skill that must be developed for future generations. However, the lack of validated CT assessments would be a major impediment in the efforts to incorporate CT into the school curriculum. This study is intended to validate the Computational Thinking Test (CTt) using the Rasch model by identifying whether the data fit the Rasch model measurement, determining the CT abilities among a small sample of Singapore secondary students through the test, and examining the presence of test items that functioned differently for gender and grade level of the students. In this study, 153 upper secondary school students from Grade 9 and Grade 10 were involved in a test that required them to do the CTt which comprises 28 test items. The performance of the students in CTt was utilized as quantitative data in this study and was analyzed using the Rasch model. The findings revealed that the data fit the Rasch model measurement. The majority of the male students and ninth-graders had a high level of CT abilities, while most of the female students and tenth-graders had a moderate level of CT abilities. Hence, the male students and ninth-graders performed better than the female students and tenth-graders. Four items functioned differently between male and female students where one gender had a better chance to get the correct answer in these four items compared to the other gender. Only one test item was functioning differently for Grade 9 and Grade 10. This means that the students of one grade-level were more likely to obtain the correct answer in this item than the students in the other grade-level. This study hopes to contribute to the literature in the area of CT assessments by providing a reference case for scholars and researchers in assessing CT abilities among the students.

Keywords: Computational thinking abilities; Computational Thinking Test; Secondary students; Rasch model measurement; Gender; Grade level

1. Introduction

In recent years, developing computational thinking (CT) in young generations has become a growing necessity in cultivating them with problem-solving and creativity skills which can be incorporated with digital technologies seamlessly (Kong, 2016). In this study, CT is regarded as a fundamental problem-solving cognitive procedure that enables the new practice of read-write. The person who is considered as code-literate is capable to write and read in the computer programming languages and other technologies, and to think computationally (Roman-Gonzalez, 2014). In this case, computer programming plays an important role as an enabler of CT (Lye and Koh, 2014), even though CT is not synonymous with computer programming (Wing, 2008). Considering such importance given to CT, many countries have widely introduced CT into their school curriculum by gradually infusing it into the Science, Technology, Engineering, and Mathematics (STEM) disciplines (Aydeniz, 2018). Arguments have been made for CT to be the integrated elements of the already existing syllabus, and not the extra part of syllabus content (Mueller, Beckett, Hennessey and Shodiev, 2017).

In scaling up CT in K-12 education, CT assessments play a vital role and thus ought to receive adequate attention in research. Various CT assessments have been developed to align with different definitions and theoretical frameworks of CT. They were utilized to assess the students from different grade levels (Adams, Cutumisu and Lu, 2019). The CT assessments were crucial to promote the understanding of the students on certain programming concepts and other CT skills such as coding tracing and debugging (Grover, 2017). Grover (2017) claimed that the CT assessments that are used to measure the learning of students should not just assess their grades, but emphasize the gaps in the students' understanding. The measures of formative and summative CT assessments that are utilized to assess the students' learning should be constructed, tested, and validated in different contexts with varied learners.

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2020). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. Journal of Computers in Education. Advance online publication. https://doi.org/10.1007/s40692-020-00177-2

Despite the increasing recognition of the significance of CT in K-12 education, researchers have yet to undertake a comprehensive psychometric validation process for the CT assessment instruments as validating the measurement instrument is valuable and essential. CT was still a poorly defined psychological construct from a psychometric viewpoint (Tang, Yin, Lin, Hadad and Zhai, 2020). Without valid and reliable assessments, it was hard for the instructors to employ these assessments to measure students' CT learning in the classroom with confidence (Tang, Yin, Lin, Hadad and Zhai, 2020) and to infuse CT into the educational system (Roman-Gonzalez, Moreno-Leon, and Robles (2019). Computational Thinking Test (CTt) was one of the CT assessments which required further validation (Cutumisu, Adams and Lu, 2019). Thus, this study sought to validate the CTt using the Rasch model by finding out whether the data fit the Rasch model measurement. This was crucial as validated assessments were pillars of effective learning as they evaluated the progress of the students in achieving the prescribed learning outcomes (Shute, Chen and Asbell-Clark, 2017).

After validating the CTt, the CT abilities among Singapore secondary students were measured using the Rasch model as well. The Rasch model analysis was employed in this study as it can deeply analyze the results as patterns among the scores of individual students, not merely as aggregated data. This study also determined the presence of test items that functioned differently for gender and grade level of the students. This was because earlier studies (e.g. Atmatzidou and Demetriadis, 2015) revealed that student's gender and age or grade level were the crucial factors in acquiring CT skills. Nevertheless, few studies are comparing the CT skills between students from different genders and grade levels. Hence, this study aimed to explore the gender and grade level differences in answering CTt. Our research questions of this study were as follows:

RQ1: To what extent, does the data collected from a sample of secondary students in Singapore fit the Rasch model?
RQ2: How do the CT abilities among secondary students vary amongst different gender and grade levels in Singapore?
RQ3: Is there any test item that functioned differently between male and female students?
RQ4: Is there any test item that functioned differently between Grade 9 and Grade 10 students?

2. Literature Review

2.1 Computational Thinking

In 1996, the term CT was utilized by Papert (1996) by focusing on how to use computation to construct new knowledge, as well as on how to use computers to promote thinking and change the way of acquiring knowledge (Tabesh, 2017). Nevertheless, it has been revitalized by Wing (2006) ten years later. Wing (2006) characterized CT as "involves solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science" (p.33). The core skill was thinking like a computer scientist when facing a problem. However, this definition has still not attained an agreement from educators. In this vein, Wing elucidated CT "is the thought processes involved in formulating problems and their solutions so that the solutions are represented in a form that can be effectively carried out by an information-processing agent" (Wing, 2011, p.1). In a simplified definition by Aho (2012), CT is considered as the thought processes involved in constructing problems so their solutions can be denoted as computational algorithms and steps. Roman-Gonzalez, Perez-Gonzalez and Jimenez-Fernandez (2017) defined CT as a fundamental problem-solving cognitive procedure that enables the new practice of read-write.

The International Society for Technology in Education and the Computer Science Teacher Association (ISTE and CSTA, 2011) created an operational definition of CT for K-12 education as a problem-solving procedure that comprises (but is not limited to) the following features: (a) creating problems that allow us to employ a computer and other tools to solve them; (b) logically organizing and analyzing data; (c) representing data via abstractions, for instance, simulations and models; (d) automating solutions via algorithmic thinking; (e) recognizing, analyzing, and applying viable solutions to accomplish the most effective and efficient combination of resources and steps; and (f) generalizing and transferring this problem-solving procedure to a myriad of problems. Although the boundaries between the formal definitions of CT were blurred, CT was usually defined as a problem-solving process that involves a set of cognitive and metacognitive activities using computational approaches, and creatively expressed as an algorithm (Cutumisu, Adams, & Lu, 2019).

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2020). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. Journal of Computers in Education. Advance online publication. https://doi.org/10.1007/s40692-020-00177-2

To evaluate CT skills among the students, a variety of CT assessments have been developed. According to Roman-Gonzalez, Moreno-Leon, and Robles (2019), there were seven types of the CT assessment tools, namely CT diagnostic tools, CT summative tools, CT formative-iterative tools, CT data-mining tools, CT skill transfer tools, CT perceptions-attitudes scales, and CT vocabulary assessment. The recent review conducted by Tang, Yin, Lin, Hadad and Zhai (2020) demonstrated that the majority of the studies assessed cognitive constructs including CT concepts and skills. Furthermore, most of the studies focused on fostering CT of the students from elementary and middle schools. The quality of the CT assessment was a crucial feature of selecting an assessment. Like any other assessment, CT assessments ought to fulfill particular criteria and psychometric standards, such as validity and reliability, to make sure that test-takers' scores are reported correctly (McMillan, Hellsten, and Klinger, 2011).

2.2 Rasch Model Measurement

Rasch model was a subset of a larger group of measurement models called item response theory (IRT) and had been used widely to analyze psychometric data in educational research (Khine, 2020). Rasch model analysis provided an extremely effective alternative to investigate the psychometric properties of measures and to address response bias (Bradley, Peabody, Akers and Knutson, 2015). The psychometric analysis approach provided by the Rasch model could be employed to develop the test items and served as a crucial tool in assessment for learning (Sumintono and Widhiarso, 2015). The Mok and Wright's five measurement principles for human science were addressed by the findings of Rasch model through logit ruler including (a) yield a linear measure; (b) overcome missing data; (c) provide a precision estimate; (d) discover outliers or misfits, and (e) replicable (Sumintono, 2018). Rasch model was chosen in this study as it can express a person's measures on the same scale regardless of which test or survey form the participants filled out (Khine, 2020). The estimates of latent traits were assessed according to the features of the person and item. By using the Rasch model, the success rate of the students in solving the test items could be examined based on the difficulty level of the items and the ability level of the students (Englehard, 2013).

2.3 Earlier Studies on Gender and Grade Level in Gaining CT skills

Several studies in the literature indicated the gender and grade level gap in gaining CT skills. For instance, Atmatzidou and Demetriadis (2015) performed a study with the 164 students of different gender and age groups (15 and 18 years old) to explore the development of students' CT skills in the educational robotics (ER) learning activity setting. During the study, the students involved in ER learning activities and different methods (oral and written) assessment tools had been utilized to assess their CT skills at different stages during the activity. The findings demonstrated that regardless of age and gender, students achieved the same level of CT skills development. However, compared with boys, girls required more training time in many cases to accomplish the same skill level. A study executed by Sullivan and Bers (2016) with the aim to analyze the results of the Ready for Robotics project to examine what gender stereotypes (if any) about technology and engineering young children that have started in kindergarten, and whether boys and girls were equally successful in mastering the robots and programming concepts, using a kit specially designed for toddlers. Although boys and girls had no significant differences in robotics and simple programming tasks, boys perform significantly better than girls on advanced programming tasks, for instance using repeated loops with sensor parameters.

