International Journal of Evaluation and Research in Education (IJERE)
Vol. 9, No. 3, September 2020, pp. 665~674
ISSN: 2252-8822, DOI: 10.11591/ijere.v9i3.20503

Validation of a digital tool for diagnosing mathematical proficiency

Putcharee Junpeng[1], Metta Marwiang[2], Samruan Chiajunthuk[3], Prapawadee Suwannatrai[4], Kanokporn Chanayota[5], Kissadapan Pongboriboon[6], Keow Ngang Tang[7], Mark Wilson[8]
[1,2,3,4,5,6]Department of Educational Measurement and Evaluation, Khon Kaen University, Thailand
[7]International College, Khon Kaen University, Thailand
[8]Graduate School of Education, University of California, United States

Article Info

Article history:

Received Jan 29, 2020
Revised May 19, 2020
Accepted Jul 15, 2020

Keywords:

Construct modeling
Digital tool
Mathematical proficiency
Rasch analysis
Test development

ABSTRACT

This study was aimed to validate a digital tool for diagnosing mathematical proficiency in the Number and Algebra strand of 1,504 Thai seventh-grade students. Researchers employed a multidimensional approach, an extension of the Rasch model to measure its quality. A design-based research method was adopted to create the diagnostic tool which consists of four components, namely register system, input data, process system, and diagnostic feedback report. The diagnostic framework consists of 18 tasks encompassing 11 and 7 tasks in the Mathematical Procedures Dimension and Structure of the Observed Learning Outcome dimension, respectively. The results revealed that there is internal structure evidence of validity based on the comparison of model fit and the Wright map. The results also indicated that the reliability evidence and item fit are compliant with the quality of the digital tool as shown in the analysis of standard error of measurement and infit and outfit of the items. In conclusion, the developed digital tool can diagnose seventh-grade students' multiple mathematical proficiencies in terms of accuracy, consistency, and stability. This implies that the digital tool can provide fruitful information, particularly to those intermediate and high mathematical proficiency levels because the error for estimating proficiency in each dimension was at the lowest value for those students.

Corresponding Author:

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

1. INTRODUCTION

Mathematical proficiency (MP) refers to a student's ability to explore, conjecture, and reason logically in cognitive processes and to understand how to solve a mathematical problem that to: apply appropriate strategies to solve the problems and reflect on the procedure used to solve the problems [1-3]. Since mathematics is important and practical knowledge as a fundamental discipline and a foundation for many other scientific disciplines, MP will be among the 21st-century skills to solve the problems in our real-life efficiently and appropriately [4-7], particularly in the Number and Algebra strand [8]. This is further supported by [9] who stated that students have to be prepared with sufficient MP in terms of skills and competencies needed for work and life in the current era of accelerating change in society, particularly due to technological development [10]. In other words, MP is considered as a tool for studying science, technology, and other topics that underline the infrastructure for developing humans and economies in the era of digital transformation [11].

The MP Assessment Framework in this study has two dimensions, namely Mathematical Procedures (MAP) and Structure of the Observed Learning Outcome (SLO). The progress maps of [11]'s study showed that there are five levels of the MAP dimension, namely non-response/irrelevance, unrecalled memory, basic memory and reproduction, simple skills and concept, and strategic or extended thinking [12, 13]. On the other hand, the SLO dimension which was adopted from the SOLO taxonomy is a model to identify, describe or explain the level of understanding to determine the quality level of students' learning results. Researchers divided SLO [14] into four levels, namely pre-structural, uni-structural, multi-structural and relational, and extended abstract.

2. RESEARCH METHOD

The research employed construct modeling that embedding pedagogy and curriculum while designing the diagnostic task [15]. A design-based research method was then adopted from Adams [16] approach with four consecutive steps to create the diagnostic tool. Besides, the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM) [17] was applied to validate the quality of the diagnostic tool.

2.1. Population and sample

The population of this research is all the seventh-grade students from Thai secondary schools in the northeastern region of Thailand, so-called Isan. This region has the largest amount of students, located in a rural area, the school facilities are not well equipped, and students' proficiencies are generally lower than the national standard particularly in the Number and Algebra Strand of seventh-grade students [18]. A total of 1,504 students with diverse capabilities levels from 119 schools were stratified randomly selected following the school size in Isan-biggest, big, medium, and small sizes to fulfill the sample size required at provincial and national levels, and the school readiness [19]. The 1,504 students were involved in the third and fourth phases of the study (shown below).

