International Journal of Evaluation and Research in Education (IJERE)
Vol. 9, No. 4, December 2020, pp. 955~962
ISSN: 2252-8822, DOI: 10.11591/ijere.v9i4.20505

Development of an assessment tool for mathematical reading, analytical thinking and mathematical writing

Nuttida Kesorn[1], Putcharee Junpeng[2], Metta Marwiang[3], Kissadapan Pongboriboon[4], Keow Ngang Tang[5], Shruti Bathia[6], Mark Wilson[7]
[1,2,3,4]Faculty of Education, Khon Kaen University, Thailand
[5]Institute for Research and Development in Teaching Profession for ASEAN, Khon Kaen University, Thailand
[6,7]Graduate School of Education, University of California, United States

Article Info

Article history:

Received Jan 30, 2020
Revised Jul 25, 2020
Accepted Sep 21, 2020

Keywords:

Analytical thinking
Assessment tool
Mathematical reading
Mathematical writing
Multidimensional model

ABSTRACT

The main objective of this research was to develop and validate the quality of an assessment tool for evaluating the mathematical reading, analytical thinking, and mathematical writing skills of fourth-grade students. We randomly selected 222 fourth grade students across multiple schools of varying sizes to take the assessment. Multidimensional Random Coefficients Multinomial Item Response Model was applied to validate the quality of the developed assessment tool. A design-based research methodology was adopted to develop the assessment tool encompassing four phases as follows: 1) analyze how students solve mathematical problems; 2) develop the assessment tool; 3) validation of the tool; and 4) reflection. The results of this research indicate that the assessment tool consisting of 19 items and two dimensions is a reliable and valid metric to measure mathematical reading, mathematical writing and analytical ability of fourth graders. The Likelihood-Ratio test showed that the multidimensional model fits better in comparison to the unidimensional model. It can be concluded that each item is qualified to assess the students and relevant to the developed dimensional examination structure.

Corresponding Author:

Putcharee Junpeng,
Faculty of Education,
Khon Kaen University,
123 Mitraphap Road, A. Muang, Khon Kaen 40002, Thailand.
Email: jputcha@kku.ac.th

1. INTRODUCTION

Mathematics is an important subject because it provides practical knowledge and plays a significant role in stimulating student’s learning [1, 2]. This can be explained by the fact that mathematics is not only a fundamental discipline but also a foundation for many other scientific disciplines [3, 4]. If mathematics teachers are to be judged by the outcomes of the students, then, at least the components making up the curriculum and the assessment tasks should be made explicit, so that the classroom activities may be aligned and reasoned judgements may be made regarding the classroom focus, so that the classroom activities may be aligned and reasoned judgments may be made regarding the classroom focus of the teachers mathematics teachers concerning their classroom focus [5, 6]. Therefore, some degree of regulation is deemed necessary in both curriculum document prescription and systematic assessment in the current global educational climate [7, 8].

According to [9, 10], students are required to analyze the situation and use complex knowledge including mathematical understanding and analytical thinking in the process of solving mathematical problems. The natural question during the assessment process is – how to also assess the process of thinking demonstrated in the solutions and not only the mathematical correctness answers [9, 11]. Analytical thinking assists students in solving problems in mathematics. Students need to understand parts of the situation, the ability to scrutinize and breakdown facts.

Apart from the above, we also need to keep in mind that there are different ways of arriving at a solution. The National Test in Thailand classified mathematical literacy into four levels – pre-analytics, partial-analytical, semi-analytical and analytical [11]. Mathematical reading is important because students have to read to work through mathematical problems, communicate their ideas coherently, organize their thoughts, structure arguments, extend their thinking and knowledge to cover other perspectives and experiences, understand their own problem-solving and thinking process as well as of others and finally develop flexibility in representing and interpreting ideas [12]. Mathematical writing is another essential ability that students need, in order to write clear mathematical explanations. If students want to contribute to the greater body of mathematical knowledge, they must be able to communicate their ideas in a way that is comprehensible to others [13].

According to [14], mathematics ability enables students to comprehend mathematics concept, to explain the correlation of concepts and to apply concept of algorithm flexibly, accurately, efficiently, and precisely in problem-solving. Currently, there is not an appropriate assessment tool to assess students' reading, analyzing, and writing skills in mathematics accordance to the record from the Office for National Education Standards and Quality Assessment [15]. If we have a high-quality tool, we can use the results from the assessment to improve teacher performance and also enable students to enhance their mathematical reading, mathematical writing and analytical skills. In this line of reasoning, current research is aimed to develop and validate the quality of an assessment tool for evaluating the mathematical reading, analytical thinking, and mathematical writing (RTW) of fourth-grade students.

