Development and Validation of Metacognitive Test in Programming Using Graded Response Model

Ni Made Sri Mertasari[1], I Made Candiasa[2]

{srimertasari@undiksha.ac.id[1], candiasa@undiksha.ac.id[2]}

[1,2]Universitas Pendidikan Ganesha

Abstract. Metacognition includes explaining knowledge that has been mastered and choosing strategies to master new knowledge. Therefore, metacognitive measurement in programming is very important because programmed problems constantly change dynamically. This study aims to develop and validate metacognitive tests in programming using the Graded Response Model (GRM). Metacognitive tests in programming were developed from indicators of cognitive knowledge and cognitive regulation in the form of descriptive tests. The probability of participants answering the test items is estimated from the likelihood of participants answering each level of completion of the test items. The levels of work on each item have a graded response, so the analysis is carried out with GRM. GRM considers two characteristics of the item: the level of difficulty and the level of discrimination. The results showed that metacognitive test items in programming were suitable to be developed with the graded responses model (GRM). The probability of participants correctly answering one test item is relevant to the test's ability and level of difficulty. In addition, the difference in the likelihood of answering one test item is also applicable to the differentiating power of the test.

Keywords: metacognitive, programming, graded response model, level of difficulty, level of discrimination

1 Introduction

Programming competence is indispensable in the current and future technological era [1]. Unfortunately, programming still needs to be improved to learn [2]. Students experience many difficulties in programming courses [3]. Early learning programming sometimes fails to motivate students to continue because students find it challenging to master programming logic [4]. Many studies have been carried out for the programming learning process to track common and individual barriers in learning programming [5]. One of the variables that can affect the success of programming lectures is metacognitive [6], [7].

Metacognition was first coined by Flavel [8] as the ability to understand one's abilities, which definition is supported by many experts, such as Van der Stel [9] and Cantoia et al. [10]. Also,

IConVET 2022, October 06, Singaraja, Indonesia

DOI 10.4108/eai.6-10-2022.2327430

very important description of metacognition is the ability to understand what is known and strategies for learning what is not known [11]. If the individual can explain the knowledge he already has and know the knowledge that must be possessed, then the individual is said to have metacognitive. As an implication, the individual can learn further to explore the next knowledge on an ongoing basis.

Metacognitive abilities are conceptualized as a set of interrelated competencies for learning and thinking and include many of the skills required for active learning, critical thinking, reflective judgment, problem-solving, and decision-making [12]. Metacognitive ability can be measured from competence to learn, think critically, solve problems, make decisions, and make choices. Metacognitive abilities develop and contribute to learning performance, partly independent of intelligence [13]. Throughout its development, metacognition becomes more explicit, powerful, and effective, as it increasingly operates under the control of the individual's consciousness [14].

This study aims to develop a cognitive ability test that can be used in programming courses. The instrument for measuring metacognitive ability in mathematics learning is guided by two indicators of metacognitive ability formulated by Schraw and Denisson [15], namely cognitive knowledge and cognitive regulation. Knowledge consists of three sub-indicators, namely declarative knowledge, procedural knowledge, and conditional knowledge. Meanwhile, the regulation consists of five sub-indicators: planning, regulation or management of information, processing or calculation, control, and evaluation. The instrument was developed in the form of non-routine mathematical problems. Students are asked to solve these problems by answering several questions related to metacognitive indicators in programming.

The instrument developed was validated using item response theory, generally abbreviated as IRT. Item response theory has an item orientation, thereby eliminating the dependence between test items and test takers (the concept of parameter invariance). In addition, the test taker's response to one test item does not affect the other test items (the concept of local independence), and the test items only measure one measurement dimension (unidimensional concept). Hambleton, Swaminathan, and Rogers suggest several characteristics of item response theory [16]. The characteristics of an item do not depend on the test taker. The score described by the test taker is independent of the test. The model in item response theory emphasizes more on the item level of the test compared to the test. The model does not require strictly parallel tests to estimate reliability. The model describes a decision measure for each ability score, namely the functional relationship between the test taker's ability level and the test taker himself.

