Advances in Social Science, Education and Humanities Research, volume 491
Proceedings of the International Joint Conference on Arts and Humanities (IJCAH 2020)

Rasch and Classical Test Theory Validation of Automated Assessment Tool for Measuring Students' Creativity in Computer Programming

Ekohariadi[1*] Yeni Anistyasari[1] Ricky Eka Putra[1]

[1] *Department* of Informatics, *Universitas Negeri Surabaya, Surabaya, East Java, Indonesia*
[*]*Corresponding author. Email: ekohariadi@unesa.ac.id*

ABSTRACT
Creative thinking is one of the skills that must be mastered by students in the 21st Century. Some students fail and do not proficient in computer programming because they have low creative thinking skills. One of the computer programming learning to increase creativity is digital storytelling. The most commonly used measurement of creativity is the Torrance Test of Creative Thinking - Figural (TTCT-F), which consists of fluency, flexibility, originality, and elaboration. The problem faced when using TTCT-F is that it requires a long time and high subjectivity, the measurement results therefore are inconsistent. This study proposes the validation process exploiting Rasch model of automatic assessment tools to measure student creativity. Discriminant validity likely exists between computational thinking and creative thinking. Values of fit indices are in the range between 0.92 and 1.22 for Infit MNSQ and between 0.93 and 1.12 for Outfit MNSQ. These mean that the items are homogeneous with other items in a measurement scale. The Range of logit measurement of creative thinking reveal the data is slightly skewed with respect to person ability. Item difficulties for the easiest item is originality scales and the most difficult item is elaboration scales.

Keywords: *Rasch model, automated assessment, creativity, computer programming*

1. INTRODUCTION

Computer programming is an applicable skill. Basic programming courses can be trained by programming languages such as C, C ++, Java, Scratch, and Python. For Informatics students, programming plays an important role in improving problem-solving skills and is an important means of interacting with computer systems. It is important for students not only to master programming principles, methods, and techniques, but also the ability to think computationally and creatively. Computer programming requires critical thinking, problem-solving, computational thinking, creative, and systems design skills. The need for computational and creative thinking has received wide attention so that students are ready to face life in the 21st century [1], [2].

Students and teachers frequently face numerous problems during learning. Programming concepts and language syntax become obstacles to learning programming. The mean score of the programming ability of first-year students was 22.89 out of 110 points [3]. Besides, the dropout rate is between 30 and 40 percent, which reveals how enthusiastic students learn to program [4]

The main challenge faced by students when programming is determining creative solutions according to certain requirements [5]. Some students fail in programming because they have poor creative thinking skills [6]. This results in low student motivation towards programming. Creative thinking is defined as thinking that is flexible, imaginative, and innovative which utilizes all person's skills and experiences [7]. In programming, creative thinking will produce smart new solutions in resolving computational problems and increase the effectiveness of a solution [8]. Creativity is thus required to create new products and technologies in which Informatics is a field with a high level of innovation. One form of learning in computer programming is to increase creativity is digital storytelling.

The most frequently used test to measure creativity is the Torrance Tests of Creative Thinking (TTCT) [9] and is adapted from Guilford's concept of divergent thinking. This test measures creativity at four scales; fluency,

flexibility, originality, and elaboration. Fluency is the ability to generate a large number of ideas or solutions to problems in a short period of time. Flexibility is defined as the ability to simultaneously propose various approaches to a particular problem. Originality is the ability to generate new and original ideas. Elaboration is the ability to systematize and organize the details of an idea in your head and implement it. One form of TTCT is a figural test (TTCT-F) with three activities: construction of images, completion of images and repeating lines [10].

Current assessment tools for measuring creativity are challenging to apply in education because they require trained people to assess responses, but this preference is slow, expensive and subjective. The results of the assessment of creativity carried out among raters can vary and depend on the individual interpretation and knowledge of the assessors, therefore the inter-rater reliability is low [11]. This initiates the measurement results to be inconsistent so as to an objective assessment tool is compulsory and objective. Therefore this study proposes a new method to utilize computer technology to measure students' creativity in computer programming using a web-based automated assessment tool.

Based on the results of the literature review, the automated assessment tool for measuring creativity based on TTCT-F developed in this study is a novel method and has not been deeply investigated by other researchers. Another novelty in this study is the algorithm in the automated assessment tool to measure fluency, flexibility, originality, and elaboration, which is a development of existing image processing algorithms that provide outperform results in image classification. Furthermore, to prove that the automated assessment tool is valid and reliable in measuring students' creativity in computer programming, an analysis was carried out using classical test theory and the Rasch model (item response theory).