2.2. Research procedure

The research procedure consisted of four phases as shown in Figure 1. In the first phase, researchers analyzed the problems of practical diagnostic tools on how to diagnose MP, namely (a) what students know and can do, as well as (b) what students need to know and can do. Researchers employed qualitative in-depth interviews, focus group interviews, and document analysis with software engineers, teachers, and educators to identify the definition of MP to create the construct map in each dimension to fit the actual context of Isan's mathematics classroom.

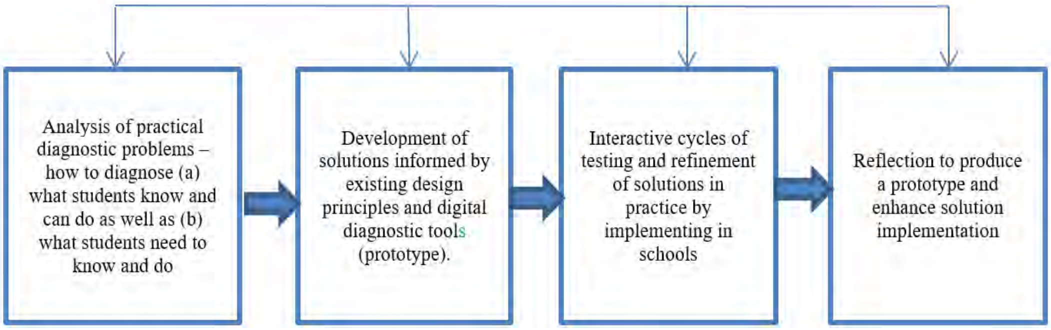
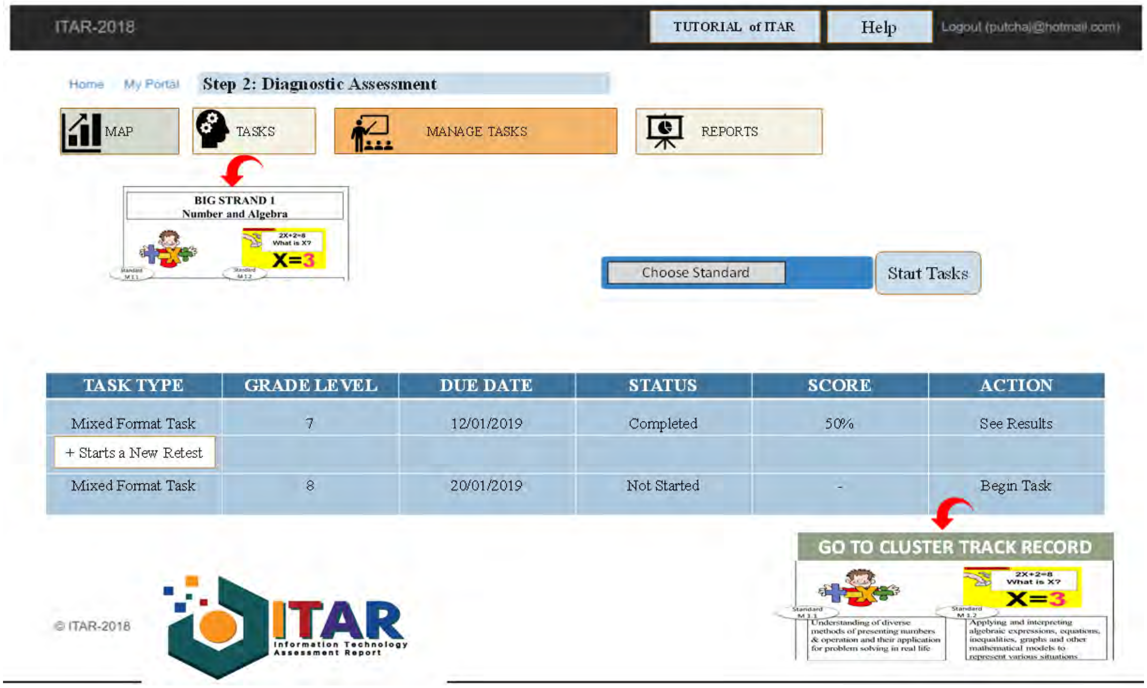