The item response theory for the dichotomous test differs from the item response theory for the polytomy test. There are several models of item response theory for the dichotomous test, namely the one-parameter logistic (1PL), two-parameter logistic (2PL) parameter, three-parameter logistic (3PL), and the latest four-parameter logistic (4PL). For the polytomy test, several models can be used, including the Partial Credit Model (PCM), Graded Response Model (GRM), Generalized Partial Credit Model (GPCM), which is similar to PCM, Modified Graded Response Model (M-GRM), and Rating Scale Model (RSM). The graded Response Model (GRM) is a model of item response theory developed from a two-parameter logistic (2PL) model that uses difficulty and discrimination index.

The one-parameter logistic model (1PL) only considers one item characteristic: the difficulty level. The two-parameter logistic (2PL) model considers two item characteristics: the level of

difficulty and discriminating power. The three-parameter logistic (3PL) model considers three item characteristics: the level of difficulty, discrimination, and the opportunity for participants who are less able to make guesses. The three-parameter logistic (3PL) model considers three item characteristics: the level of difficulty, discriminating power, and the opportunity for participants to make guesses. The three-parameter logistic (3PL) model considers three item characteristics: the level of difficulty, discrimination, and the opportunity for participants who are less able to make guesses. The four-parameter logistic (4PL) logistic model considers four item characteristics: the level of difficulty, discriminatory power, opportunities for participants who are less able to make guesses, and opportunities for participants who are able to make mistakes.

The graded Response Model (GRM) is a model of item response theory for polytomy tests which is a development of the two-parameter logistic model (2PL). GRM uses the difficulty index and discriminatory parameters, as with the 2PL model, but the difficulty index parameter is obtained from each category on an item. The parameters of the difficulty index and the discriminating power of the questions will determine the probability of a test taker correctly answering a test item. In GRM each item can be obtained an estimate of the differentiating power of the items, and the difficulty index of each category in a test item is arranged sequentially so that the answers of the test takers must be sorted from the lowest category to the higher category. GRM was first developed by Samejima to be applied to item analysis with several levels in its completion, and these levels have a graded response [17].

The graded Response Model (GRM) is the right model to be used to analyze items that have categorical responses. According to Widhiarso, the GRM model is an indirect approach, so before determining the category response function (CRF) it must first be determined the characteristic operating function (OCF) of each category on a test item as the basis for deciding CRF [18]. The formula calculates OCF:

$$P\_ij^{*}(\theta)=\exp\left[(\alpha\_i(\theta\_n-\beta\_ij)\right] / (1+\exp\left[(\alpha\_i(\theta\_n-\beta\_ij)\right])$$

Where: $P\_ij^{*}(\theta)$= the probability that respondents who have the ability can answer the jth category on item i correctly; yθ_n= respondent ability parameters; β_ij= the difficulty index parameter of the jth category on item I; α_i= i-th discriminating index parameter. CRF is calculated by the formula: $P\_ij(\theta)=P\_ij^{*}(\theta)-P\_(i(j+1))^{*}(\theta)$ Under the condition: $P\_i0^{*}(\theta)=1$ and $P\_(i(j+1))^{*}(\theta)=0$. Where: P_ij (θ) = probability of item i-category j; $P\_ij^{*}(\theta)$= probability of item i of the earlier category; $P\_(i(j+1))^{*}(\theta)$ = probability of item i last category.

2 Method

The development model used in developing this product is a formative Research type development model with three stages: the preliminary stage, the self-evaluation stage, and the formative evaluation stage [19]. At the Preliminary stage, a review of several reference sources related to this research was carried out. Furthermore, a self-evaluation was carried out, including the design of the test grid, test preparation, and scoring guidelines. After the test is compiled, a formative evaluation is carried out through expert reviews, one-to-one, and a small group. The formative evaluation results are used as material for the final revision. The revised test results are subject to a field test. An important step in this field test is the GRM analysis. GRM is used to display the estimation of item parameters and students' abilities [20].

GRM is an IRT model for polytomy data developed for item responses characterized by categorical order [17].

Metacognitive tests in programming are tested on students who are taking programming courses. The test results were analyzed using the Grade Response Model (GRM). Two characteristics of the items needed in GRM are the level of difficulty and discriminating power. Therefore, the GRM calculation is preceded by calculating the item difficulty level and the item discriminative power. GRM is a model for stratified solutions with polytomy scores, where the item parameters are interpreted as the steps' difficulty level [21]. GRM is used for the observed polytomous ordered variables, implemented to estimate the item parameters and students' abilities [20]. GRM holds the important assumption that for a single item, the entire set of categories has homogeneous reasoning [21].