The mathematical reading dimension consists of interpreting the problems and capturing the points. The analytical thinking, which is common to both the dimensions includes problem-solving and rational thinking indicators. Mathematical writing covers report this first and then report analytical thinking dimension. In the first pilot study, we found that mathematical thinking is a multi dimensional construct that comprises of two sub dimensions – mathematical reading and mathematical writing. We justify the above argument by providing evidence for validity, reliability and item fit. As the results, this research was focused on reading and analytical thinking (RAT), and writing and analytical thinking (WAT) dimensions. We used the between-item multidimensional model. This means that each item mapped only to one dimension as shown in Figure 1.

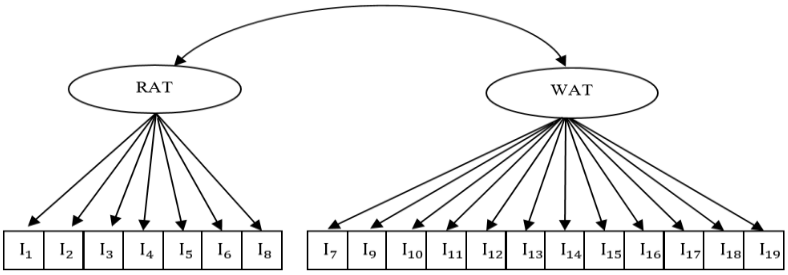


Figure 1. Between-item multidimensional model for assessing RTW

2. RESEARCH METHOD

Researchers adopted the multidimensional modeling approach [16-19] and design-based research [20, 21] method to develop the assessment tool. We used the Multidimensional Random Coefficients Multinomial Logit Model (MRCML) to validate the quality of the developed assessment tool.

2.1. Multidimensional random coefficients multinomial logit model (MRCML)

We use Item Response Theory methods to analyze the items. IRT methods have advantages over classical test theory approaches [17, 18]. Within the family of IRT models we use the Multidimensional Random Coefficients Multinomial Logit Item Response Model (MRCML) [19] because it retains the estimate for each item while modeling dependencies between them.

2.2. Population and sample

A total of 222 fourth grade primary school students of varying abilities studying in schools of varying sizes (small, medium, and big) in the Nakhon Ratchasima Province, Thailand were randomly selected. This is the minimum sample size required for using multidimensional item response model in order to get quality information.

2.3. Research procedure

The design-based research procedure consisted of four phases as follows: 1) analyze how students solve mathematical problems; 2) develop the assessment tool; 3) validation of the tool; and 4) reflection. In the first phase, we collaborate with mathematics teachers to develop the conceptual model of mathematical reading, mathematical writing and analytical ability in line with the core curriculum in basic education 2008 (revised edition 2017). We collect data through interviews and using think-aloud techniques. In the second phase, we develop a prototype guided by the test blueprint to assess student's RAT and WAT. A total of 19 items were developed to measure student's ability in the two dimensions.

In the third phase, researchers validated the quality of the developed assessment tool by considering its validity and reliability. There were three validity evidence that researchers took into consideration, namely 1) test content by experts and the Wright Map; 2) students' response processes as the characteristics of students' thinking reflected in the Think-aloud's Form; and 3) internal structure through the Wright Map using ACER Conquest 2.0 [22, 23]. Moreover, the reliability of assessment tool that researchers took into consideration were: 1) reliability of the Expected-A-Posteriori and Separation (EAP/PV) which is a measurement of the consistency of multidimensional analysis; 2) internal consistency using Cronbach's Alpha Coefficient; and 3) Standard error of measurement (SEM) in line with the educational and psychological assessment standards [24]. Finally, we reflect on the assessment and propose changes if necessary.