1.1. Related Works

1.1.1. Classical Test Theory

Validity, as applied to a test, is a judgment or estimate of how well a test measures what it purports to measure in a particular context. More specifically, it is a judgment based on evidence about the appropriateness of inferences drawn from test scores. One way measurement specialists have traditionally conceptualized validity is according to three categories:1). content validity, 2). criterion-related validity, and 3). construct validity

Construct validity is a judgment about the appropriateness of inferences drawn from test scores regarding individual standings on a variable called a *construct.* A number of procedures may be used to provide different kinds of evidence that a test has construct validity. The various techniques of construct validation may provide evidence, for example, that validity coefficient showing little (that is, a statistically insignificant) relationship between test scores and/or other variables with which scores on the test being construct-validated should *not* theoretically be correlated provides discriminant evidence of construct validity (also known as *discriminant validity*) [12].

A successful evaluation of discriminant validity shows that a test of a concept is not highly correlated with other tests designed to measure theoretically different concepts.

In showing that two measures do not correlate, it is necessary to correct for attenuation in the correlation due to measurement error. It is possible to calculate the extent to which the two measures overlap by using the following formula where r_{xy} is the observed correlation between x and y, r_{xx} is the reliability of x, and r_{yy} is the reliability of y:

$$\hat{r}_{xt\,yt} = \hat{r}_{xy} / (\hat{r}_{xx}\hat{r}_{yy})^{1/2}$$

The r_{xtyt} is the correlation between the true scores of the measures x and y or the construct underlying the measure *x* dan the construct underlying the measure *y* [13]. The above formula is called the *attenuation formula,* because it shows how measurement error in the *x* and *y* measures reduces the observed (computed) correlation (r_{xx}) below the true score correlation *(r_{xy}).* Although there is no standard value for discriminant validity, a result less than 0.85 suggests that discriminant validity likely exists between the two measures.

1.1.2. Rasch Model

Initially, the item response theory (IRT) model was developed to deal with items that were recorded as true-false or dichotomous. In IRT, the mathematical model for grain characteristic curves is a cumulative form and a logistic function. There are three models, namely the one-parameter, two-parameter, and three-parameter logistic model (1-PL, 2-PL, and 3-PL).

Rasch measurement theory (RMT) is a simple logistic unidimensional measurement model that satisfies fundamental measurement requirements [14]. RMT is applied when a set of items in a scale are intended to be summed together to represent a common unidimensional latent variable [15]. Unless unidimensionality has been established, it is not valid to add together the scores for any set of items [16].

2. METHOD

2.1. Research Design

This research has been conducted in the Department of Informatics, Faculty of Engineering, Unesa. The number of respondents is 70 students. The study used a one-group pretest-posttest design [17]. Subjects carry out basic programming learning with visual programming language

(X). Furthermore, students are given individual assignments (O). Each student is given the freedom to determine what they will do. The project assignment is digital storytelling.

2.2. Creativite Thinking Skills Assessment

The project assignments specified during the lesson are individual tasks, each student is offered the self-determination to decide what they do. The project assignment is in digital storytelling. Through digital storytelling, students can express their creativity. The indicators of creativity are fluency, flexibility, originality, and elaboration. The results of the creativity indicators and assessment rubric are presented in Table 1.

Table 1. Creative thinking scoring indicators and rubrics

Indicator	Rubric
Fluency	Fluency is the total number of responses, therefore, student's digital story-telling time-duration is counted. The scoring of fluency is as follows. The score lies in 0-2. Maximum score is awarded if the digital story-telling is more than 120 seconds. Conversely, minimum score is given if its time duration is less than 60 seconds. Thus, student whose digital story-telling results are between 60 and 120 seconds is granted score 1.
Flexibility	Flexibility is defined as the degree of difference of the responses,that is the number of categories that the answers cover. This definition is assumed as number of created characters in digital story-telling. If students are able to create more than 20 kinds of characters, they honor maximum score or 2. Otherwise, if they create less than 10 characters, a minimum score is given. This is to say, score 1 is obtained if 10 - 20 kinds of characters are created.
Originality	Originality is expressed by statistical infrequence of each response, each response was compared to the total amount of responses from all the participants. In other words, originality is measured by the similarity response among participants. The greater similarity result, the less originality score. The similarity score lies between 0 - 1. In similarity, score 1 implies the response is similar to others and vice versa. Since originality score is iversely proportional to similarity score, maximum score of originality (i.e. 2) is assigned if the score of similarity is 0. The other way about, score 0 of originality is allocated if the similarity stays in 1-0.5 . In addition, score 1 of originality is honored if the similarity score is between 0.51 and 0.05.
Elaboration	Elaboration is the level of detail of the response. It can be assumed as image details or image intensity which the score lies between 0-256. Score 0 of elaboration is granted if the image details score is less then 100; score 1 of elaboration is assigned if image details score is greater than 200. Score 1 of elaboration , thus, is obtained if the details score is between 100-200.