Another study was implemented by Rijke, Bollen, Eysink and Tolboom (2018) to investigate the development of students' CT skills during their primary school years. The respondents were 200 primary school students from the ages of 6 and 12. Two CT skills were introduced to these students, i.e. decomposition and abstraction. It was found that older students performed better on abstraction tasks than students in the youngest age group. After the age of 9.5, female students started to perform better than their male peers on abstraction tasks. Furthermore, Angeli and Valanides (2020) conducted to discover the impact of learning using Bee-Bot on the CT of boys and girls in the setting of two scaffolding approaches. The results showed statistically significant learning outcomes between the initial assessment and final assessment of children's CT abilities. Even though both boys and girls benefited from the scaffolding approaches, a statistically significant interaction was found between gender and scaffolding strategies, indicating that

boys benefited from individualistic, kinesthetic, spatially-oriented, and manipulative-based activities with cards, while girls benefited more from collaborative writing activities.

3. Materials and Methods

3.1 Research Design

The quantitative approach was utilized in this study to answer the research questions, which was utilizing a cognitive test to measure the CT abilities of students. The data were produced in the form of a number in quantitative research whereas score was gained for the correct answer to the test items. Normally, this occurred by shifting deductively from abstract concepts to specific data collection approaches and obtained the exact numerical information generated from that approach. Numerical data signified unified, standardized, and compact technique to represent abstract concepts empirically (Neuman, 2014). In the present study, a diagnostic tool, CTt was used to collect data from Singapore secondary students. The data collected were then analyzed quantitatively using the Rasch model technique.

3.2 Respondents

In this study, the respondents who took part in the study were 153 upper secondary school students. They were chosen from four secondary schools in Singapore through a stratified random sampling technique. This technique was employed as it allowed the researchers to acquire a sample population that best represents the entire population being studied. The population of the students was divided into subgroups and a random sample was randomly selected from the subgroups (Neuman, 2014). The students who participated were recruited based on classroom sampling. The number of chosen respondents came from four schools: School A (19 Secondary Three), School B (39 Secondary Three and 37 Secondary Four), School C (23 Secondary Three), School D (35 Secondary Four). 81 of the respondents were students from Grade 9 and 72 of them were students from Grade 10. They consisted of 124 males and 29 females. Their age was in between 15 to 16 years old. All of the respondents were taking O-level Computing subjects. They were selected to identify their CT abilities. Regarding Institutional Review Board (IRB) regulations, the participation of all the respondents was voluntary and they agreed to be part of the study after explanations were provided to them such as that the identities of the students will be kept anonymous and confidential. The code was utilized to represent the students, for instance, M referred to male, F referred to female, X referred to Grade 9, and Y referred to Grade 10.

3.3 Instrumentation

For instrumentation, the CTt was used to identify the secondary students' ability of CT. CTt was a CT diagnostic tool that which developed by Roman Gonzalez (2015) intended to measure the ability to create and solve problems by drawing on the basic concepts of computing, as well as utilizing the logic-syntax of programming languages, such as loops, functions, basic sequences, variables, conditionals, and iterations. CTt was the most famous block-based assessment which did not link to a certain programming language or subject (Cutumisu, Adams and Lu, 2019). This test had 28 multiple-choice items with four options, i.e. A, B, C, and D. CTt covered seven computational concepts which arranged in increasing difficulty order, i.e. Basic directions and sequences (BD) (4 items); Loops-repeat times (4 items); Loops-repeat until (4 items); If-simple conditional (4 items); If/else-complex conditional (4 items); While conditional (4 items); and Simple functions (4 items). The computational concepts were associated with the CSTA Computer Science Standards for Grade 7 and 8 (CSTA, 2011), as well as some of the CT framework from Brennan and Resnick (2012). There were two types of the environment-interfaces for CTt, namely 'The Canvas' (5 items), and 'The Maze' (23 items).

Furthermore, CTt also comprised of two styles of response alternatives in every item, which were Visual blocks (20 items), and Visual arrows (8 items). Three cognitive tasks were embedded in CTt, i.e. Debugging (5 items): amending the wrong commands; completion (9 items): completing unfinished commands, and sequencing (14 items): arranging the commands in an orderly way (Roman-Gonzalez, Perez-Gonzalez and Jimenez-Fernandez, 2017). The sample of items as shown in Figure 1 and Figure 2. Item 11 was regarded as the maze, visual arrows, loops-repeat until, yes-nesting, and debugging. This item required the students to identify the step of instructions to take 'Pac-Man'

to the ghost by the path marked out that has a mistake. Meanwhile, item 26 was deemed as the canvas, visual blocks, simple functions, yes-nesting, and completion. The students were required to find out the missing step of instructions to make the artist draw the triangles in this item.

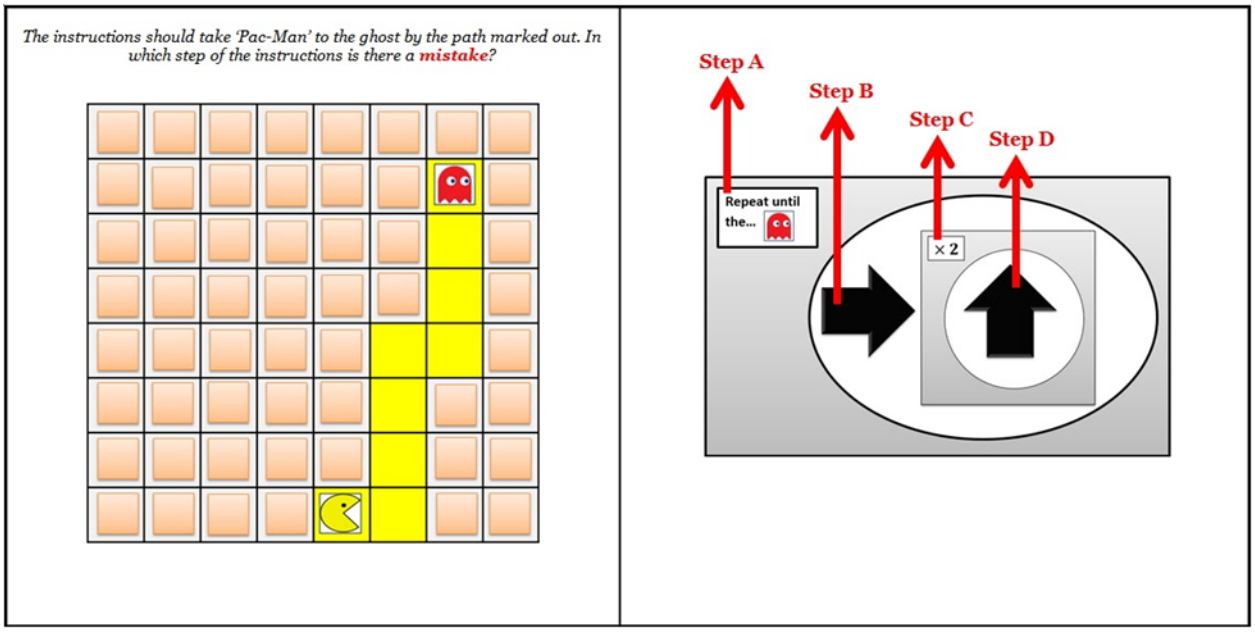

Figure 1. Item 11: Loops-repeat until; The Maze; visual arrows; yes-nesting; debugging

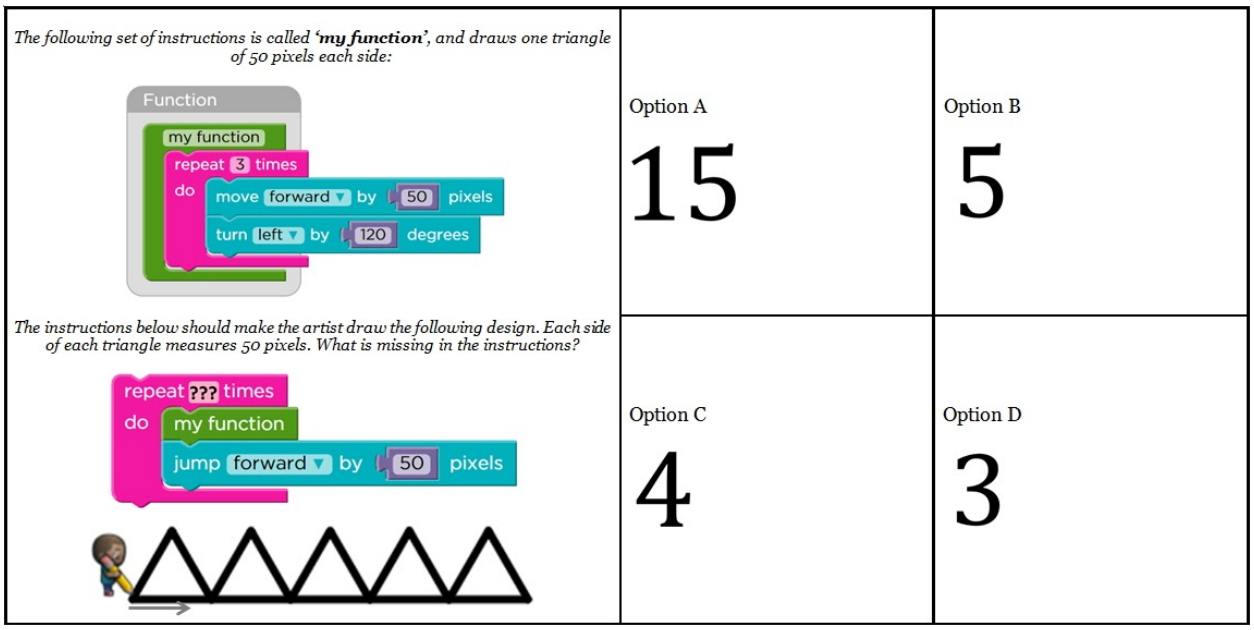

Figure 2. Item 26: Functions-simple functions; The Canvas; visual blocks; yes-nesting; completion

3.4 Procedure

This study took four days to complete the data collection at four secondary schools in Singapore. The CTt was distributed through Google forms to the students during the Computing subject periods in the school. The students

completed the test using a laptop in the computer laboratory. They were given 45 minutes to complete CTt. In the google forms, the students have to give their demographic information including gender, school name, and grade level. Before starting the CTt, the researcher gave some instructions to the students and provided three examples of items, so that the students could familiarize themselves with the kind of questions and the characters that will appear in the CTt. After completing all the questions, the students clicked on the 'Submit' button to save their answers. The answers of the students were analyzed as the quantitative data in this study.