Analysis of practical diagnostic problems – how to diagnose (a) what students know and can do as well as (b) what students need to know and do
Development of solutions informed by existing design principles and digital diagnostic tools (prototype).
Interactive cycles of testing and refinement of solutions in practice by implementing in schools
Reflection to produce a prototype and enhance solution implementation

Figure 1. The four steps of digital tool development

In the second phase, researchers developed a digital diagnostic tool as a prototype, so-called Information Technology Assessment Report (ITAR). ITAR is comprised of four components, namely register system, input data, process system, and diagnostic feedback report. The register system is accessible through the website's uniform resource locator as http://itassess.com that allows electronic direct registration of securities of the users including students, teachers, parents, and guests. In the input data component, teachers can manage tasks by themselves or students can access the tasks following the assignment of teachers.

Firstly, students have to login with their own username and password. Secondly, they can select the ‘Task’ icon from the above menu and view the ‘My Portal’ panel indicates the diagnostic assessments, big strands, and learning standards that have been assigned to the students. The mini-map in the corner highlights the clusters that have been expanded in the list. In addition, the ‘My Portal’ panel displays critical information about each assigned assessment where ‘Grade Level’ shows the intended grade level of the tasks as shown in Figure 2.



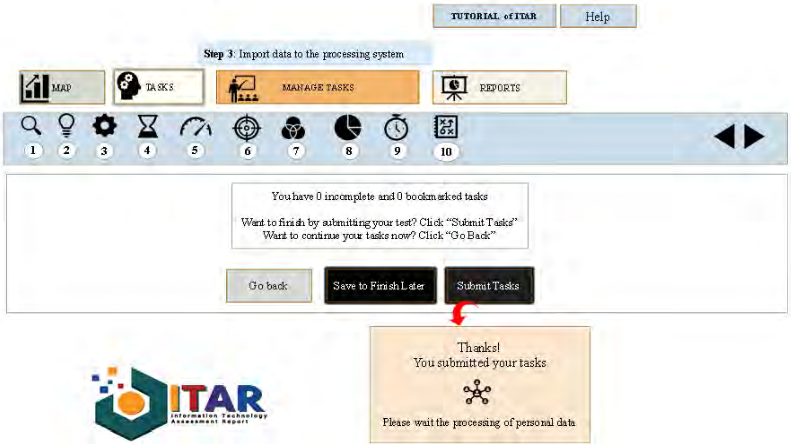
TASK TYPE	GRADE LEVEL	DUE DATE	STATUS	SCORE	ACTION
Mixed Format Task	7	12/01/2019	Completed	50%	See Results
+ Starts a New Retest					
Mixed Format Task	8	20/01/2019	Not Started	-	Begin Task

Understanding of diverse methods of presenting numbers & operation and their application for problem solving in real life

Applying and interpreting algebraic expressions, equations, inequalities, graphs and other mathematical models to represent various situations

Figure 2. Functions of input data component

The next component is the process system. After students have finished all the tasks in an assessment, they can scroll back to check through the items, review what they wrote and change their answers if necessary. Finally, students can submit their assessment if all the tasks are finalized. Figure 3 shows the submission of assessment for scoring by tapping the Pause/Submit button to pause the assessment. Once the assessment is completed, it can be submitted by clicking the submit task button. ITAR will automatically process the data once the students have submitted their assessment [20, 21].



You have 0 incomplete and 0 bookmarked tasks

Want to finish by submitting your test? Click “Submit Tasks”
Want to continue your tasks now? Click “Go Back”

Thanks!
You submitted your tasks

Please wait the processing of personal data

Figure 3. Submission of diagnostic assessment

The final component of ITAR is a diagnostic report. This component can show the individual report and the class diagnosis report. The individual report is a data-driven estimation of the student's current level of MP according to MAP and SLO dimensions. The class diagnostic report is a data-driven estimation of the whole class students' current MP performance according to the multiple dimensions. The diagnostic report provides the precise details in the real time [22] regarding the four key aspects: (i) what students learn; (ii) what students need to learn; (iii) how to learn it, and (iv) how well students understand it.