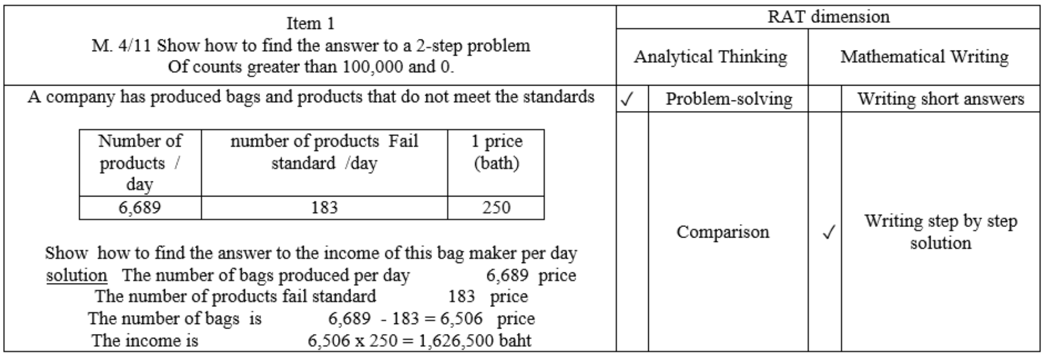
3. RESULTS AND DISCUSSION

3.1. Development of assessment tool

Researchers developed an assessment tool consisting of 19 items to evaluate fourth-grade students' Mathematical capabilities in terms of their RTW. This assessment tool consists of seven items in RAT and 12 items in WAT. The item format is the 4-choices question and also essay questions. All the items were analyzed using MRCML. The assessment tool consists of two dimensions and six mathematical indicators as shown in Table 1. We present in Figure 2 an example item for fourth graders in the eleventh indicator (M.4/11). This item maps onto the RAT dimension with focus on problem-solving, analytical thinking and the ability to write step by step processes.

Table 1. Assessment tool of RTW

Mathematical Indicators	RAT (Item)	WAT (Item)	No. of Items
M. 4/2 Compare and sort out more than 100,000 numbers from different situations.	(1)	(7)	2
M. 4/7 Estimate the result of addition, subtraction, multiplication, and division from reasonable situations.	(2)	(10), (13), (16), (18)	5
M. 4/8 Find the value of an unknown character in a symbolic sentence. Show addition and subtraction by writing a symbolic sentence of a number greater than 100,000 and 0.	(3)	(14), (17)	3
M. 4/10 Find the result of addition, subtraction, multiplication, and division between the numbers and 0.	(4),(5), (6)	(9), (12)	5
M. 4/11 Show how to find the answer to a 2-step problem of counts greater than 100,000 and 0	(8)	(11), (15), (19)	4
Item Total	7	12	19



Item 1 M. 4/11 Show how to find the answer to a 2-step problem Of counts greater than 100,000 and 0.	RAT dimension			
	Analytical Thinking		Mathematical Writing	
A company has produced bags and products that do not meet the standards	✓	Problem-solving		Writing short answers
		Comparison	✓	Writing step by step solution

Number of products / day	number of products Fail standard /day	1 price (bath)
6,689	183	250

Show how to find the answer to the income of this bag maker per day
solution The number of bags produced per day 6,689 price
The number of products fail standard 183 price
The number of bags is 6,689 - 183 = 6,506 price
The income is 6,506 x 250 = 1,626,500 baht

Figure 2. Examples of an item to assess RAT dimension in RTW

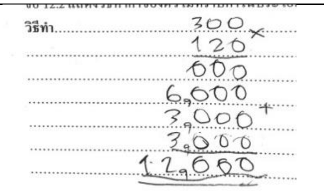
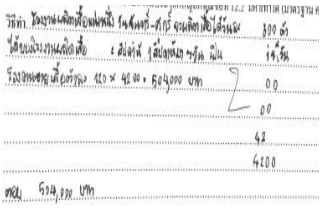
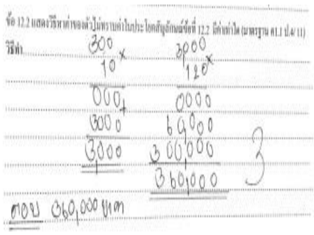
3.2. The validity of assessment tool

A total of four experts validated the content of the assessment tool [9]. Content validity index (CVI) is the degree to which an instrument has an appropriate number of items for construct being measured indicating an excellent level of content validity [25]. The CVI has 2 indices--the individual content validity index (I-CVI) and the overall content validity index (S-CVI). The I-CVI which is the proportion of content experts giving item a relevance rating of 3 or 4 was 1.00. Moreover, the S-CVI as the content validity of the overall scale was equal to 1.00 as well. This implies that the assessment tool is found to be valid in terms of its content [25].