3.5 Data Analysis

Rasch model analysis was employed in this study to analyze the data collected from the CTt. The data collected were entered into Microsoft Excel and then imported into the software of Winsteps version 3.73. To validate CTt, the content validity and internal consistency reliability of CTt were determined. The item reliability, item separation, person reliability, and person separation were examined as well. The appropriateness at the item quality was checked using several criteria: Outfit mean square (MNSQ), Outfit z-standardized (ZSTD), and Point-measure correlation (Pt-Measure Corr) for each item. The Wright map was displayed to demonstrate comprehensively about the item difficulty and abilities of students. Furthermore, logit value person (LVP) analysis was performed to identify the Singapore secondary students' CT abilities. The person-fit analysis was also examined using three criteria, i.e. Outfit mean square (MNSQ), Outfit z-standardized (ZSTD), and Point-measure correlation (Pt-Measure Corr). To explore the presence of test items that functioned differently for gender and grade level of the students, Differential Item Functioning (DIF) analysis was also executed using Winsteps.

4. Results and Discussion

In this study, the quantitative data were analyzed to validate the CTt using the Rasch model by finding out whether the data fit the Rasch model measurement, to examine the CT abilities among Singapore secondary students, as well as to determine the presence of test items that functioned differently for gender and grade level of the students. The results for these purposes were reported in the following sections.

RQ1: To what extent does the data collected from a sample of Singapore students fit the Rasch model?

The content validity of the CTt was identified by assessing fit validity (Baghaei, 2008). The assessment of Rasch item-fit statistics was utilized to determine the degree to which item fitted the model and therefore fit the concept of a single attribute (Boone, Staver and Yale, 2014). Two fit-indices were examined for CTt, namely Infit MNSQ (mean-square) and outfit MNSQ (mean-square). In Table 1, it was found that the Outfit MNSQ of CTt item mean was 1.01 which was very close to the ideal value of 1.0, within the acceptance ranges of 0.5 to 1.5 (Boone, Staver and Yale, 2014). Besides that, Cronbach Alpha (KR-20) person raw score test reliability was utilized to assess the internal consistency of the students' responses in the test (Sekaran, 2003). The value of Cronbach's Alpha for the CTt was 0.77 which was considered good (Isa and Naim, 2016). Furthermore, the raw variance explained by measures was 32.6% which was more than 20% as suggested by Sumintono and Widhiarso (2015). Hence, the data were fit to the model as the items were productive for measurement and had a reasonable prediction.

The item reliability index was the estimate of the replicability of item placement within a hierarchy of items along with the measured variable if these same items were to be given to another sample of people who had a similar ability (Bond and Fox, 2015). From Table 1, the CTt had a highly reliable value of item reliability, i.e. 0.95. An item separation index was an estimate of the separation or spread of items along with the measured variable (Bond and Fox, 2015). The CTt also had enough spread as it was greater than three, which was 4.36. As shown in Table 1, the mean measure (logit) of the item is 0.00 logit and the standard deviation is more than one logit (1.53), which suggests a very wide dispersion of measures across the logit scale in item difficulty level. This shows that the instrument can measure a larger spectrum of student ability in terms of CTt.

Table 1. Summary Statistics of Person and Items

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2020). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. Journal of Computers in Education. Advance online publication. https://doi.org/10.1007/s40692-020-00177-2

	Person	Item
N	148*	27*
Measures (logit)		
Mean	1.86	0.00
SD, standard deviation	1.13	1.53
SE, standard error	0.09	0.30
Outfit Mean Square		
Mean	1.01	1.01
SD	1.00	0.46
Separation	1.47	4.36
Reliability	0.68	0.95
Alpha cronbach	0.77	
Raw variance explained by measures	32.6%	

* item and person outlier drop from this table

On the other hand, the person reliability index was the estimate of the replicability of person placements that can be anticipated if a certain sample of people were to be given another set of appropriate items measuring the same construct (Bond and Fox, 2015). In Table 1, the person reliability for CTt was 0.68 which was regarded as fair reliable (Isa and Naim, 2016). A person separation index was an estimate of the separation or spread of persons on the measured variable (Bond and Fox, 2015). The value of person separation for CTt was more than one, i.e. 1.47 which indicated the samples were sufficient to separate the ability of a person (Gracia, 2005). Person logit mean was +1.86 logit showing all respondents were considered to have above-average ability (higher than item mean) for the CTt. Its standard deviation is 1.13 indicating a wide dispersion level of ability in the students.

The criteria used to check the appropriateness at the item level were including the values of Outfit MNSQ, Outfit ZSTD, and Pt-Measure Corr for each item. A range between 0.5 and 1.5 for Outfit MNSQ of the item and person proposed a suitable fit of the data to the model (Boone, Staver and Yale, 2014). The value of the Outfit ZSTD should be between -1.9 to 1.9 to determine that the items had reasonably predictability (Boone, Staver and Yale, 2014). The Pt-Measure Corr was employed to examine whether all the items were functioning in the intended direction. The positive value of Pt-Measure Corr was considered as acceptable, but the negative value indicated that the items were not functioning as compared with the others (Bond and Fox, 2015).

Table 2. Item-fit analysis

Item	Measure	Infit MNSQ	Outfit MNSQ	Outfit ZSTD	Pt-Measure Corr
Q1	***-4.94***	***1.00***	***1.00***	***0.00***	***0.00***
Q2	-3.73	1.00	0.34	-0.59	0.11
Q3	-1.40	1.01	0.83	-0.13	0.23
Q4	0.10	0.96	0.95	-0.14	0.38
Q5	-2.29	1.01	0.95	0.19	0.14
Q6	-2.05	1.12	3.01	2.12	-0.01
Q7	-1.68	1.01	0.72	-0.28	0.22
Q8	0.92	0.97	0.89	-0.68	0.46
Q9	-2.29	0.91	0.34	-0.91	0.25
Q10	0.20	1.05	1.00	0.10	0.34
Q11	-1.06	1.12	1.20	0.55	0.16
Q12	2.24	1.07	1.05	0.43	0.45
Q13	-0.49	1.04	1.09	0.36	0.28
Q14	0.20	0.88	0.68	-1.48	0.48

Q15	2.41	1.09	1.30	2.28	0.41
Q16	0.65	0.81	1.00	0.07	0.52
Q17	0.85	1.13	1.12	0.80	0.35
Q18	0.25	0.90	0.85	-0.64	0.45
Q19	1.26	0.98	0.92	-0.65	0.48
Q20	-0.17	0.88	0.62	-1.46	0.45
Q21	0.52	1.02	0.98	-0.05	0.40
Q22	1.44	0.93	0.87	-1.13	0.52
Q23	3.15	1.15	1.42	2.04	0.37
Q24	0.20	1.09	1.26	1.10	0.29
Q25	1.00	0.96	0.89	-0.77	0.47
Q26	0.43	0.99	1.22	1.10	0.38
Q27	-0.01	0.94	0.76	-0.90	0.42
Q28	-0.63	1.00	0.91	-0.13	0.30

Note: Q1: outlier item, too easy item

In Table 2, even though Q2 and Q9 were not in the range of Outfit MNSQ, but their Outfit ZSTD and Pt-Measure Corr were still within the acceptable range. The values of the Outfit ZSTD for items Q15 and Q23 were out of the range, but the values of Outfit MNSQ and Pt-measure Corr were in the range. Hence, all these items were retained and did not need to be removed. The item will only be considered as misfit when three criteria (Outfit MNSQ, Outfit ZSTD, and Pt-Measure Corr) are not met. But if only one or two criteria are not met, then the item still can be used for measurement purposes. It was noticed that item Q6 was regarded as "misfitting" as it did not meet the requirement for these three criteria with the value of Outfit MNSQ (3.01), Outfit ZSTD (2.12), and Pt-Measure Corr (-0.01). Thus, this item was omitted and excluded in the subsequent analysis. So, a total of 27 items were analyzed in the survey. In short, the overall data collected from the Singapore secondary students fit the Rasch model measurement.