The steps for implementing GRM are as follows.

Calculate the item difficulty index. The item difficulty index can be calculated using the classical test model, but it must be converted to a modern test model.

Calculate the item discriminatory index. The item difference index can be calculated using the classical test model, but it must be converted to a modern test model.

Calculate the Operating Characteristic Function (OCF) for each category on an item with the formula: $P_{ij}^*(\theta)=\exp[(\alpha_i(\theta_n-\beta_{ij})]/(1+\exp[(\alpha_i(\theta_n-\beta_{ij})])$.

Calculate Category Response Function (CRF) all items using the formula: $P_{ij}(\theta)=P_{ij}^*(\theta)-P_{i(j+1)}^*(\theta)$ under the condition: $P_{i0}^*(\theta)=1$ and $P_{i(j+1)}^*(\theta)=0$.

3 Result and Discussion

The item difficulty index is in the range of -2 to 2. Meanwhile, all item discrimination indexes exceed 0.25. After calculating the difficulty index and discrimination index, the probability of the test takers begins with calculating the OCF with the assumption that the test participant's ability level is = 0.5. Thus, the OCF obtained from item 1 has a discriminatory index of 0.307 and the difficulty indices -0.55, -0.32, -0.08, 0.102, 0.122, 0.431, 0.521, 0.722 are as follows.

$P_{11}^*(\theta)=\exp[(0.307(0.5-(-0.55))]/(1+\exp[(0.307(0.5-(-0.55))])=0.57990$

$P_{12}^*(\theta)=\exp[(0.307(0.5-(-0.32))]/(1+\exp[(0.307(0.5-(-0.32))])=0.56260$

$P_{13}^*(\theta)=\exp[(0.307(0.5-(-0.08))]/(1+\exp[(0.307(0.5-(-0.08))])=0.5440$

$P_{14}^*(\theta)=\exp[(0.307(0.5-(0.102))]/(1+\exp[(0.307(0.5-(0.102))])=0.53051$

$P_{15}^*(\theta)=\exp[(0.307(0.5-(0.122))]/(1+\exp[(0.307(0.5-(0.122))])=0.52898$

$P_{16}^*(\theta)=\exp[(0.307(0.5-(0.431))]/(1+\exp[(0.307(0.5-(0.431))])=0.50530$

$P_{17}^*(\theta)=\exp[(0.307(0.5-(0.521))]/(1+\exp[(0.307(0.5-(0.521))])=0.49839$

$P_{18}^*(\theta)=\exp[(0.307(0.5-(0.722))]/(1+\exp[(0.307(0.5-(0.722))])=0.48297$

With the same approach, OCF was obtained for all items, as listed in the following table.

Table 1. OCF Metacognitive Test Items

Item Number	Category							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0.579897	0.562605	0.544398	0.530509	0.528979	0.505296	0.498388	0.482968
2	0.601419	0.577652	0.551242	0.523621	0.516649	0.49647	0.475689	0.469186
3	0.582139	0.571648	0.546681	0.529055	0.528137	0.509899	0.487569	0.477681
4	0.595517	0.569542	0.547974	0.526684	0.517492	0.498465	0.482202	0.473852
5	0.594777	0.584825	0.571122	0.546453	0.528673	0.505909	0.499923	0.484041
6	0.604359	0.582214	0.55890	0.530585	0.521247	0.502609	0.480975	0.467580
7	0.577726	0.571949	0.549039	0.52852	0.527219	0.511127	0.490638	0.490638
8	0.588547	0.578625	0.555188	0.528826	0.515882	0.497314	0.481512	0.472092

OCF cannot be used to compare the probability of each item category. Therefore it is necessary to continue to the next step by calculating the item CRF. Based on the OCF of each item, the CRF calculation for item number 1 is then carried out as follows.

$P\_11\ (\theta)=1-P\_i1^*\ (\theta)=1-0.579897=0.42010$

$P\_12\ (\theta)=P\_i1^*\ (\theta)-P\_i2^*\ (\theta)=0.579897-0.562605=0.01729$

With the same approach, CRF was obtained for all items, as listed in the following table.