2.3. Analysis

Data from 4 items were analyzed exploiting the Rasch-Masters Partial Credit model. In general, the IRT models a test situation where a person answers a set of items. In this study the higher the potential for creativity, the greater the probability of people getting high scores. The probability also depends on the difficulty of the item.

In the Rasch model the item has only the difficulty level characteristic (b), therefore the Rasch model is termed a one parameter model. The Rasch model can be used for dichotomous items (scores 0 and 1) and polytomous items (scores 0, 1, 2) as in this study. Rasch Model for polytomous items Rasch-Masters Partial Credit Model [18].

The PC model was developed to analyze test items which required several steps in the completion process. The PC model can be considered an extension of the Rasch model. Item i is scored $x = 0, 1, \ldots, m_i$ for an item with k response categories ($k = m_i + 1$). Masters proposes a general equation about the probability that a test taker has the ability θ and gets a score of x in item i as follows:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^{x}(\theta - \delta_{ij})\right]}{\sum_{k=0}^{m_i}\left[\exp\sum_{j=0}^{k}(\theta - \delta_{ij})\right]} \quad x = 0, 1, \ldots, m_i$$

The notation δ_{ij} is defined as step difficulty of the item which relates to the transition from one category (score) to the next. For the notation convention, $\sum (\theta - \delta_{ij})$ is determined to be zero when $k = 0$ [19].

3. RESULT AND DISCUSSION

3.1. Discriminant Validity

The two tests being compared are creative thinking and computational thinking. Creative thinking construct consists of fluency, flexibility, originality, and elaboration. On the other hands, computational thinking construct consists of flow control, data representation, abstraction, user interactivity, synchronization, and logic. Furthermore, computational thinking is a major competency taught in computer programming courses. Tables 2, 3, and 4 respectively depict the reliability values of computational thinking, creative thinking, and correlation between computational thinking and creative thinking.

Table 2. Summary of reliability of computational thinking

Item	Item-test correlation	Alpha
Flow_Control	0.59	0.75
Data_Representation	0.80	0.70
User_Interactivity	0.72	0.72
Synchronization	0.56	0.76
Parallelism	0.71	0.74
Logic	0.81	0.72
Test scale		0.77

Table 3. Summary of reliability of creative thinking

Item	Item-test correlation	Alpha
Fluency	0.80	0.12
Flexibility	0.82	0.01
Originality	0.52	0.63
Elaboration	0.18	0.52
Test scale		0.45

Table 4. correlation between computational thinking and creative thinking

	Computational thinking	Creative thinking
Computational thinking	1.000	
Creative thinking	0.041	1.000

From data above, we can find the correction for attenuation formula as follow:

$$0.041/(0.45x0.77)^{1/2} = 0.07$$

Although there is no standard value for discriminant validity, a result less than 0.85 suggests that discriminant validity likely exists between computational thinking and creative thinking.

3.2. Fit of Items in Measuring Constructs

Examination of the conformity of the constructed item refers to the value recorded in the infit and outfit MNSQ index. We have to look carefully at the values of these indeces to determine whether the item developed appropriate (item fit) to measure a latent variable or construct. Based on the Bond & Fox [20], a study to determine the suitability item built, the infit and outfit MNSQ should be in the range between 0.6 to 1.4. If the items are out of the range, it must be separated, modified or rephrased [21].

Table 5 shows the measurement of fit items that fit the Rasch measurement model for creative thinking. Values are in the range between 0.92 and 1.22 for Infit MNSQ and between 0.93 and 1.12 for Outfit MNSQ. These mean that the items are homogeneous with other items in a measurement scale.