The Wright map in Figure 3 presented the distribution of item difficulty and students' ability on the same logit scale (Bond and Fox, 2015). The item difficulty was placed on the left side of the Wright map, while the ability of the students was located on the right side of the Wright map. The higher logit implied the more difficult items and students with higher ability. Easier items and lower ability students were represented by the lower logit (Boone, Staver and Yale 2014). Figure 3 demonstrated that the most difficult item was item Q23 and the easiest item was item Q1. Meanwhile, there were five students with the highest ability 003MX, 001MX, 044Mx, 107MY, and 110MY. Student 021MX was the student with the lowest ability among 153 students. There were a few items free person which indicated that all the students were able to answer these items correctly, i.e. Q1, Q2, Q3, Q5, Q7, Q9, and Q11. There were some items with the same level of measurement, for instance, Q10, Q14, Q18, and Q24. Three items were difficult for the students, namely Q12, Q15, and Q23 as these item difficulties were above the person measure average. It means that the probabilities of being able to answer these items correctly were less than 50%.

The items of basic directions and sequences (BD), i.e. Q1, Q2, and Q3 were able to be solved by all the students. Eight of the 153 students (5%) could not answer Q4 correctly. Two of the items of loops-repeat times (LT) (Q5, and Q7) were answered by the students correctly. But 31 students out of 153 students (20%) were unable to solve item Q8. Such a situation perhaps due to the students did not understand the concept of loops-repeat times such as which command should be repeated and which command should not be repeated. Besides that, two items of loops-repeat until (LU) (Q9 and Q11) were answered correctly by all the students. However, another two items Q10 and Q12 were failed to be solved by 14 students (9%) and 90 students (59%). The students might not fully comprehend the concept of loops-repeat until, for instance, repeat how many times, and move forward to how many pixels. For items of If-simple conditions (SC), only three students (034FX, 101MY, and 021MX) who cannot solve the item Q13 accurately. 14 students (9%) failed to answer item Q14 precisely and 17 students (11%) solved item Q16 unsuccessfully. 97 of 153 students were unable to answer item Q15 in a correct way, i.e. 63% which was more than half of the students. This may be as a result of the students did not know how to apply the concept of If in the test. It

was surprising that the students were able to solve the items of If/else-complex conditional (CC) (Q17, Q18, Q19, and Q20) compared to If-simple conditional (SC). For item Q20, there were 3 students (2%) solved it incorrectly; for item Q18, there were 14 students (9%) who answered it wrongly. 31 students (20%) unable to solve item Q17 and 44 students (29%) answered item Q19 wrongly.

Regarding While conditional (WC), 14 students (9%) cannot solve item Q24 correctly, 17 students (11%) failed to answer item Q21 accurately, and 54 students (35%) solved item Q22 wrongly. Only 27 out of 153 students were able to solve the most difficult item (Q23), i.e. 18%. Even though the items in the CTt were ordered in increasing difficulty, but most of the students were capable to solve the items of simple functions (SF) with higher difficulty such as Q25 (80%), Q26 (91%), Q27 (97%), and Q28 (99%) compared to other items of while conditional such as Q22 and Q23. The possible reason for this phenomenon was the students were possibly tricked by the items of Q22 and Q23 by relating the five spaces that PacMan would need to move with the need to repeat the "move forward" command five times. In Figure 3, most of the students were located above the item logit average. This indicated a high probability of the students gained a high score for the CTt. Additionally, the person measure average was placed at logit +1.90 which meant that the average achievement of the students was above the average level of item difficulty. In other words, the CT abilities of the secondary students in Singapore was good as they were over the expected performance.

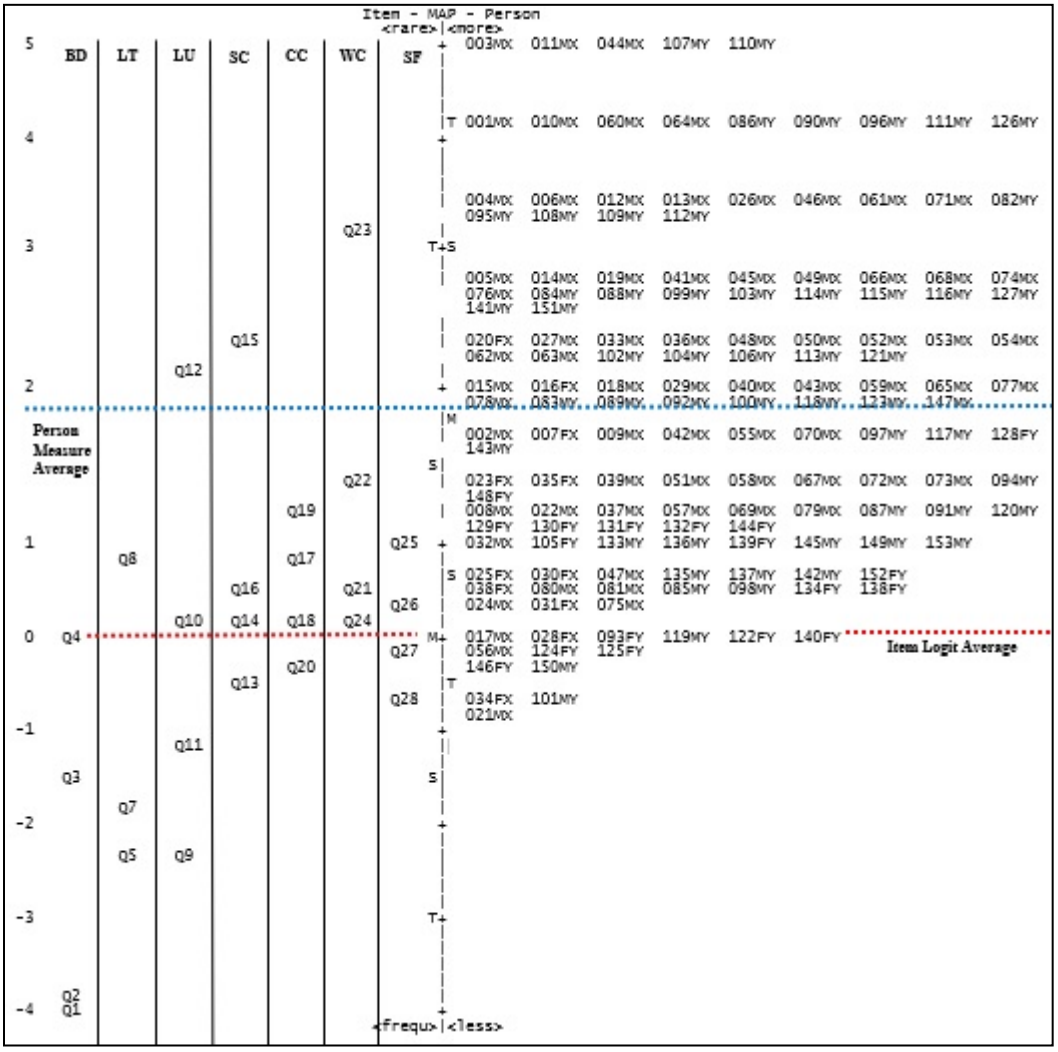

Figure 3. Wright map

RQ2: How are the CT abilities among secondary students vary amongst different gender and grade levels in Singapore?

Table 3 presents an analysis of the logit value person (LVP). Grouping of students' abilities was based on the value of the mean and standard deviation of all LVP. 50 out of 124 of the male students (40%) were in the group of the high ability of CT, compared to only two out of 29 female students (7%). There were 27 male students (22%) at the very high ability of CT, while there were no female students in this category. In the moderate CT ability group in this study, it consists of 34 (27%) male and 16 (55%) female. There were 13 males (11%) and 11 females (38%) in the group of low CT ability. It can be said that the CT abilities of most of the male students were at a high level and the CT abilities of most of the female students were at a moderate level. This demonstrated that the CT abilities of male students were superior to the CT abilities of female students in this study. This finding was not consistent with the outcome from the study of Atmatzidou and Demetriadis (2015) where the students reached the same level of CT skills regardless of gender and age. The performance of male students was better than the female students in this study most probably because the female students were less interested in the programming, and they had more anxiety and lack of confidence in using programming (Stoilescu, & Egodawatte, 2010).

For Grade 9 students, 30 of them (37%) obtained high person measures, followed by a group of very high ability (n = 15, 19%), moderate ability (n = 24, 30%) and low ability (n =12, 15%). High person measures achieved by 22 Grade 10 students (31%), very high person measures accomplished by 12 Grade 10 students (17%), and moderate person measures attained by 26 Grade 10 students (36%). There are 12 Grade 10 students who obtained low person measures, i.e. 17%. The CT abilities of the majority of the Grade 9 students were at a high level and the CT abilities of the majority of the Grade 10 students were at a moderate level. Such a situation revealed that Grade 9 students had higher CT abilities than Grade 10 students. It was surprising that the younger students performed better than the older students, which was contrary to the results from the study of Rijke, Bollen, Eysink and Tolboom (2018). This situation might be due to the younger students were less influenced by their existing knowledge, and their minds and brains were intrinsically more flexible and more open to new ideas of computing (Gopnik, Griffiths, & Lucas, 2015). Additional work such as interviews is needed to provide a warrant for the claim and rule out rival explanations.

Table 3. Logit value person (LVP) analysis (N =153)

Demographic	**Very High** $\boldsymbol{LVP > +2.94}$	**High** $\boldsymbol{+2.94 \geq LVP \geq +1.79}$	**Moderate** $\boldsymbol{+1.79 \geq LVP \geq +0.64}$	**Low** $\boldsymbol{LVP \leq +0.64}$
Gender				
Male	27	50	34	13
Female	0	2	16	11
Grade level				
Grade 9	15	30	24	12
Grade 10	12	22	26	12

There was one “misfitting” student as his Outfit MNSQ, Outfit ZSTD, and Pt-Measure Corr were not in the range, i.e. student 100MY (4.63, 2.41, -0.03). This was because he had response patterns in the CTt which were out of ordinary as shown in the person diagnostic PKMAPs in Figure 4. The diagram of person diagnostic PKMAPs indicated the easiest item was at the bottom and the most difficult item was at the top of the diagram. The items that were answered correctly were located on the left side of the diagram, while the items that were answered incorrectly were located on the right side of the diagram. In Figure 4, two easy items, i.e. Q5 and Q10 were answered wrongly by the student 100MY. This might be due to the carelessness made during the test. However, student 100MY can answer the items of 22 which had a higher difficulty level than his logit of ability, indicating a lucky guess made by the student.

