Kong, S.C., Andone, D., Biswas, G., Hoppe, H.U., Hsu, T.C., Huang, R.H., Kuo, B.C., Li, K.Y., Looi, C.K., Milrad, M., Sheldon, J., Shih, J.L., Sin, K.F., Song, K.S., & Vahrenhold, J. (Eds.). (2019). *Proceedings of International Conference on Computational Thinking Education 2019.* Hong Kong: The Education University of Hong Kong.

The Measurement of Computational Thinking Performance Using Multiple-choice Questions

Yerkhan MINDETBAY[1*], Christian BOKHOVE[2], John WOOLLARD[3]
[123] University of Southampton, The United Kingdom
y.a.mindetbay@soton.ac.uk, c.bokhove@soton.ac.uk, j.woollard@soton.ac.uk

ABSTRACT
This study investigates the measurement of computational thinking performance of secondary school students using multiple-choice questions. The sample group of 775 grade eight students are drawn from 28 secondary schools across Kazakhstan. Students responded to a Computational Thinking Performance test of 50 multiple-choice questions. The test covers the concepts: logical thinking, generalisation and abstraction. The validity and reliability of the multiple-choice questions are determined using an Item Response Theory model. The item difficulty and discrimination coefficients are calculated, and the item characteristic curves for each question and test information functions for each quiz are generated. The results of the study show that the multiple-choice question assessment is a valid and reliable tool to measure computational thinking performance of students.

KEYWORDS
computational thinking, measurement, evaluation, multiple-choice questions, item response theory

1. INTRODUCTION
As computational thinking is becoming more popular trend in education, many countries integrated it into their national curricula. The most common way of delivering computational thinking in schools is through teaching computer programming, in some cases applying the pair programming technique and using unplugged activities (Bell, Witten, & Fellows, 2015) to teach computer science concepts in classrooms. The increased use of educational robots and programmable kits is also spreading the teaching of computational thinking. However, teaching methods are still in the early stage of development. The evaluation of computational thinking is as important as its integration into curricula, as without clear and verified assessment, attempts to integrate computational thinking into any curriculum cannot be verified. Moreover, in order to judge the effectiveness of computational thinking teaching strategies, measures must be approved that would allow teachers to assess what children learn (Grover & Pea, 2013). There is a need for standardized tests that can assess whether students can think computationally (Linn et al., 2010). The aim of this research is to establish a valid measurement of computational thinking performance of students by using multiple-choice questions.

1.1. Computational Thinking and Evaluation
Thinking is a mental process with a high-order cognitive function used in the process of making choices and judgments (Athreya & Mouza, 2017). The thinking process consists of lower-order and higher-order sub-processes, where a higher-order thinking is related to problem-solving, critical thinking, creative thinking, and decision-making. Computational thinking is a cognitive process, which reflects the ability to think in abstractions, algorithmically and in terms of decomposition, generalisation and evaluation (Selby, 2014, p.38). Computational thinking is related to spatial ability (Ham, 2018), academic success (Ambrosio, Almeida, Franco, & Macedo, 2014; Durak & Saritepeci, 2017; Gouws, Bradshaw, & Wentworth, 2013) and problem-solving ability (Román-González, Pérez-González, & Jiménez-Fernández, 2016).

2. METHODOLOGY
A bespoke computational thinking assessment was designed because most of the assessment tools for computational thinking are based on particular programming languages (Jamil, 2017) or some specific tools (Moreno-Leon & Robles, 2015; Oluk & Korkmaz, 2016; Seiter, 2015; Weese, 2016; Zhong, Wang, Chen, & Li, 2015). Context-specific evaluations of computational thinking might be biased due to students' prior knowledge and experience in those particular programming languages or tools. In this study, the test is more neutral as it is not a language-specific measurement. The national curricula of the Kazakhstani schools, the annual plans of "Bilim Innovation" Lyceums and students' problem-solving experience have been explored in order to construct test questions. The multiple-choice test is written taking into consideration the national curriculum, annual plans for informatics and Informatics textbooks (Shaniyev et al. 2017) and students' experience with problem solving.