Table 2. CRF Metacognitive Test Items

Item Number	Category							
	1	2	3	4	5	6	7	8
1	0.42010	0.01729	0.01821	0.01389	0.00153	0.02368	0.00691	0.01542
2	0.39858	0.02377	0.02641	0.02762	0.00697	0.02018	0.02078	0.00650
3	0.41786	0.01049	0.02497	0.01763	0.00092	0.01824	0.02233	0.00989
4	0.40448	0.02597	0.02157	0.02129	0.00919	0.01903	0.01626	0.00835
5	0.40522	0.00995	0.01370	0.02467	0.01778	0.02276	0.00599	0.01588
6	0.39564	0.02215	0.02331	0.02831	0.00934	0.01864	0.02163	0.01339
7	0.42227	0.00578	0.02291	0.02052	0.00130	0.01609	0.02049	0.00683
8	0.41145	0.00992	0.02344	0.02636	0.01294	0.01857	0.01580	0.00942

The table above shows the probability of test-takers answering the eight questions, divided into eight categories with the assumption that the standard test-taker's ability is =0.5. For item 1, category 1 has a difficulty index of -0.55 and a discrimination index of 0.307. The probability of test takers in correctly answering item 1 category 1 is 0.42010. It means, at the level of ability of test takers = 0.5 to get a perfect score, it is relatively easy. However, in category 2, with a difficulty index of -0.32 and a discrimination index of 0.307, the probability of a test taker with an ability of 0.5 to answer the overall test correctly is only 0.01729. The ability level of test takers =0.5 to get a perfect score is quite difficult. Based on the results of the calculations in Table 2, the same interpretation can be given for other test items.

The opportunity for test takers to answer the test items correctly is tiered from category 1 to category 2, from category 2 to category 3, and so on. The metacognitive test developed from

eight indicators formulated by Schraw and Denisson [15] does have levels. Declarative knowledge, procedural knowledge, and conditional knowledge in cognitive knowledge require test takers' ability to be tiered. Conditional knowledge requires relatively more abilities than procedural knowledge, and procedural knowledge requires relatively higher abilities than declarative knowledge. The same thing happened to the regulation, which consists of sub-indicators planning, regulating or managing information, processing or calculation, controlling, and evaluating. In this case, the evaluation demands the highest ability of the test takers.

The graded abilities required in the metacognitive tests match the GRM. GRM from Samejima is designed for items with some a priori order associated with the measured latent variable [22]. The discriminatory index for each category in one item is constant, but it does not mean it is constant for all items [21]. Suppose the level of item diffi-culty is relatively the same for category 1. In that case, the probability of participants answering correctly at category 1 is relatively the same for participants with the same ability. The probability of test takers answering the test items perfectly depends on the difficulty level in each item's subsequent categories.

This article is still limited to studying test takers with an ability of 0.5. Further analysis is still needed to get a complete picture of the probability of participants answering correctly for all categories on each item with varying student abilities. Thus obtained a clearer picture of the effectiveness of the application of GRM for the analysis of metacognitive tests in programming. In general, for the implementation of GRM, Lautenschlager, Meade, and KIM note that GRM is already in widespread use. However, performance in various conditions needs to be better understood [23].

4 Conclusion

Metacognitive tests in programming have levels, which in this study are divided into eight according to metacognitive sub-indicators. The cognitive knowledge indicator consists of three sub-indicators: declarative, procedural, and conditional. Meanwhile, cognitive regulation consists of five sub-indicators: planning, regulation or management of information, processing or calculation, control, and evaluation. These levels are used as a stage in analyzing test items with GRM. Preliminary research results show that GRM is relevant for analyzing metacognitive tests in programming. Metacognitive tests that are polyatomic are suitable for analysis with GRM. Two characteristics are considered in GRM: difficulty index and discriminating power. Further analysis is still needed to obtain comprehensive results involving participants with greater and more heterogeneous abilities.

References

[1] World Economic Forum. (2020). The Future of Jobs Report 2020. OCTOBER 2020.