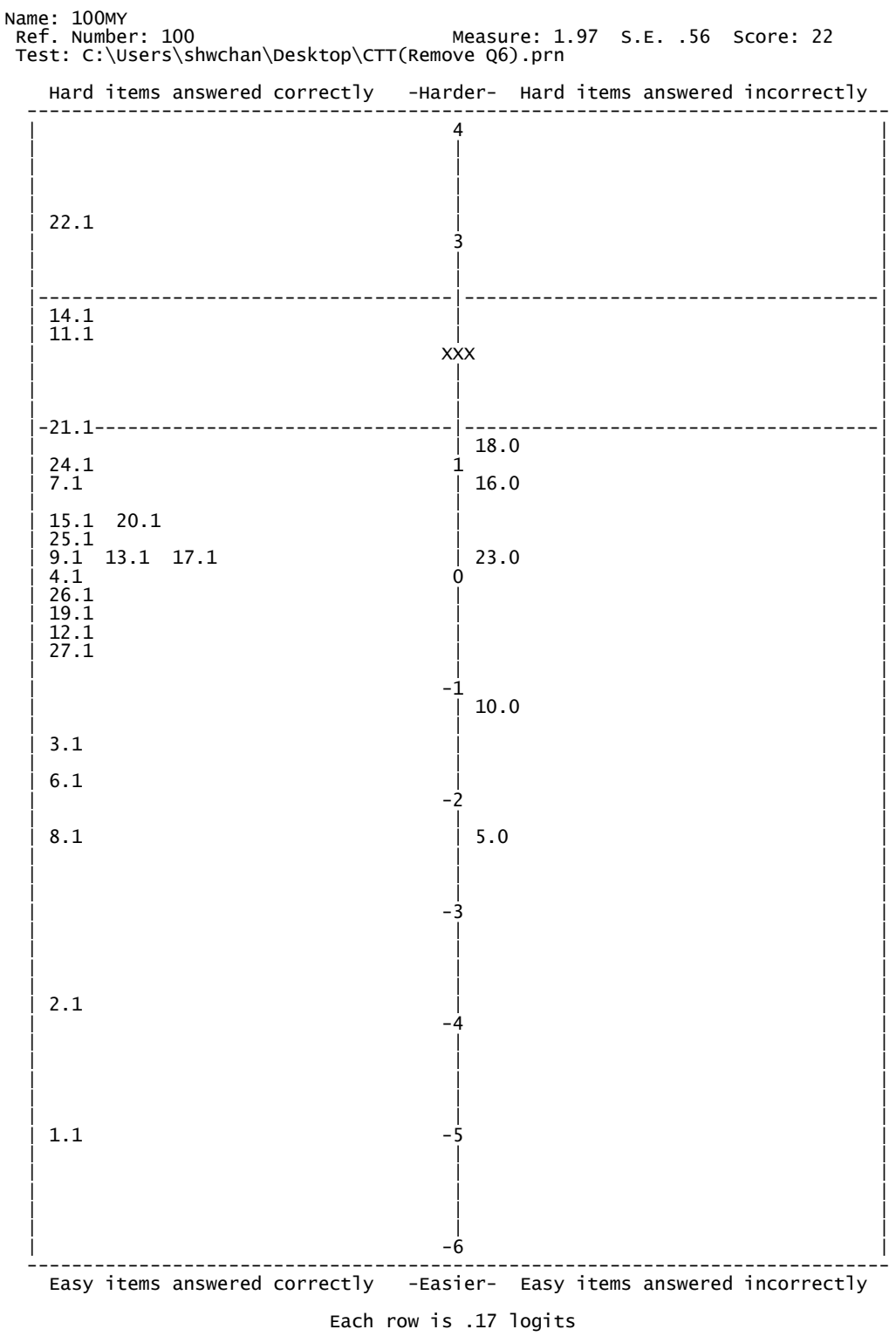

Figure 4. Response from student 100MY

RQ3: Is there any test item that functioned differently between male and female students?

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2020). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. Journal of Computers in Education. Advance online publication. https://doi.org/10.1007/s40692-020-00177-2

Differential Item Functioning (DIF) analysis was conducted to examine the differences in test item responses owing to gender. An item was considered as having DIF if it had the t value of less than −2.0 or more than 2.0, DIF contrast value of less than −0.5 or more than 0.5, and the p (Probability) value of less than 0.05 or greater than −0.05 (Boone, Staver and Yale, 2014; Bond and Fox, 2015). Therefore, four items were detected to have DIF as shown in Table 4, which were item Q15 (Figure 5), Q16 (Figure 6), Q17 (Figure 7), and Q23 (Figure 8).

Table 4. Differential item functioning (DIF) analysis by gender in CTt (N = 153).

Item	DIF		DIF Contrast	t	Prob.
	Male (M)	Female (F)			
Q15	2.53	1.48	-1.06	-2.32	0.0237
Q16	0.22	1.48	1.26	2.60	0.0114
Q17	1.01	0.05	-0.96	-2.06	0.0437
Q23	3.28	2.03	-1.25	-2.44	0.0179

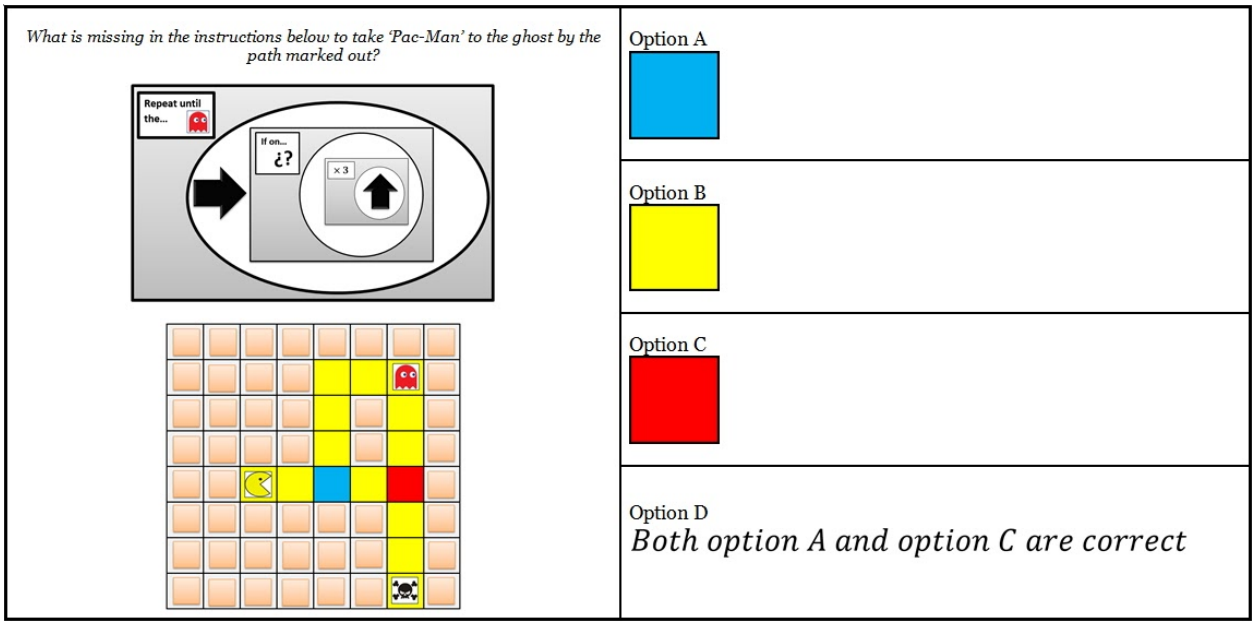

Figure 5. Item 15: If-simple conditional; The Maze; visual arrows; yes-nesting; completion

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2020). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. Journal of Computers in Education. Advance online publication. https://doi.org/10.1007/s40692-020-00177-2