In the third phase, researchers implemented and tested the prototype with 1,504 Thai seventh-grade students. If the digital tool was inappropriate for teachers and students, researchers conducted iterative cycles of testing and refinement of solutions to find suitable ways to implement in the actual classroom context. In the final phase, researchers generated the quality evidence of the prototype by validating it through a measurement model. Researchers employed MRCMLM to examine the internal structure evidence of validity based on the comparison of model fit to ensure that the structure of the diagnostic tool in two-dimensions (i.e. MAP and SLO) fits better than one dimension. The required sample size for estimation of item parameters in the multidimensional model of the Rasch-family models is 400 to 500 to provide accurate parameter estimates [19, 23]. In addition, a Wright map was used to support the validation tool because it combined the construct for measuring MAP and SLO idea with the MRCMLM model, a powerful means to interpret the students' MP in each dimension [15]. Moreover, the low standard error of measurement (SEM (θ)) and the acceptable values of infit and outfit means would determine whether the digital tool has accuracy, consistency, and stability to diagnose in multiple proficiencies.

2.3. MP instrument

All MP tasks are created according to the core curriculum. The tasks were re-designed according to teachers and content experts' feedback as well as an initial empirical analysis of the pilot testing, using the Wright map, item fit, and step fit for validation purpose and the structural model of measurement, the internal consistency and split-half reliability coefficients for reliability information. Based on the pilot testing results, researchers deleted those tasks that overlapped the content knowledge between items and also some tasks that were found inappropriate for seventh-grade students. Subsequently, a mixed format was developed including open-ended questions and selected-response test items.

A specific scoring guide was used to assess students' MAP and SLO as the two MP dimensions. The scores of this scoring guide ranged from 0 to 4 and 0 to 3 for MAP and SLO, respectively, indicating inappropriate, partly appropriate, most appropriate, and beyond proficiency levels. The scoring guide was used for the assessment tool for the entire class.

3. RESULTS AND DISCUSSION

After researchers implemented the ITAR in the mathematics classroom for a year in the Isan region, Thailand, researchers aimed to validate the created digital tool for diagnosing mathematical proficiency in the Number and Algebra Strand of Thai seventh-grade students in terms of its accuracy, consistency, stability using validation based on internal structure, reliability, and item fit. The results of this study are presented by following the three methods of the validation analysis, namely validity evidence, reliability evidence, and item fit.

3.1. Validity evidence

After researchers tried out the created digital diagnostic tool, researchers interviewed those students regarding their understanding of the contents and the relevancy of the tasks in the digital diagnostic tool. The results revealed that students understand well about the items as expected by researchers. Besides, researchers also utilized their feedback to improve the tasks and scoring before conducting in the actual classroom context.

The second validation on the internal structure of the digital tool in terms of its accuracy of the MP construct was conducted by comparing the model fit, for the unidimensional and multidimensional models. The unidimensional model means a composition of all the tasks into one dimension while the multidimensional model means separation of the tasks into 11 tasks and 7 tasks for the respective MAP and SLO dimensions as shown in Figure 4 and Figure 5. The results revealed that multidimensional model had a statistical fit significantly better than unidimensional model through the Likelihood Ratio Chi-Squared G^2 (χ^2=589.142, df=2) [24] as well as the Akaike Information Criterion (AIC) [25] and Bayesian Information Criterion (BIC) [26] had lower value in multidimensional constructs for diagnosing MP, as shown in Table 1.

Table 1. The comparation of model fit

Model	Deviance	N of Parameter	AIC	BIC
Unidimensional	46095.39	37	46169.39	46212.95
Multidimensional	45506.25	39	45584.25	45630.16

Likelihood Ratio Chi-Squared $G^2 = \chi^2$=589.14, df=2, p = .01
AIC = 45584.25 < 46169.39
BIC = 45630.16 < 46212.95