The next step of validation was based on students' response processes reflecting in the think-aloud forms. Students from three different ability levels, namely, good, moderate and weak took part in the think-aloud to explain their learning behavior. Researchers analyzed the qualitative data based on their responses. This validation method is known as the "think-aloud" protocol. The researchers synthesized the results and used them to improve the items to ensure that students can understand completely the content of each item in the assessment tool. Table 2 shows an example think-aloud procedure.

Table 2. Examples of think-alound protocol results

Open-ended Question of Item 2
One day shirt factory produces 300 shirts for Monday-Friday. The factory is closed every Saturday and Sunday. The factory sells shirts at 120 baht per shirt. What this is the factory income of -2week shirt?

Level	Person	Example conversation	Interpretation of protocol	Picture of student answers
Weak	1	Teacher: From the problem to the situation. How to find an answer? Students: 300x 120is the answer.	Students were able to read the problem and solve the problem but incorrectly. Because they cannot interpret and cannot find symbol sentences to find the correct answer	 วิธีทำ 300 × 120 000 6,000 + 3,000 3,000 12,000
	2	Teacher: What is the question given and what is the question? Students: the factory produces 300 shirts a day for Monday-Friday .The factory sells shirts at 120 baht per shirt.		
Medium	1	Teacher: From the problem to the situation. How to find an answer? Students: 300x120x14 is the answer.	In this group, it was found that students were able to read the problem and solve the problem but cannot interpret and cannot find symbol sentences to find the correct answer	 120 × 4200 = 504000 300 14 00 00 42 4200 504,000
	2	Teacher: What is the question given and what is the question? Students: the factory produces 300 shirts a day for Monday-Friday. The factory sells shirts at 120 baht per shirt. You like to know what the income of the -2week shirt factory is.		
Good	1	Teacher: From the problem to the situation. How to find an answer? Students: 300x 120x10 is the answer.	In this group, students can read the problem correctly and find the symbolic sentence to find the correct answer.	 300 × 10 000 3000 3000 3000 × 120 0000 6000 300000 360000 3 360,000
	2	Teacher: What is the question given and what is the question? Students: the factory produces 300 shirts a day for Monday-Friday .The factory is closed every Saturday and Sunday. The factory sells shirts at 120 baht per shirt. You like to know what the income of the -2week shirt factory is		

Results of validation based on the internal structure of assessment tool revealed that the multidimensional approach has produced a better AIC and BIC values to evaluate RTW when compared to the unidimensional approach, as shown in Table 3. The comparative analysis of the two models, unidimensional approach versus multidimensional approach showed that the deviance statistic was 5984.93 and 5881.76, the number of parameters was 42 and 44, AIC values were 6068.93 and 5969.76, and BIC values were 6083.47 and 5985.00 respectively. It can be concluded that the multidimensional approach is found to be the most relevant model [26]. Additionally, the results of the covariance/correlation matrix of RAT and WAT showed that there is a correlation between the two dimensions that are RAT and WAT as 0.68. This implies that the correlation between the two dimensions is medium.

Table 3. Results of validation based on internal structure

Assessment tool	Device statistics	No of parameter	AIC	BIC
Unidimensional approach	5984.93	42	6068.93	6083.47
Multidimensional approach	5881.76	44	5969.76	5985.00

Likelihood Ratio Chi-Squared $G^2 = \chi^2$=103.17, df=2, p = .01; AIC = 5969.76 < 6068.93; BIC = 5985.00 < 6083.47

The Wright map was used to provide a picture of the assessment tool by placing the difficulty of the items on the same measurement scale as the capability of the students. This provides the researchers with a comparison of students and items, to better understand the appropriateness of the assessment tool [27]. Researchers observe that the mean location increase and banding of thresholds of the Wright Map support construct validity. However, we notice some overlap between level 2 and level 3 of items in the WAT dimension. We expected to observe a monotonic increase in mean WLE as levels increase within each item [28]. We notice that the mean WLE is increasing for each item. The respondents are distributed normally between a range of around -5 logits to +4 logits. Wright map of the assessment tool is shown in Figure 3.