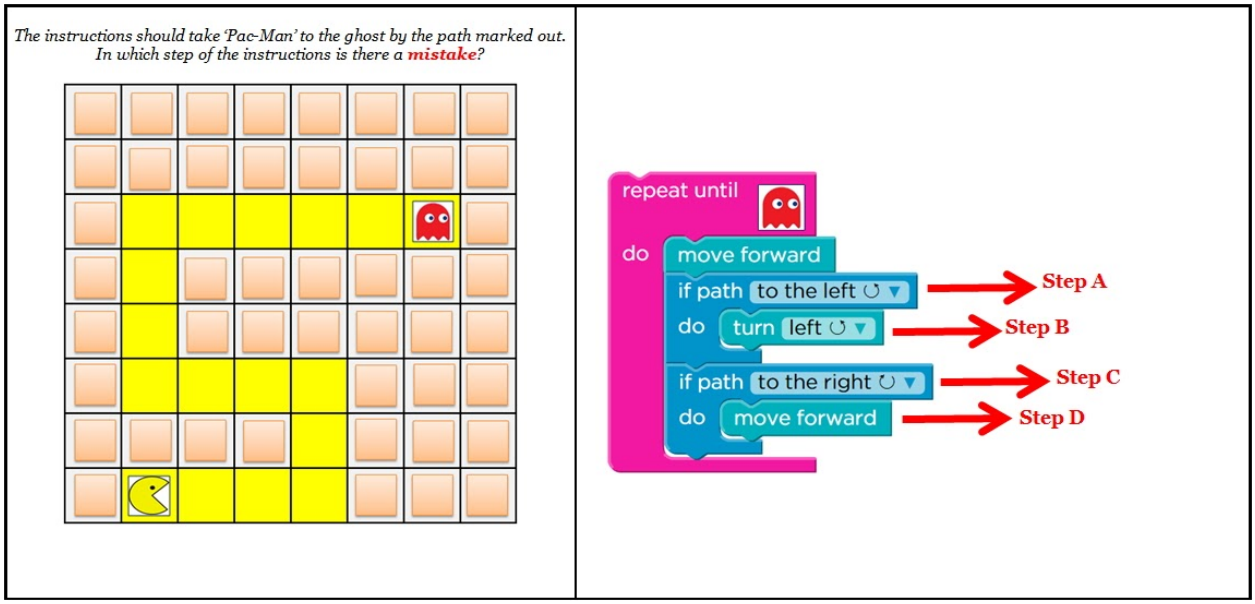

Figure 6. Item 16: If-simple conditional; The Maze; visual blocks; yes-nesting; debugging

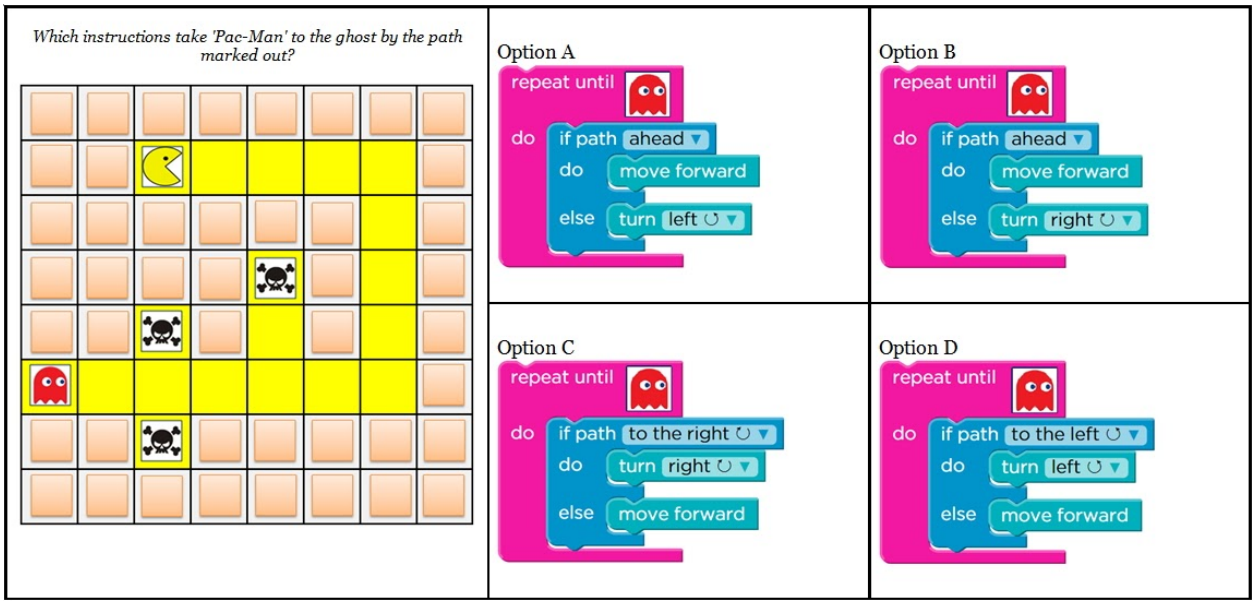

Figure 7. Item 17: If/else-complex conditional; The Maze; visual blocks; yes-nesting; sequencing

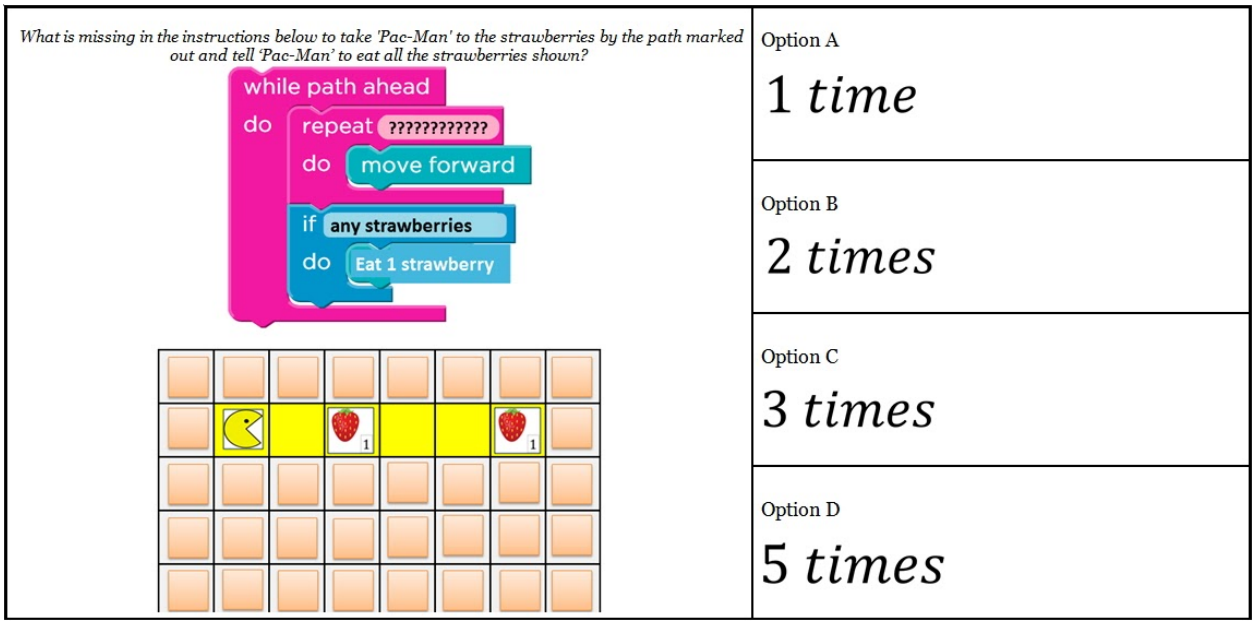

Figure 8. Item 23: while-conditional; The Maze; visual blocks; yes-nesting; completion

Further, Figure 9 shows the items Q15, Q17, and Q23 are more challenging for male students to get the correct answer compared to female students. These three items were classified as If-simple conditional, If/else-complex conditional, and while-conditional respectively. For Q15, the students ought to complete the missing instruction to take 'Pac-Man" to the ghost by the path marked out. The students had to determine the sequence of the given set of commands with the item Q17 of 'Which instructions take 'Pac-Man' to the ghost by the path marked out?'. In Q23, the students were required to find out the missing instructions to take 'Pac-Man' to the strawberries by the path marked out and inform the 'Pac-Man' to eat all the strawberries shown. But for item Q16, male students found it is easier compared to female students. Q16 was under the category of If-simple conditional where the students had to do the debugging to identify which step of the instructions has a mistake. It can be asserted that the achievements of male students and female students for items of Q15 and Q16 were different. For item Q15, the majority of the male students inclined to select answer C, and this demonstrated that they didn't fully understand that the external loop will keep moving the Pac-Man to the right until it reaches the ghost. For item Q16, the female students tended to choose answer C which indicated that they did not know how to take the sprite's perspective to identify which was ahead, left, and right in conjunction with a select statement to decide whether to turn left or right.

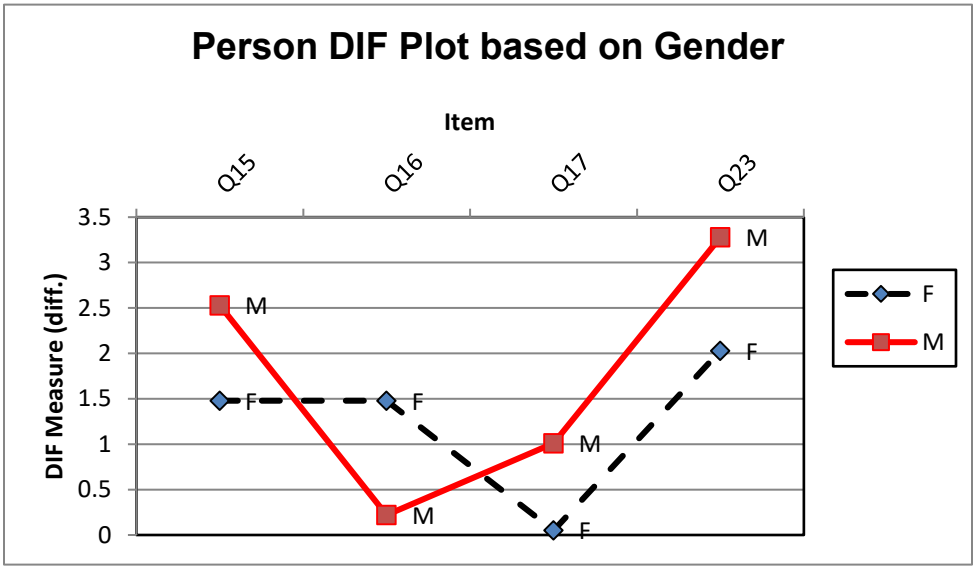

Figure 9. Person four item DIF Plot based on gender

RQ4: Is there any test item that functioned differently between Grade 9 and Grade 10 students?