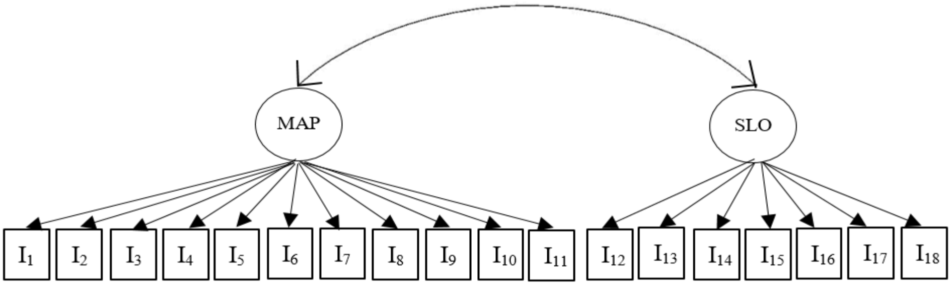


Figure 4. MAP and SLO dimensions of multidimensional model for diagnosing MP

The third validity evidence was examined using the Wright map. The validity argument of the Wright map is a graphical representation that links the item difficulties and student ability estimates on the common scale as the quality evidence. In other words, Wright map is comprised of a distribution of item difficulties, distribution of student ability estimates, and how well the item difficulty distribution is matching with the student ability estimates. Therefore, the items should match with the student ability estimates to justify that the test is maximally informative.

Results of the Wright map indicated that the distribution of item difficulties is well matching with both MAP and SLO dimensions. Both dimensions of MP show good variability item difficulties or student ability estimates. However, most of the student ability estimates were clustered in the range of hard item difficulties but did not cover low student ability for the MAP dimension. This result is expected as the tasks in the MAP dimension are suitable for the median and high MP levels but did not capture the lowest MP level. This implies that easier items should be added in the MAP dimension in the digital diagnostic tool in the future. Although the distribution of item difficulties and student ability estimates were generally appropriate with the construct to measure MP for the SLO dimension, the thresholds were sometimes lower than the empirical data when compared to other items. For example, item 16 has a threshold for score Level 2, represented by 16.2. It means that MP level at which student has a 50 percent chance of obtaining a score of 3. In the expected results, the item score threshold should be lower than the item score threshold of 15.3 and 18.3 because the chance of obtaining a score of 4 should be lower than a score of 3.

The result showed that the SLO dimension revealed some problems in the fifth and fourth levels of MAP dimension and SLO dimension, respectively. Nevertheless, the result revealed that the distribution between item score thresholds and student ability distribution is quite similar in the high level of the Wright map. In addition, the chance of obtaining a high score in SLO seemed to be harder than for the MAP. This implies that more MP levels should be added to the digital diagnostic tool in the future. This is in line with the arguments made by [11, 27] that the internal structure of assessment tools should derive from the construct maps and that skills are addressed in an orderly way at different stages in the Wright map. Generally, the skills representing the lower levels on the construct map are ones generally associated with items targeted at lower grade levels, and skills representing higher levels on the construct map are ones generally associated with items targeted at higher grade levels. In this line of reasoning, validity evidence to support the internal structure in this study is provided as the item calibrations supported the dimensions for diagnosing MP and items design.

This argument was also supported the previous validity evidence for this study based on a digital tool's content. As indicated in [28], there is a relationship between an MP assessment tool's content and the construct that is intended to measure, and items can be interpreted as the assessment tool is valid to use [15], which has been explored through the Wright map on the common scale, as shown in Figure 5. The item locations on the right cover the respondent locations on the left in the Wright map, indicating that the digital tool consisting of two dimensions, namely 11 tasks for MAP (item 1-11) and 7 tasks for SLO (item 12-18), represented the proficiency range of the students. However, the future study should add easier tasks in the MAP dimension and more level of the construct in the SLO dimension as stated in the above discussion.