2.1. Multiple-choice questions
As a frequently used assessment type in school, multiple-choice questions (MCQ) have several advantages including: efficiency for large-scale studies (Becker & Johnston, 1999; Dufresne, Leonard, & Gerace, 2002; Roberts, 2006); accuracy (Holder & Mills, 2001); objectivity (Becker & Johnston, 1999; Haladyna & Steven, 1989; Simkin & Kuechler, 2005; Zeidner, 1987); and compatibility with classical and item response theories (Haladyna & Steven, 1989). Multiple-choice questions are the most suitable format for assessment of higher-order cognitive skills and abilities (Downing & Haladyna, 2006), such as problem-solving, synthesis, and evaluation; and they are more effective on improving learning (Haynie, 1994; Smith & Karpicke, 2014). The multiple-choice questions for this study have been carefully constructed in line with the

context relevant recommendations on writing good multiple-choice items provided by the authors Downing & Haladyna (2006), Frey et al. (2005), Gierl et al. (2017) and Reynolds et al. (2009). In addition, two experts with experience in assessing computational thinking reviewed these test questions. Each item in this multiple-choice test has four response options, with one correct answer and three distractors. There are 50-multiple-choice questions (set of 5 quizzes with 10 questions each) in this test with maximum score of 50. It is conducted online with a duration of 100 minutes.

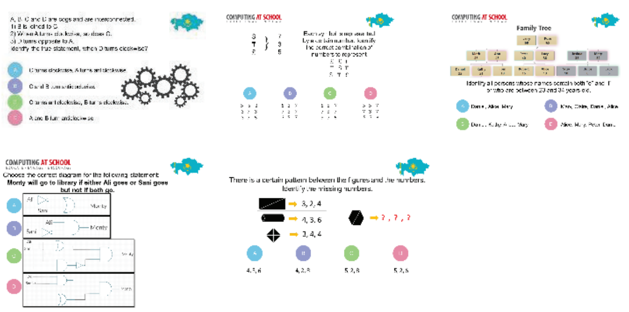

Figure 1. Sample questions.
Note: The sample questions can be accessed through:
http://bit.ly/SampleQs

2.2. Item Response Theory

Item Response Theory (IRT) is a paradigm for the design analysis and scoring of test instruments that measures attitudes, abilities and other variables. This theory is based on the relationship between person's performance on a test item and the person's performance level on an overall measure of the ability the item was constructed to measure. IRT is based on a mathematical model, which describes in probabilistic terms, how a test taking person with a higher standing on a trait is likely to respond in a different response category to a person with a low standing on the trait (Ostini & Nering, 2006). IRT has several advantages over traditional test theory, such as, sample independency, measurement of range of different abilities, accounting item difficulty, accounting for guessing, and supporting adaptive testing (Thissen & Wainer, 2001).

3. DATA ANALYSIS

The responses for multiple-choice questions were converted into dichotomous items, 0s for wrong responses and 1s for correct responses. These data from 775 13-14 year old participants are tested according to 2-parameter and 3-parameter IRT models. These data are collected from 775 (549 boys, 226 girls) 8^{th} grade students aged 13-14 years from Kazakhstan. The relationship between the probability of correct response to an item and the ability scale is described by the item characteristic curve (Baker & Kim, 2017). The item difficulty is a location index that shows where the item is located along the ability scale. An easy item functions among the low-ability students, a hard item functions among the high-ability students. The discrimination of an item, tells how well an item can differentiate between students with the abilities below the item location and those with the abilities above the item location. The item discrimination shows the steepness of the item characteristic curve in its middle section of the plot. The steeper the curve the better the item can discriminate; the flatter the curve the less the item can discriminate (Baker & Kim, 2017). The item discrimination parameter is "a". The item difficulty parameter is "b". The guessing parameter is "c". A 2-parameter IRT model suits better in this study, as the guessing parameter "c" is found as non-significant in 3-parameter model. The coefficients of item difficulty and item discrimination are presented in tables 1 and 2. The item characteristic curve plots are presented for each quiz in figures 2-6. The Cronbach Alpha is calculated for the items based on the responses from the sample size of 775. For the IRT analysis, the "mirt" and "ltm" libraries were used in RStudio.

4. RESULTS

The difficulty coefficients of majority of items are between the range of -0.8 and 1.3. All item characteristic curves for items fit well except for three items, item1(X1-black curve in Figure 4) and item6(X6-pink curve in Figure 4) in Abstraction quiz (Q3), item7(X7-yellow curve in Figure 5) in Pattern Numbers (Q4) quiz with the difficulty coefficients of 3.0, 1.8 and 2.0 respectively as shown in Table 1. The Cronbach Alpha (Field, 2013) coefficient for all 50 items is 0.87 (>0.7).

Table 1. The coefficients of item difficulty.

	Q1	Q2	Q3	Q4	Q5
Item1	-0.5	-0.8	**3.0**	0.9	-0.4
Item2	-0.2	-0.4	-0.8	0.1	0.2
Item3	0.2	0.1	-0.6	0.2	0.5
Item4	-0.3	-0.1	-0.1	0.0	0.4
Item5	0.2	-0.4	0.3	0.1	-0.1
Item6	0.0	0.0	**1.8**	-0.4	0.4
Item7	-0.2	1.0	-0.2	**2.0**	0.4
Item8	0.6	0.0	0.3	-0.1	0.4
Item9	1.3	0.2	0.5	0.2	1.1
Item10	1.2	0.5	0.2	0.0	0.3

The discrimination coefficients are between the range of 0.3 and 2.3 as shown in Table 2. The test information functions for each quiz show that the average ability respondent is tested the best.