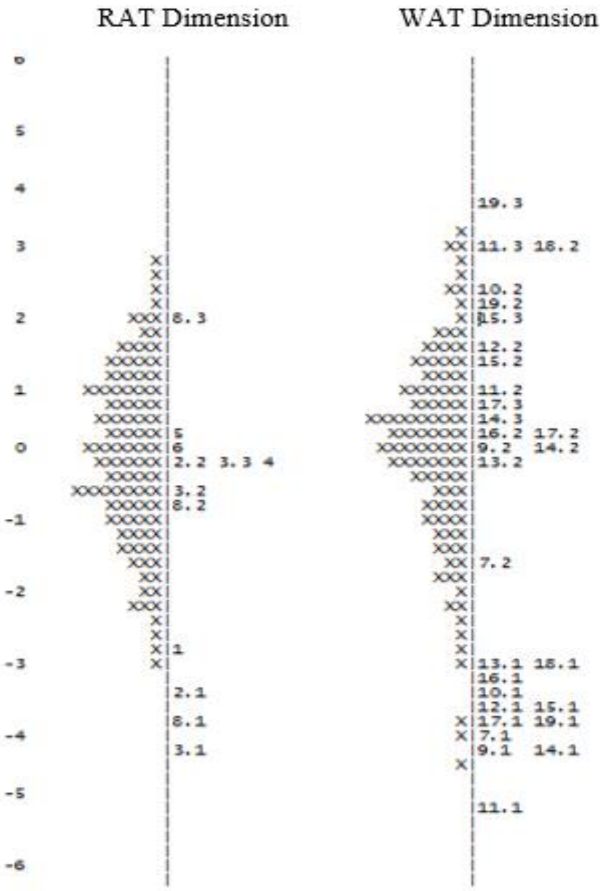


Figure 3. Wright map of the assessment tool

3.3. Reliability of the assessment tool

Researchers analyze the reliability coefficient using Rasch analysis by identifying Expected-A-Posteriori and Seperation (EAP/PV). Results of EAP/PV were highest when we used a multidimensional model for analysis. The EAP/PV values of RAT and WAT were 0.77 and 0.84 respectively. This implies that the two dimensions are considered as suitable precision to use as a research tool which is consistent with the criteria set by [29] who suggested that the precision of the measuring coefficient should be greater than 0.70. The accuracy of the reliability coefficient is considered acceptable because the assessment tool is not a measurement that has a large impact on the sample [29].

Next, researchers analyzed internal consistency using True Score Model as Classical Test Theory (CTT) by identifying the Cronbach's Alpha Coefficient (α). Similar results were found as the reliability values as 0.87 for assessment tool of RTW. This is once again consistent with the criteria set by [29] that the precision of the measuring coefficient should be greater than 0.70. The accuracy of the reliability coefficient is considered acceptable because the assessment tool is not a measurement that has a large impact on the sample [29, 30]. Finally, researchers utilized the standard deviation graph SEM to investigate the reliability of the assessment tool by examining the standard error of measurement. When the multidimensional model was separated into two related sub-dimensions, namely θ_{RAT}, and θ_{WAT}, the latent parameter of each student would have a different standard error of measurement (SEM). Figure 4 illustrates the SEM for the two separated sub-dimensions.

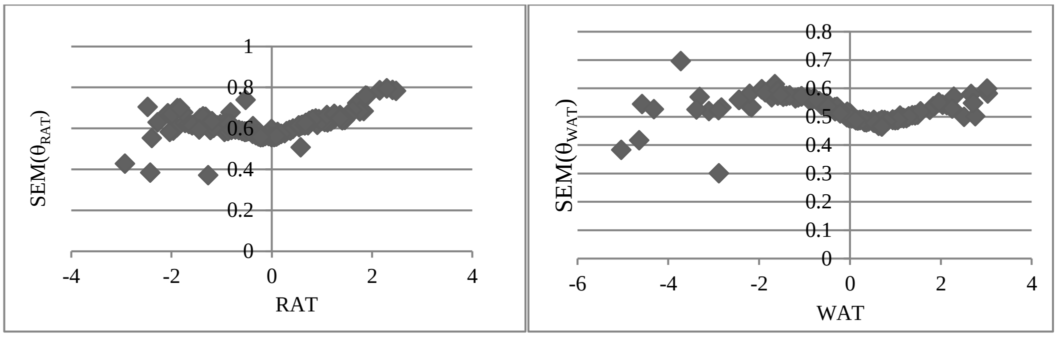


Figure 4. The standard error of measurement for RAT and WAT

The reliability evidence of RAT and WAT's standard error of measurement (SEM θ) showed that SEM (θ_{RAT}) and SEM (θ_{WAT}) are ranged from 0.37 to 0.80 and 0.30 to 0.70, respectively. This implies that the SEM values for both dimensions were at acceptable level and small error for estimating RTW, particularly for intermediate to the high level of RTW. This is because both SEM values had the lowest error if the student ability (θ) were ranged from -0.50 to +0.50 logits. However, the errors seemed to increase when estimating the high level of RAT and the low level of WAT. Results of the overall SEM values from the multidimensional model revealed that students have the same mean score for both RAT and WAT dimensions. The latent dimension values on both sides of the relationship graph between SEM values were flat as obtained from the multidimensional model. Therefore, this is the evidence of the assessment tool's reliability that consistent to the sub-dimensional estimate of each side.