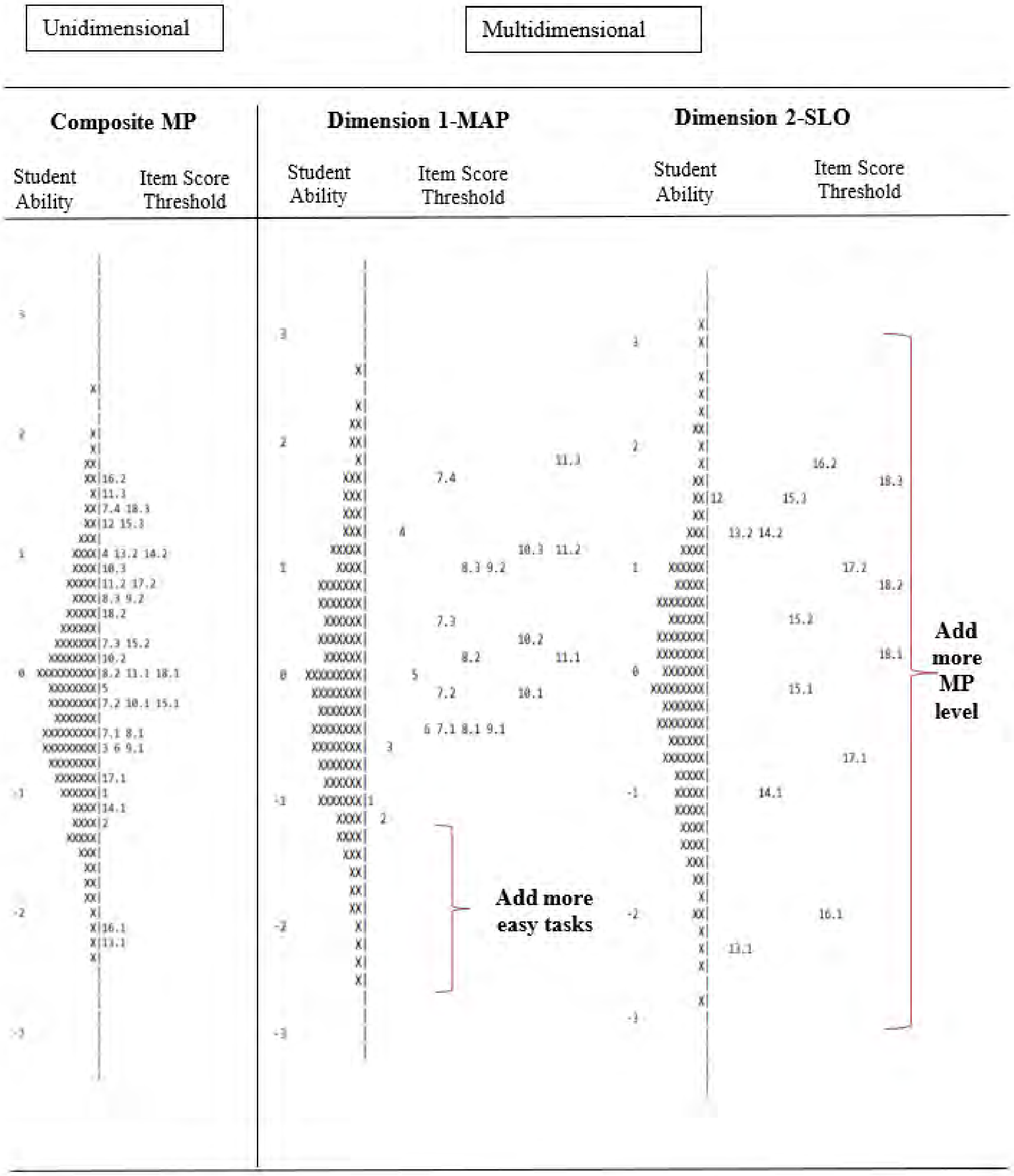
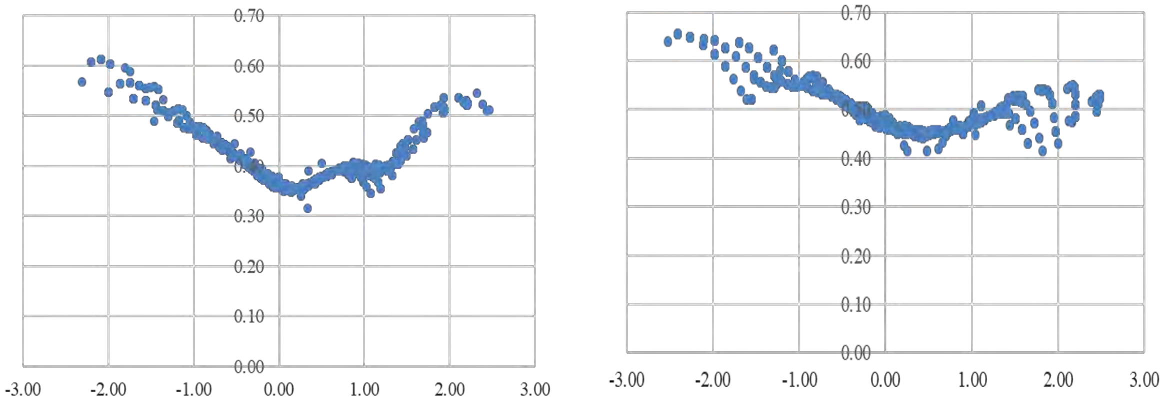