[2] Webber, C.G., Possamai, R.: An Immune-based Approach to Evaluate Programming Learning. In 9th IFIP World Conference on Computers in Education (2009).

[3] Koscianski, A., Bini, E.: Tackling Barriers in the Learning of Computer Programming. In 9 th IFIP World Conference on Computers in Education (2009).

[4] Holvikivi, Jaana. (2010). Conditions for successful learning of programming skills. IFIP TC 3 International Conference on Key Competencies in the Knowledge Society (KCKS) / Held as Part of World Computer Congress (WCC), Sep 2010, Brisbane, Australia. pp.155-164, ff10.1007/978-3-642-15378-5_15ff. ffhal01054703.

[5] Vainio, V., Sajaniemi, J. (2007). Factors in Novice Programmers' Poor Tracing Skills. ACM SIGCSE Bulletin. Volume 39, Issue 3

[6] Rum, Siti Nurulain Mohd and Maslina Zolkepli. (2018). Metacognitive Strategies in Teaching and Learning Computer Programming. International Journal of Engineering & Technology, 7 (4.38) (2018) 788-794.

[7] ÇAKIROĞLU, Ünal and Betul ER. (2020). Effect of Using Metacognitive Strategies to Enhance Programming Performances. Informatics in Education, 2020, Vol. 19, No. 2, 181–200

[8] Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American Psychologist, 34(10), 906–911.

[9] Van der Stel, Manita. (2011). Development of Metacognitive Skills in Young Adolescents: A Bumpy Ride to the High Road. Disertasi, Universiteit Leiden, Oktober 2011.

[10] Cantoia, M., B. Colombo, A. Gaggioli, B. Girani De Marco. (2012). Metacognition 2012. Proceedings of the 5th Biennial Meeting of the EARLI Special Interest Group 16 Metacognition, September 5-8, 2012.

[11] Kornell, Nate, Lisa K. Son, and Herbert S. Terrace. (2007). Transfer of Metacognitive Skills and Hint Seeking in Monkeys. Psychological Science, Volume 18, Number 1, 2007.

[12] Dawson, Theo L. (2008). Metacognition and learning in adulthood Developmental. Testing Service, LLC. Northampton, MA.

[13] Veenman, Marcel V.J., Pascal Wilhelm, Jos J. Beishuizen (2004) The relation between intellectual and metacognitive skills from a developmental perspective, Learning and Instruction 14 (2004) 89–109.

[14] Kuhn, Deanna (2000). Metacognitive Development. Current Directions in Psychological Science, Vol. 9, No. 5 (Oct., 2000), pp. 178-181

[15] Schraw, Gregory & Rayne Sperling Dennison. (1994). Assessing Metacognitive Awareness. Contemporary Educational Psychology, Volume 19, 1994, 474-475.

[16] Hambleton, Ronald K; Swaminathan, H; and Jane Rogers, H. 1991. Fundamentals of Item Response Theory. London: SagePublications.

[17] Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Re-trieved from http://www.psychometrika.org/journal/online/MN17.pdf.

[18] Widhiarso, Wahyu. 2010. "Model Respons Bergradasi Graded Response Model (GRM)". http://widhiarso.staff.ugm.ac.id/files/Model%20Respons%20Bergradasi%20GRM.pdf .

[19] Tessmer, Martin. 1993. Planning and Conducting Formative Evaluations: Improving the Quality of Education and Training. Abingdon: Routledge.

[20] Matteucci, Mariagiulia. (2006). Student assessment via graded response model. STATIS-TICA, anno LXVI, n. 4, 2006.

[21] Cohen, Allan S. , Seock-Ho Kim, and Frank B. Baker. (1993). Detection of Differential Item Functioning in the Graded Response Model. Applied Psychological Measurement. Volume 17 Number 4 December 1993.

[22] David Thissen, Li Cai, R. Darrell Bock. 06 Apr 2010, The Nominal Categories Item Response Model from: Handbook of Polytomous Item Response Theory Models Routledge Accessed on: 11 Sep 2022 https://www.routledgehandbooks.com/doi/10.4324/9780203861264.ch3

[23] Lautenschlager, G. J., Meade, A. W., & Kim, S.-H. (2006, April). Cautions Regarding Sample Characteristics When Using the Graded Response Model. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dal-las, TX.