Table 2. The coefficients of item discrimination.

	Q1	Q2	Q3	Q4	Q5
Item1	0.9	1.1	0.3	1.2	0.8
Item2	1.3	1.2	1.0	1.6	0.9
Item3	1.2	1.6	1.5	2.1	1.1
Item4	1.8	1.9	1.4	1.9	1.1

Item5	1.2	2.1	1.0	2.3	1.2
Item6	1.9	1.5	0.6	1.7	1.2
Item7	2.3	1.4	1.6	0.6	1.5
Item8	1.4	1.9	1.4	1.4	1.3
Item9	0.8	1.7	1.0	2.2	1.4
Item10	1.1	1.3	1.2	1.6	1.1

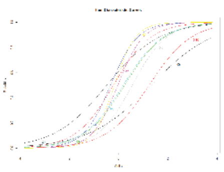

Figure 2. Logic quiz (Q1). http://bit.ly/Q1Logic

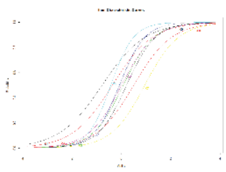

Figure 3. Logic quiz (Q2). http://bit.ly/Q2Logic

Figure 2 shows the individual item characteristic curves for the 10 items of the Logic Narrative quiz (Q1). Figure 3 shows the individual item characteristic curves for the 10 item of the Logic quiz (Q2). All lines ascend steeply showing a good discrimination coefficient.

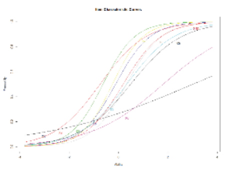

Figure 4. Abstraction quiz (Q3). http://bit.ly/Q3Abstraction

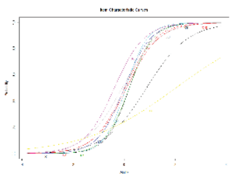

Figure 5. Pattern quiz (Q4). http://bit.ly/Q4Pattern

Figure 4 shows the individual item characteristic curves for the 10 items of the Abstraction quiz (Q3). The 2 outlier questions (items 1 and 6) being clearly identified by their more horizontal nature. Figure 5 shows the individual item characteristic curves for the 10 items of the Pattern quiz (Q4). The more difficult item being clearly identified by its displacement.

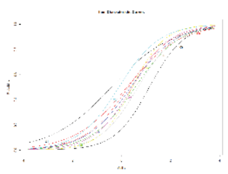

Figure 6. Pattern Figures quiz (Q5). http://bit.ly/Q5Pattern

5. DISCUSSION AND CONCLUSION

In this research, to measure students' computational thinking performance, 50 multiple-choice questions were specially designed with the focus on the concepts: logics, abstraction and generalisation. For the validity and reliability of the measurement 2-parameter IRT model and Cronbach Alpha test were used. Out of 50 items, 3 items were outliners as they were found difficult and were less informative in measurement. No too easy items were found. Test information functions for each quiz show that the most information is obtained for the average ability. The Cronbach Alpha result, the item difficulty and discrimination coefficients, the test information functions and the item characteristic curves are indications to justify the establishment of the validity and reliability of the multiple-choice questions to measure computational thinking performance of students.

6. REFERENCES

Ambrosio, A. P., Almeida, L. da S., Franco, A., & Macedo, J. (2014). Exploring Core Cognitive Skills of Computational Thinking. *Proceedings of PPIG 2014 - 25th Annual Workshop*. Sussex: University of Sussex.

Athreya, B. H., & Mouza, C. (2017). *Thinking Skills for the Digital Generation: The Development of Thinking and Learning in the Age of Information*. Cham: Springer International Publishing.

Baker, F. B., & Kim, S. H. (2017). *The Basics of Item Response Theory Using R*. Cham: Springer International Publishing.

Becker, W. E., & Johnston, C. G. (1999). The Relationship between Multiple Choice and Essay Response Questions in Assessing Economics Understanding. *Economic Record*, *75*(231), 348-357.

Bell, T., Witten, I. H., & Fellows, M. (2015). *CS Unplugged*. *Computer Science Unplugged*. Retrieved March 5, 2017, from https://www.csunplugged.org.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of Test Development* (1st ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Dufresne, R. J., Leonard, W. J., & Gerace, W. J. (2002). Marking Sense of Students' Answers to Multiple-choice Questions. *The Physics Teacher*, *40*(3), 174-180.