Figure 5. Wright map of unidimensional and multidimensional models for diagnosing MP

3.2. Reliability evidence

Researchers started to analyze the reliability coefficient using Item Response Theory (IRT) by identifying Expected-A-Posteriori and Separation (EAP/PV) value as shown in Figure 6. The EAP/PV reliabilities of MAP and SLO dimensions were 0.84 and 0.80 respectively, within the acceptable criteria and the internal consistency equals 0.85 was also acceptable [16].

Besides, reliability evidence of MAP and SLO's standard error of measurement (SEM θ) showed that SEM (θ_{MAP}) and SEM (θ_{SLO}) are ranged from 0.31 to 0.61 and 0.42 to 0.65 respectively. This implies that the SEM values for both dimensions were acceptable with a small error for estimating MP, particularly for intermediate to the high level of MP. This is because both SEM values had the lowest error if the student ability (θ) were in the range from 0.0 to 0.5 logits. However, the errors seemed to increase when estimating the low level of MP. The reliability evidence suggested that the digital diagnostic tool has high precision, stability, and consistency to diagnose MP in each dimension. Hence, this tool is found appropriate for a student in the intermediate to the high level of MP more than the low level. This is because the lowest MP level of students showed the highest error of SEM value. Figure 6 shows the results of the standard error of measurement for both MAP and SLO dimensions.



(a) SEM(θ_{MAP})

(b) SEM(θ_{SLO})

Figure 6. Standard error of measurement of MAP and SLO dimensions

3.3. Item fit

All the 18 item difficulties in the MP digital diagnostic tool consists of Item 1 to 11 of MAP items and Item 12 to 18 of SLO items were examined as elucidated in Table 2. For the MAP dimension, the item difficulties ranged from –1.16 logits (SE=0.06) to + 1.27 logits (SE=0.07). However, five out of the items are polytomous scored and the difficulty should be focused on the item score threshold. In this case, the range of item difficulty in MAP ranged from -1.16 to + 1.80, and the student measures ranged from –2.31 logits to +2.47 logits. There are some students (16.50%) whose mathematics abilities are more than +1.80 logits (3.80%) and less than –1.16 logits (12.70%) and hence not 'matched' against an item location on the scale. In Figure 6 (b), there are no items matching students at either the lowest end (-1.17 to –2.31 logits) or the highest end (+1.28 to +2.47 logits) of the scale, indicating some improvements are needed for the test. That is, both easy and hard items need to be added to improve the targeting of the items for diagnosing MP, particularly easy items. There are approximately 191 students who found these test items easy and approximately 57 of them who found the items were hard. The item difficulties were appropriate for the rest of the students, approximately 1,256 students. The evidence supports the validity argument based on internal structure, as shown in Figure 5.

On the other hand, for the SLO dimension, the item difficulties were ranged from –2.27 logits (SE=0.05) to + 1.90 logits (SE=0.05), and the student ability was ranged from –2.52 logits to +2.48 logits. Results revealed that there are some students (5.90%) whose MP is more than +1.90 logits (5.60%) and less than –-2.27 logits (0.30%) and hence not 'matched' against an item distribution on the scale. In Figure 6 (b), there are no items matching students at either the lowest end (-2.27 to -2.52 logits) or the highest end (+1.91 to +2.48 logits) of the scale, indicating some improvements are needed for the test. Nevertheless, the percentage of item difficulties were quite small, especially easy items. The easy item difficulties were inappropriate for matching to only five students. This means that the item difficulty is adequate for the digital diagnostic tool. However, when researchers considered the item score threshold as shown in Table 2, Item 16 seemed to be a very hard item. The student had a 50 percent chance of obtaining a score of 3 lower than the other items at the same score. In addition, this item was the hardest item. It should be re-oriented for diagnosing MP in the future. Table 2 shows the results of the item fit statistic analysis.