3.4. Quality of the assessment tool

The quality of the assessment tool was examined using the item fit based on the MRCML through Conquest 2.0 [16]. The essential result of this research is an assessment tool developed to evaluate the mathematical reading, analytical thinking, and mathematical writing of fourth-grade students. This assessment tool is found to have high precision, stability and consistency to assess RTW in each dimension. As a result, the developed assessment tool can overcome the 21 st century classroom whereby mathematics teachers have to achieve some congruence between tests used for monitoring or summative purposes, for the active classroom and classroom-based assessment [6]. In this line of reasoning, researchers propose that mathematics teachers can use this assessment tool to develop and improve students' capabilities in terms of their RTW in their teaching and learning process. Besides, researchers would like to suggest that mathematics teachers are required to do their alignment for these modes of assessment and also critically engaged in their professional development to learn how to develop a quality assessment tool [1, 2]. Given the importance of alignment assessment practices with classroom practices, mathematics teachers must have a reference that is explicit, and in some respect common to their settings as indicated in the results of this research [5]. Table 4 shows the result of item fit statistic analysis for multidimensional model.

Table 4. Results of item fit statistic analysis for multidimensional model

Item	Estimate	Error	Unweight Fit (OUTFIT)			Weight Fit (INFIT)		
			MNSQ	CI	T	MNSQ	CI	T
1	-2.86	0.25	0.73	(0.81, 1.19)	-3.10	1.00	(0.67, 1.33)	0.10
2	-1.77	0.14	0.96	(0.81, 1.19)	-0.40	0.93	(0.83, 1.17)	0.80
3	-1.60	0.10	1.22	(0.81, 1.19)	2.20	1.10	(0.82, 1.18)	1.10
4	-0.00	0.16	0.94	(0.81, 1.19)	-0.60	0.93	(0.87, 1.13)	-1.00
5	0.21	0.16	0.96	(0.81, 1.19)	-0.40	0.95	(0.87, 1.13)	-0.80
6	0.12	0.16	1.18	(0.81, 1.19)	1.80	1.07	(0.87, 1.13)	1.10
7	-2.82	0.17	1.49	(0.81, 1.19)	4.50	1.31	(0.75, 1.25)	2.30
8	-0.89	0.12	1.06	(0.81, 1.19)	0.60	1.06	(0.82, 1.18)	0.70
9	-2.08	0.15	1.25	(0.81, 1.19)	2.50	1.12	(0.85, 1.15)	1.50
10	-0.50	0.17	0.81	(0.81, 1.19)	-2.10	0.85	(0.77, 1.23)	-1.30
11	0.35	0.13	0.96	(0.81, 1.19)	-0.40	0.88	(0.80, 1.20)	-1.20
12	-1.06	0.16	1.22	(0.81, 1.19)	2.20	1.18	(0.82, 1.18)	1.80
13	-1.53	0.14	0.73	(0.81, 1.19)	-3.10	0.80	(0.82, 1.18)	-2.40
14	-1.18	0.10	0.94	(0.81, 1.19)	-0.70	0.97	(0.83, 1.17)	-0.40
15	-0.10	0.12	0.95	(0.81, 1.19)	-0.50	1.01	(0.78, 1.22)	0.10
16	-1.48	0.14	0.90	(0.81, 1.19)	-1.10	0.92	(0.84, 1.16)	-1.00
17	-1.00	0.10	1.48	(0.81, 1.19)	4.50	1.31	(0.83, 1.17)	3.20
18	-0.09	0.18	0.89	(0.81, 1.19)	-1.20	0.95	(0.75, 1.25)	-0.30
19	0.64	0.16	1.01	(0.81, 1.19)	0.20	1.05	(0.73, 1.27)	0.40

4. CONCLUSION

The results indicate the importance of mathematics teachers to relate their thought to its diagnostic relevance in the classroom while they are constructing an assessment. Therefore, mathematics teachers have to consider an appropriate balance and coverage of the curriculum and attempts to cover different types of cognitive engagement while dealing with the content validity of the assessment tool. Further mathematical insight is required to populate such a multidimensional model with appropriate items. Moreover, the validity evidence suggested that the assessment tool is found appropriate for a student in the intermediate to the low level more than high level in RAT and in the intermediate to the high level more than low level in WAT. This is because the lowest RTW level of students showed the highest error of SEM value.