Table 5. Item fit index of creativity scale

Item	Estimate	Infit MNSQ	Outfit MNSQ
Fluency	-0.225	0.92	0.93
Flexibility	-0.745	0.93	0.94
Originality	-1.130	1.22	1.12
Elaboration	2.100	1.21	1.06

Table 5 shows the fit statistics for creative thinking scale items. MNSQ values of between 0.92 and 1.22 indicate that the items are useful for measurement [21].

3.3. Item Difficulty and Person Ability

The Rasch model can be used to compare item difficulty and person ability by placing them on the same logit scale in the form of a vertical plot referred to as a Wright map. Figure 1 shows the Wright map for the creative thinking data.

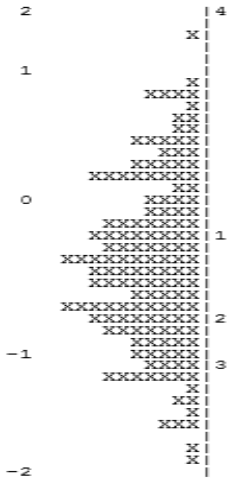

Figure 1. Wright map of creative thinking data showing logit values for person ability and item difficulty

The Wright map shows that the range of logit measures is from a low of −2 to a high of +2. This shows that the data is slightly skewed with respect to person ability. Item difficulties range from about −1.13 logits for the easiest item (originality) to about 2.10 logits for the most difficult item (elaboration).

3.4. Discussion

Researchers have long explored how to measure creativity. The question of how to assess creativity is still challenging [22]. Various approaches have been developed, ranging from methods based on expert judgment for assessing product quality (eg, Consensual Assessment Technique [23]) to frequency-based methods that employ standardized norms [24]. Although each method has several uses for creativity research, each has its limitations. The two most common limitations are problems of subjectivity (raters don't always agree on what creativity is) and labor cost (raters often score multiple responses). Both of these limitations reduce a reliable and valid assessment of creativity [25]. To solve this problem, this study investigates whether the creative quality scoring process can be done automatically using computational methods.

Creative thinking is not an innate talent but a process which is an integral component of human intelligence that can be practiced, encouraged and developed in the context of computer programming. Otherwise, learning computer programming also emphasizes computational thinking competencies that must be mastered by students. The question is whether the two constructs are different or the same. The results of the discriminant validity test show that creative thinking is different from computational thinking.

The development of creative thinking has become the attention of the global community. Students should play an active role in acquiring these skills by being knowledge creators, not information consumers. In the context of computer programming, creative thinking skills can be enhanced by computational thinking which includes the process of identifying aspects of computation and applying existing tools and techniques in computer science. In this context, computational problems can be solved, computational artifacts can be created and students have the opportunity to express themselves creatively.

The Rasch model entails unidimensional and additive requirements. Unidimensional means that all test items measure a construct. Additive refers to the properties of the unit of measurement. This unit is called the logit and has the property of maintaining the same size (interval) over the entire continuum. The interval size was then used for statistical analysis. In the Rasch model, person size and grain size each have an order on a common logit scale.

An important characteristic of a set of test items that measure a construct is that these items are unidimensional. In Rasch's analysis, if all coherent items form one scale, then the items are unidimensional. Keeves and Masters [26] proposed that item fit was used to check for unidimensionality. The research findings show that the creative thinking instrument is unidimensional. The average level of student creativity is parallel with the average level of item difficulty.

4. CONCLUSION

Discriminant validity likely exists between computational thinking and creative thinking. Values of fit indices are in the range between 0.92 and 1.22 for Infit MNSQ and between 0.93 and 1.12 for Outfit MNSQ. These mean that the items are homogeneous with other items in a measurement scale.

Range of logit measures of creative thinking is from a low of −2 to a high of +2. This shows that the data is slightly skewed with respect to person ability. Item difficulties range from about −1.13 logits for the easiest item (originality) to about 2.10 logits for the most difficult item (elaboration).

REFERENCES

[1] Vaca-Cardenas *et al.*, “Coding with Scratch: The design of an educational setting for Elementary pre-service teachers,” in *Proceedings of 2015 International Conference on Interactive Collaborative Learning, ICL 2015*, 2015, pp. 1171–1177.

[2] L. Ma, J. Ferguson, M. Roper, and M. Wood, “Investigating and improving the models of programming concepts held by novice programmers,” *Comput. Sci. Educ.*, vol. 21, no. 1,

pp. 57–80, Mar. 2011.

[3] T. Menzles and F. Shull, *Making Software What Really Works, and Why We Believe It*. Cambridge: O'Reilly, 2011.