According to Boone, Staver and Yale (2014), an item was considered as exhibiting DIF if the Mantel-Haenszel probability was less than 5% or 0.05. Then, the particular item would be classified as negligible, moderate or large DIF based on the values of Mantel-Haenszel chi-square. If the value was smaller than 0.43, the item was viewed as revealing negligible DIF. If the value was greater than 0.64, the item was regarded as indicating a large DIF. The item that had value in between 0.43 to 0.64 was deemed as showing moderate DIF (Zwick, Thayer and Lewis, 1999). As displayed in Table 5, item Q24 was having Mantel-Haenszel probability that less than 0.05 for both Grade 9 (X) and Grade 10 (Y). This indicated that item Q24 was exhibiting DIF and it was large DIF as the value of Mantel-Haenszel chi-square was larger than 0.64. Item 24 was categorized as a while-conditionals. The students were required to finish an incomplete given set of commands with the item of 'Which steps are missing in the instructions below to take 'Pac-Man' to the strawberries by the path marked out and tell 'Pac-Man' to eat all the strawberries?" as shown in Figure 10.

Table 5. DIF items for grade level

Person Class	Item	Mantel-Haenszel Probability	Mantel-Haenszel Chi-Square	DIF Type
X	Q24	0.0004	12.3207	large
Y	Q24	0.0004	12.3207	large

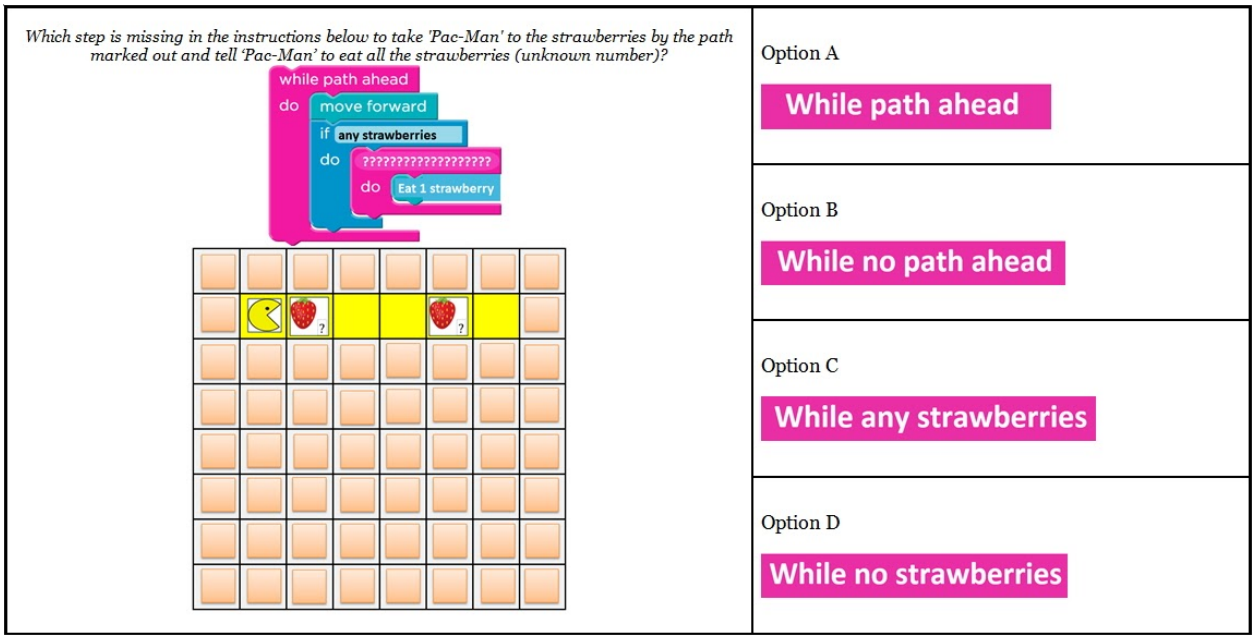

Figure 10. Item 24: while-conditional; The Maze; visual blocks; yes-nesting; completion

Item Q24 was further explored in the graph representation of person DIF in Figure 11. The curve of an item that approaching the upper limit such as item 23 presented a high level of difficulty. Meanwhile, the curve which approaching lower limit such as item Q1 demonstrated the low level of difficulty or easy item. The DIF measure of Grade 9 students for item Q24 was +0.93, while the DIF measure of Grade 10 students for item Q24 was -0.80. This implied that item Q24 was easier for students from Grade 10 to answer compared to Grade 9. Such a situation may occur because students in Grade 10 have learned while-conditional concepts in the previous year and have some understanding of it compared with students in Grade 9. For other items, the difference in the ability to do the items correctly does not differ much in terms of grade level.

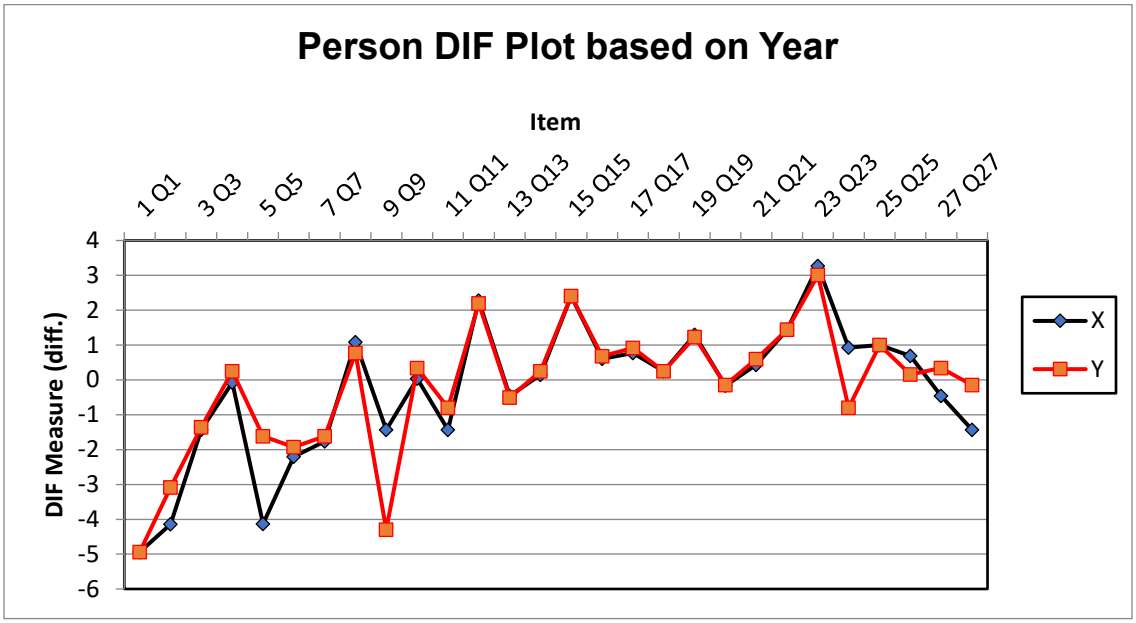

Figure 11. Person DIF Plot based on year

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2020). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. Journal of Computers in Education. Advance online publication. https://doi.org/10.1007/s40692-020-00177-2

5. Conclusion

This study has been conducted on a sample of 153 students of different gender and grade levels from four secondary schools in Singapore. For the research question one, the findings indicated that the data of this study fit the Rasch model measurement as the items of CTt were acceptable for good measurement and productive for measurement. They were also reasonably predictable and did not show under predictability or over predictability. With regard to research question two, the CT abilities for most of the male students and Grade 9 students were considered at a high level. Meanwhile, the CT abilities for most of the female students and Grade 10 students were deemed at a moderate level. Therefore, it can be concluded that the male students and Grade 9 performed better than the female students and Grade 10 in this study.

Concerning research question three, the DIF analysis for gender demonstrated that four items of CTt functioned differently between male and female students, namely item 15, item 16, item 17, and item 23 as they were not within the range of t value, DIF contrast value, and p-value. This indicated that one of the genders tended to obtain the correct solution in these four items than another gender. The person DIF plot displayed that the female students found item 15, item 17, and item 23 easier to be answered compared to the male students. However, the male found item 16 easier to be solved than female students. Three criteria (t value, DIF contrast value, and p-value) had been used to identify the DIF items. The more criteria we use, the better the measurement model.

Regarding research question four, an item, Q24, was exhibiting DIF as the Mantel-Haenszel probability was less than 0.05 and it was a large DIF due to the Mantel-Haenszel chi-square was greater than 0.64. The person DIF plot revealed that ten-graders can answer the item Q24 easily compared to nine-graders. This meant that the item Q24 functioned differently between Grade 9 and Grade 10 students. There was only one item out of 27 items that had DIF, this means that other items gave similar a chance for the students from two grade levels to get the correct solution. In sum, the DIF analyses for gender and grade level indicated that the quality of the instrument was fairly good in the sense that only a small number of items had DIF. Rasch measurement model was an appropriate way to analyze the data as it was sensitive enough to detect differences for gender and grade level.

This study contributes to the field of CT assessments by validating the CTt through psychometric analysis techniques using Rasch model measurement. The additional validation of CTt in this study fills the gap or missing literature of CTt. This will further address the problem of integrating CT into the school curriculum. It also allows the researchers and instructors to use CTt in the classroom confidently. This study serves as a guideline for the researchers and scholars in assessing the abilities of CT among the students as well. It provides valuable information about the differences in gender and grade level in acquiring CT skills. These findings will help the researchers and instructors in modifying CTt and developing new assessments. The results of this study have implications for the empirical evidence about the application of CT assessments in secondary schools towards facilitating the incorporation of CT in the school curriculum.