Durak, H. Y., & Saritepeci, M. (2017). Analysis of the Relation between Computational Thinking Skills and Various Variables with the Structural Equation Model. *Computers & Education*, *116*, 191-202.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics (3rd ed.).* London: SAGE Publications Ltd.

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing Rules: Collective Wisdom. *Teaching and Teacher Education*, *21*(4), 357-364.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-choice Tests in Education: A Comprehensive Review. *Review of Educational Research, 87*(6), 1082-1116.

Gouws, L., Bradshaw, K., & Wentworth, P. (2013). First Year Student Performance in a Test for Computational Thinking. *Proceedings of SAICSIT '13*. ACM, 271-277.

Grover, S., & Pea, R. (2013). Computational Thinking in K–12: A Review of the State of the Field. *Educational Researcher*, *42*(1), 38-43.

Haladyna, T. M., & Steven, M. (1989). Validity of a Taxonomy of Multiple-choice Item-writing Rules.

Applied Measurement in Education, *7347*(January 2015), 37-41.

Ham, D. A. (2018). Spatial Thinking as a Path Towards Computational Thinking. In *Computational Thinking in Primary Education*, 103-122. Hershey: IGI Global.

Haynie, W. (1994). Effects of Multiple-choice and Short-answer Tests on Delayed Retention Learning. *Journal of Technology Education*, *6*(1), 1-7.

Holder, W. W., & Mills, C. N. (2001). Pencils Down, Computer Up: The New CPA Exam. *Journal of Accountancy*, *191*(3), 57-60.

Jamil, H. M. (2017). Smart Assessment of and Tutoring for Computational Thinking MOOC Assignments using MindReader. Retrieved October 17, 2017, from http://arxiv.org/abs/1705.00959

National Research Council (2010). *Report of a Workshop on The Scope and Nature of Computational Thinking*. National Academies Press.

Moreno-Leon, J., & Robles, G. (2015). Dr. Scratch: A Web Tool to Automatically Evaluate Scratch Projects. *Proceedings of the Workshop in Primary and Secondary Computing Education.* ACM, 132-133.

Oluk, A., & Korkmaz, Ö. (2016). Comparing Students' Scratch Skills with Their Computational Thinking Skills in Terms of Different Variables. *International Journal of Modern Education and Computer Science*, *8*(11), 1-7.

Ostini, R., & Nering, M. (2006). *Polytomous Item Response Theory Models*. California: SAGE Publications Ltd.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and Assessment in Education (2nd ed.)*. New Jersey: Pearson.

Roberts, T. S. (2006). The Use of Multiple Choice Tests for Formative and Summative Assessment. *Proceedings of Conferences in Research and Practice in Information Technology Series*. Australian Computer Society, Inc., 175-180.

Román-González, M., Pérez-González, J.C., & Jiménez-Fernández, C. (2016). Which Cognitive Abilities Underlie Computational Thinking? Criterion Validity of the Computational Thinking Test. *Computers in Human Behavior, 72,* 678-691.

Seiter, L. (2015). Using SOLO to Classify the Programming Responses of Primary Grade Students. *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. ACM, 540-545.

Selby, C. (2014). *How can the Teaching of Programming be Used to Enhance Computational Thinking Skills?* The United Kingdom: University of Southampton.

Shaniyev, Y., Gesen, I., Aidarbayev, N., Akhmetov, N., & Yerzhanov, E. (2017). *Informatics - A Bilingual Textbook - Grade 8* (1st ed.). Astana: Astana Kitap.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice Tests and Student Understanding: What is the Connection? *Decision Sciences Journal of Innovative Education*, *3*(1), 73-98.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval Practice with Short-answer, Multiple-choice, and Hybrid Tests. *Memory*, *22*(7), 784-802.

Thissen, D., & Wainer, H. (2001). *Test Scoring*. New Jersey: Lawrence Erlbaum Associates.

Weese, J. L. (2016). Mixed Methods for the Assessment and Incorporation of Computational Thinking in K-12 and Higher Education. *Proceedings of the 2016 ACM Conference on International Computing Education Research.* ACM, 279-280.

Zeidner, M. (1987). Essay Versus Multiple-choice Type Classroom Exams: The Student's Perspective. *Journal of Educational Research*, *80*(6), 352-358.

Zhong, B., Wang, Q., Chen, J., & Li, Y. (2015). An Exploration of Three-dimensional Integrated Assessment for Computational Thinking. *Journal of Educational Computing Research*, *53*(4), 562-590.