However, there are still some limitations in this research because researchers used only three validity quality methods to measure the assessment tool. Therefore, future researchers can consider other criteria to determine the coefficient between conditional accuracy and predictive validity by looking into the relationship between the constructed test and students' standardized examination. Multidimensional Item Response Theory that utilized in this research needs to span the range of item difficulties for accurate estimation of the item parameters. Therefore, future researchers have to use non-random sample so that the estimated parameter could later cover the whole range of diverse capabilities levels from the lowest level (logit ≤ =3) to the highest level (logit ≥ +3).

MRCML is a general and flexible model that has been used by researchers to design matrices to specify the relationship between responses to the items and structural parameters for the given measurement situation that allows for the specification of a large number of multidimensional item response models. Consequently, researchers would like to suggest to the Ministry of Education, Thailand to conduct the related training for mathematics teachers so that they know how to utilize the MRCML model whenever they involve in assessing their students' mathematical learning problems. Ultimately, the assessment tool will assist them to assess their students' multidimensional mathematical proficiencies and improve their overall mathematical proficiencies as a total.

ACKNOWLEDGEMENTS

Researchers would like to take this opportunity to thank the Thailand Research Fund (TRF) Advanced Research Scholar, and Khon Kaen University, Thailand (Grant No: RSA6080074), Faculty of Education, Khon Kaen University to make this research a success. Special thanks to Graduate School of Khon Kaen University, Thailand for providing financial support to attend at the 3rd World Conference on Education 2019 (WCEDU 2019) at Kualar Lumpur, Malaysia.

REFERENCES

[1] P. Junpeng, *et al.*, “Constructing progress maps of digital technology for diagnosing mathematical proficiency,” *Journal of Education and Learning,* vol. 8, no. 6, pp. 90-102, 2019.

[2] J. Adler, “Learning about Mathematics Teaching and Learning from Studying Rituals and Ritualization? A Commentary,” *Educational Studies in Mathematics*, vol. 101, no. 2, pp. 291-299, 2019.

[3] H. Lattimer, “Translating Theory into Practice: Making Meaning of Learner Centered Education Frameworks for Classroom-based Practitioners,” *International Journal of Educational Development*, vol. 45, pp. 65-76, Nov. 2015.

[4] Y-M. Huang, S-H. Huang and T-T. Wu, “Embedding Diagnostics in a Digital Game for Learning Mathematics,” *Education Technology Research and Development*, vol. 62, no. 2, pp. 187-207, 2014.

[5] Z, Aksu, “Pre-service mathematics teachers’ pedagogical content knowledge regarding student mistakes on the subject of circle,” *International Journal of Evaluation and Research in Education*, vol. 8, no. 3, pp. 440-445, 2019.

[6] C. Long, T. Dunne and H. de Kock, “Mathematics, Curriculum and Assessment: The Role of Taxonomies in the Quest for Coherence,” *Pythagoras*, vol. 35, no. 2, pp. 1-14, 2014.

[7] G.M. Richardson, L.L. Byrne and L.L. Liang, “Making Learning Visible: Developing Preservice Teachers’ Pedagogical Content Knowledge and Teaching Efficacy Beliefs in Environmental Education,” *Applied Environmental Education & Communication*, vol. 17, no. 1, pp. 41-56, 2018.

[8] W. Kuiper, N. Nieveen and J. Berkvens, “Curriculum Regulation and Freedom in the Netherlands – A Puzzling Paradox,” in W. Kuiper and J. Berkvens (Eds.), *Balancing Curriculum Freedom and Regulation across Europe. CIDREE Yearbook 2013*, pp. 139-162, 2013.

[9] K. Bulková and S. Čeretková, “Rubrics as Assessment Tool of Mathematical Open-ended Problems,” in *16th Conference on Applied Mathematics APLIMAT 2017 Proceedings*, pp. 235-244, 2017.