[4] T. Beaubouef and J. Mason, "Why the High Attrition Rate for Computer Science Students: Some Thoughts and Observations," 2005.

[5] P. Schaumont and I. Verbauwhede, "The exponential impact of creativity in computer engineering education," in *2013 IEEE International Conference on Microelectronic Systems Education, MSE 2013*, 2013, pp. 17–20.

[6] L. D. Miller *et al.*, "Improving learning of computational thinking using creative thinking exercises in CS-1 computer science courses," in *Proceedings - Frontiers in Education Conference, FIE*, 2013, pp. 1426–1432.

[7] D. C. McClelland, *Human Motivation*. Cambridge University Press, 1988.

[8] M. S. Peteranetz, A. E. Flanigan, D. F. Shell, and L. K. Soh, "Computational Creativity Exercises: An Avenue for Promoting Learning in Computer Science," *IEEE Trans. Educ.*, vol. 60, no. 4, pp. 305–313, Nov. 2017.

[9] J. P. Guilford, *Fundamental Statistics in Psychology and Education*, vol. 41, no. 3. Wiley, 1957.

[10] K. H. Kim, "The Torrance Tests of Creative Thinking - Figural or Verbal: Which One Should We Use?," *Creat. Theor. – Res. - Appl.*, vol. 4, no. 2, pp. 302–321, Feb. 2018.

[11] D. H. Cropley and J. C. Kaufman, "Measuring functional creativity: Non-expert raters and the creative solution diagnosis scale," *J. Creat. Behav.*, vol. 46, no. 2, pp. 119–137, Jun. 2012.

[12] R. J. Cohen, M. E. Swerdlik, and S. M. Phillips, *Psychological testing and assessment: An introduction to tests and measurement, 3rd ed. - PsycNET*. Mayfield Publishing Co, 1996.

[13] F. L. Schmidt and J. E. Hunter, "Measurement error in psychological research: Lessons from 26 research scenarios.," *Psychol. Methods*, vol. 1, no. 2, pp. 199–223, 1996.

[14] "Rasch Models for Measurement | SAGE Publications Inc." [Online]. Available: https://us.sagepub.com/en-us/nam/rasch-models-for-measurement/book2266. [Accessed: 29-Aug-2020].

[15] A. Tennant and P. G. Conaghan, "The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper?," *Arthritis Care and Research*, vol. 57, no. 8. Arthritis Rheum, pp. 1358–1362, 15-Dec-2007.

[16] David L. Streiner, Geoffrey R. Norman, and John Cairney, *Health Measurement Scales: A practical guide to their development and use - Oxford Medicine*. Oxford: Oxford University Press, 2014.

[17] Geoffrey E. Mills and L. R. Gay, *Educational Research: Competencies for Analysis and Applications | Pearson*. Florida: Pearson, 2019.

[18] G. N. Masters, "A rasch model for partial credit scoring," *Psychometrika*, vol. 47, no. 2, pp. 149–174, Jun. 1982.

[19] S. E. Embretson and S. P. Reise, *Item response theory for psychologists.* . Washington DC: Lawrence Erlbaum Associates Publishers., 2000.

[20] Trevor G Bond and Christine M. Fox, *Applying the Rasch Model; Fundamental Measurement in the Human Sciences | Request PDF*. Routledge, 2015.

[21] John Michael Linacre, "Test validity and rasch measurement: Construct, content, etc | Request PDF," 2004.

[22] B. Barbot, R. W. Hass, and R. Reiter-Palmon, "Creativity assessment in psychological research: (Re)setting the standards," *Psychol. Aesthetics, Creat. Arts*, vol. 13, no. 2, pp. 233–240, May 2019.

[23] T. M. Amabile, "The social psychology of creativity: A componential conceptualization," *J. Pers. Soc. Psychol.*, vol. 45, no. 2, pp. 357–376, Aug. 1983.

[24] B. Forthmann, S. H. Paek, D. Dumas, B. Barbot, and H. Holling, "Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates," *Br. J. Educ. Psychol.*, vol. 90, no. 3, pp. 683–699, Sep. 2020.

[25] R. Reiter-Palmon, B. Forthmann, and B. Barbot, "Scoring divergent thinking tests: A review and systematic framework," *Psychol. Aesthetics, Creat. Arts*, vol. 13, no. 2, pp. 144–152, May 2019.

[26] G.N. Masters and J.P. Keeves, *Advances in Measurement in Educational Research and Assessment*. Emerald Group Publishing Limited, 1999.