6. Limitations of Study and Future Direction

Nevertheless, there are some limitations to this study. One of them was the small sample size of students who participated in this study which is unable to represent the whole population. Hence the suggestion is to involve more students in future studies. Future research studies also ought to be conducted to further validate CTt for a wide range of students with different backgrounds and demographics such as the non-computing background. Furthermore, the results indicated that some items functioned differently for the students. Such a situation needs to be inspected through a further interview with the students in future studies to find out the reason which caused such situations to happen. The limitation of CTt is that it is a static and decontextualized assessment which only focused on computational concepts, and not computational practices and computational perspectives from Brennan and Resnick's (2012) framework (Roman-Gonzalez, Moreno-Leon, and Robles, 2019). Consequently, it is recommended that the researchers and instructors can further revise CTt to address this limitation in future investigation.

Another limitation was that there was only one CT assessment being utilized in this study. Roman-Gonzalez, Moreno-Leon and Robles (2019) argued using only one CT assessment is not sufficient to acquire the complete view

of CT performance amongst students. This might result in a misunderstanding of the CT development of the students which may result in making incorrect educational decisions. Grover (2015) asserted the use of multiple complementary assessments or 'systems of assessments' which provide convergent measures would lead to a more comprehensive understanding of the students. Thus, it is recommended to combine several CT assessments in future studies to implement a more comprehensive evaluation.

Acknowledgements We like to thank Marcos Roman-Gonzalez for providing the CTt test to us and answering questions about their use.

Funding Support for this paper was provided by the project grant for: Researching and developing pedagogies using unplugged and computational thinking approaches for teaching computing in the schools (Project Number: OER 04/16 LCK).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

Ethical Approval All procedures performed in studies involving human participants were approved by the Institutional Review Board (IRB) of the Nanyang Technological University (Ethics approval number: IRB-2016-10-023).

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

Adams, C., Cutumisu, M. & Lu, C. (2019). Measuring K-12 computational thinking concepts, practices and perspectives: An examination of current CT assessments. In K. Graziano (Ed.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 275-285). Las Vegas, NV, United States: Association for the Advancement of Computing in Education (AACE).

Aho, A. V. (2012). Computation and computational thinking. *The Computer Journal*, 55(7), 832-835.

Angeli, C., & Valanides, N. (2020). Developing young children's computational thinking with educational robotics: An interaction effect between gender and scaffolding strategy. *Computers in Human Behavior*, 105.

Atmatzidou, S., & Demetriadis, S. (2015). Advancing students' computational thinking skills through educational robotics. A study on age and gender relevant differences. *Robotics and Autonomous Differences*, 75, 661–670.

Aydeniz M. (2018). Integrating computational thinking in school curriculum. In M. Khine (Eds) *Computational Thinking in the STEM Disciplines*. Springer, Cham.

Baghaei, P. (2008). The Rasch Model as a construct validation tool. *Rasch Measurement Transactions*, 22, 1145–1146.

Bond T. G., & Fox C. M. (2015) *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.

Boone, W.J., Staver, J.R., & Yale, M.S. (2014). *Rasch Analysis in the human sciences*. Dordrecht, The Netherlands: Springer.

Bradley, K., Peabody, M., Akers, K., & Knutson, N. (2015). Rating scales in survey research: Using the Rasch model to illustrate the middle category measurement flaw. *Survey Practice*, 8(2), 1-14.

Brennan, K., & Resnick, M. (2012). *New frameworks for studying and assessing the development of computational thinking.* In Annual American Educational Research Association meeting, Vancouver, BC, Canada.

CSTA. (2011). *K-12 computer science standards.* Retrieved from http://csta.acm.org/Curriculum/sub/CurrFiles/CSTA_K-12_CSS.pdf

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2020). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. Journal of Computers in Education. Advance online publication. https://doi.org/10.1007/s40692-020-00177-2

Cutumisu, M., Adams, C., & Lu, C. (2019). A scoping review of empirical research on recent computational thinking assessments. *Journal of Science Education and Technology*, 28(6), 651-676.

Englehard, G. (2013). *Invariant measurement, using Rasch models in the social, behavioral and health sciences*. New York: Routledge.

Gopnik, A., Griffiths, T., & Lucas, C. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, 24(2), 87–92.

Gracia, S. (2005). *Analyzing CSR implementation with the Rasch model.* Rhode Island College.

Grover, S. (2015). "Systems of assessments" for deeper learning of computational thinking in K-12. In *Proceedings of the 2015 Annual Meeting of the American Educational Research Association* (pp. 15–20). Retrieved from https://www.sri.com/sites/default/files/publications/aera2015_systems_of_assessments_for_deeper_learning_of_computational_thinking_in_k-12.pdf

Grover, S. (2017). Assessing algorithmic and computational thinking in K-12: Lessons from a middle school curriculum. In P. Rich & C. Hodges (Eds). *Emerging research, practice, and policy on computational Thinking. Educational Communications and Technology: Issues and Innovations* (pp. 269-288). Springer, Cham.

Isa N.M., & Naim H.A. (2016). Science process skill assessment: Teachers practice and competency. In Q. Zhang (Eds) *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings* (pp. 251-266). Springer, Singapore.

International Society for Technology in Education & the Computer Science Teachers Association (ISTE & CSTA) (2011). *Operational definition of computational thinking for K–12 education.* Retrieved from http://www.iste.org/docs/ct-documents/computational-thinking-operational-definition-flyer.pdf

Khine, M. S. (2020). Objective measurement in psychometric analysis. In Khine, M. S. (Ed.). *Rasch measurement applications in quantitative educational research* (pp. 3-7). Springer Nature Singapore Pte Ltd.

Kong, S.-C. (2016). A framework of curriculum design for computational thinking development in K-12 education. *Journal of Computers in Education*, 3(4), 377-394.

Lye, S. Y., & Koh, J. H. L. (2014). Review on teaching and learning of computational thinking through programming: What is next for K-12? *Computers in Human Behavior*, 41, 51-61.

McMillan, J. H., Hellsten, L. M., & Klinger, D. A. (2011). *Classroom assessment: Principles and practice for effective standards-based instruction (Canadian ed.)*. Toronto, ON: Pearson.

Mueller J., Beckett D., Hennessey E., & Shodiev H. (2017). Assessing computational thinking across the curriculum. In P. Rich & C. Hodges (Eds). *Emerging research, practice, and policy on computational thinking. Educational Communications and Technology: Issues and Innovations* (pp. 251 – 267). Springer, Cham.

Neuman, W. L. (2014). *Social research methods: Qualitative and quantitative approaches (7th Edition)*. United States of America: Pearson Education Limited.

Papert, S. (1996). An exploration in the space of mathematics educations. *International Journal of Computers for Mathematical Learning*, 1(1), 95─123.

Rijke, W.J., Bollen, L., Eysink, T.H., Tolboom, J.L. (2018). Computational thinking in primary school: An examination of abstraction and decomposition in different age groups. *Informatics in Education*, 17(1), 77-92.

Roman-Gonzalez, M. (2014). Aprender a programar 'apps' como enriquecimiento curricular en alumnado de alta capacidad. Bord_on. *Revista de Pedagogía*, 66(4), 135-155.

Roman-Gonzalez, M. (2015). Computational thinking test: Design guidelines and content validation. *Proceedings of EDULEARN15 Conference*, 2436-2444.

Roman-Gonzalez, M., Perez-Gonzalez, J.-C., & Jimenez-Fernandez (2017). Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test. *Computers in Human Behvavior*, 72, 678-691.

Román-González M., Moreno-León J., & Robles G. (2019) Combining assessment tools for a comprehensive evaluation of computational thinking interventions. In S. C. Kong & H. Abelson (Eds) *Computational thinking education* (pp. 79-98). Springer, Singapore.

Sekaran, U. (2003). *Research method for business—A skill building approach (4th ed.).* John Wiley & Sons, Inc.

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2020). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. Journal of Computers in Education. Advance online publication. https://doi.org/10.1007/s40692-020-00177-2

Shute, V. J., Chen, S., & Asbell-Clark, J. (2017). Demystifying computational thinking. *Educational Research Review*, 22(2017), 142–158.

Stoilescu, D., & Egodawatte, G. (2010). Gender differences in the use of computers, programming, and peer interactions in computer science classrooms. *Computer Science Education*, 20(4), 283-300.

Sullivan, A., & Bers, M. U. (2016). Girls, boys, and bots: Gender differences in young children's performance on robotics and programming tasks. *Journal of Information Technology Education: Innovations in Practice*, 15, 145–165.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan [Application of rasch modelling in educational measurement]*. Cimahi: Trimkom Publishing House.

Sumintono, B. (2018). Rasch Model measurements as tools in assessment for learning. *Proceedings of 1st International Conference on Education Innovation (ICEI 2017).* Atlantis Press.

Tabesh, Y. (2017). Computational thinking: A 21st Century Skill. *Olympiads in Informatics*, 11, Special Issue, 65–70.

Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. *Computers & Education*, 148.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33─35.

Wing, J. M. (2008). Computational thinking and thinking about computing. Philosophical Transactions. Series A, Mathematical. *Physical, and Engineering Sciences*, 366(1881), 3717e3725.

Wing, J. M. (2011). *Research notebook: Computational thinking- what and why? The link.* The magazine of the Carnegie Mellon University School of Computer Science. Retrieved from http://www.cs.cmu.edu/link/research-notebookcomputational-thinking-what-and-why

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel–Haenszel DIF analysis. *Journal of Educational Measurement*, 36, 1–28.