[10] R.Y. Gazali, “The Development of Learning Materials for Mathematics for Junior High School Students based on Ausubel’s Theory of Learning,” *Pythagoras: Mathematics Education Journal*, vol. 11, no. 2, pp. 182-192, 2016.

[11] A. Qolfathiriyus, I. Sujadi and D. Indriati, “Characteristic Profile of Analytical Thinking in Mathematics Problem Solving,” *Journal of Physics: Conference Series,* vol. 1157, no. 3, pp. 1-6, 2019.

[12] D. Metsisto, “Reading in the Mathematics Classroom,” in J. M. Kenney, *et al.* (Eds), *Literacy Strategies for Improving Mathematics Instruction, Chapter 2 (The Association for Supervision and Curriculum Development ASCD),* 2005.

[13] Kevin P. Lee, “A Guide to Writing Mathematics,” pp. 1-17, 2010. [Online]. Available: https://web.cs.ucdavis.edu/~amenta/w10/writingman.pdf

[14] Mutmainah, Rukayah and M. Indriayu, “Effectiveness of Experiential Learning-based Teaching Material in Mathematics,” *International Journal of Evaluation and Research in Education*, vol. 8, no. 1, pp. 57-63, 2019.

[15] Office of Academic Affairs and Education Standards, “Guidelines for the Development and Evaluation of Reading, Thinking, and Writing according to the Core Curriculum of Basic Education,” 2011.

[16] R. J. Adams, M. Wilson and W. Wang, “The Multidimensional Random Coefficients Multinomial Logit Model,” *Applied Psychological Measurement*, vol. 21, no. 1, p. 1-23, 1997.

[17] D. C. Briggs and M. Wilson, “An Introduction to multidimensional measurement using Rasch models,” *Journal of Applied Measurement*, vol. 4, no. 1, pp. 87-100, 2003.

[18] D. J. Cooke, *et al.*, “Evaluating the Screening Version of the Hare Psychopathy Checklist—Revised (PCL:SV): An item response theory analysis,” *Psychological Assessment*, vol. 11, no. 1, pp. 3–13, 1999.

[19] S. E. Embretson and S. P. Reise, *Item response theory for psychologists*. Lawrence Erlbaum Associates, Inc., Mahwah, 2000.

[20] T. Amiel and T. C. Reeves, “Design-based Research and Educational Technology: Rethinking Technology and the Research Agenda,” *Educational Technology & Society*, vol. 11, no. 4, pp. 29-40, 2008.

[21] T. Stemberger and M. Cencic, “Design Based Research: The Way of Developing and Implementing,” *World Journal on Educational Technology: Current Issues*, vol. 8, no. 3, pp. 180-189, 2016.

[22] M. Custer, “Sample size and item parameter estimation precision when utilizing the one-parameter ‘Rasch’ model,” Paper presented at *The annual meeting of the mid-western Educational Research Association, Evanston*, 2015. [Online]. Available: https://files.eric.ed.gov/fulltext/ED562560.pdf

[23] M. L. Wu, R. J. Adams, M. R. Wilson and S. A. Haldane, *ACER ConQuest Version 2: Generalized Item responses modeling software*. Camberwell: Australian Council for Educational Research, 2007.

[24] American Educational Research Association, *Standards for Educational and Psychological Testing* (6th ed.). Washington DC, 2014.

[25] M. R. Lynn, “Determination and quantification of content validity,” *Nursing Research*, vol 35, no. 6, pp. 382-385, 1986.

[26] R. P. McDonald, “A Basis of Multidimensional Item Response Theory,” *Applied Psychological Measurement*, vol. 24, no. 2, pp. 99-114, 2000.

[27] M. E. Lunz, “Using The Very Useful Wright Map,” *Measure Research Associates Test Insight*, 2010. [Online]. Available: https://www.rasch.org/mra/mra-01-10.htm

[28] B. Duckor, K. Draney, and Mark Wilson, “Assessing assessment literacy: An item response modeling approach for teacher educators,” *Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana*, vol 54, no. 2, pp. 1-25, 2017.

[29] J. C. Nunnally, *Psychometric theory* (2nd ed.). New York, McGraw-Hill, 1978.

[30] F. B. Baker and S-H Kim, *The Basics of Item Response Theory Using R.* Switzerland: Springer International, 2017. [Online]. Available: https://doi.org/10.1007/978-3-319-54205-8