The results of the digital tool inspection are found in accordance with acceptable criteria. The test statistics consist of Unweighted Fit MNSQ (outfit) between 0.76-1.26 and Weight Fit MNSQ (infit) between 0.82-1.10 which were within the acceptable range that is between 0.75 and 1.33 [28-30] as shown in Table 1. Therefore, all 18 tasks are found in compliance with the fit.

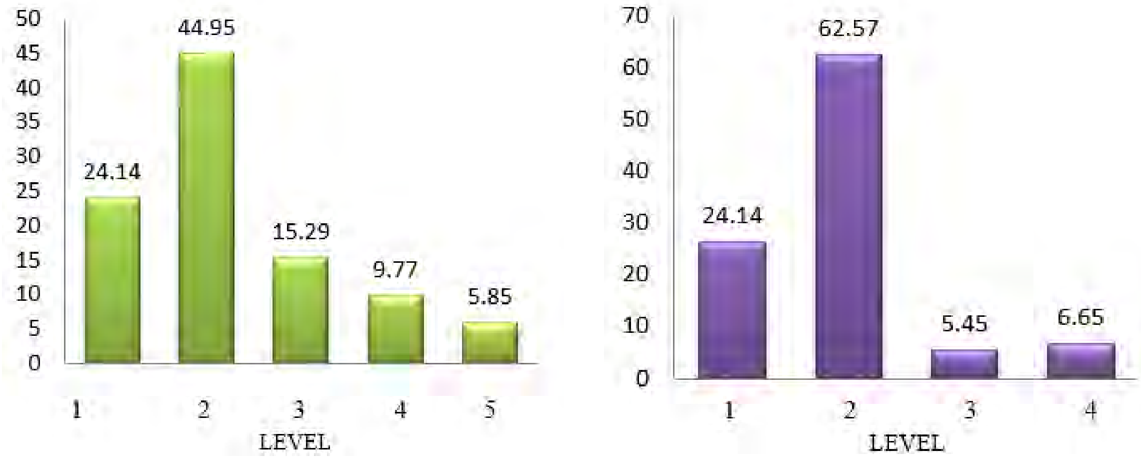
Table 2. Results of item fit statistic analysis

Item	Difficulty	SE	Outfit		Infit		Item Score Threshold			
			MNSQ	CI	MNSQ	CI	1	2	3	4
1	-0.92	0.06	1.13	(0.93, 1.07)	1.05	(0.95, 1.05)				
2	-1.16	0.06	1.26	(0.93, 1.07)	1.08	(0.94, 1.06)				
3	-0.50	0.06	1.09	(0.93, 1.07)	1.10	(0.95, 1.05)				
4	1.27	0.07	1.16	(0.93, 1.07)	1.04	(0.94, 1.06)				
5	0.03	0.06	1.06	(0.93, 1.07)	1.05	(0.96, 1.04)				
6	-0.44	0.06	1.05	(0.93, 1.07)	1.06	(0.96, 1.04)				
7	0.48	0.03	1.01	(0.93, 1.07)	0.95	(0.93, 1.07)	-0.34	-0.05	0.55	1.67
8	0.26	0.03	0.80	(0.93, 1.07)	0.82	(0.93, 1.07)	-0.34	0.16	0.93	
9	0.24	0.04	0.96	(0.93, 1.07)	0.95	(0.94, 1.06)	-0.45	0.93		
10	0.48	0.63	1.04	(0.93, 1.07)	0.99	(0.93, 1.07)	-0.09	0.32	1.14	
11	1.06	0.03	0.76	(0.93, 1.07)	0.85	(0.92, 1.08)	0.27	1.11	1.80	
12	1.50	0.07	1.16	(0.93, 1.07)	1.01	(0.93, 1.07)				
13	-0.50	0.05	0.96	(0.93, 1.07)	0.96	(0.93, 1.07)	-2.27	1.27		
14	0.12	0.04	1.03	(0.93, 1.07)	1.03	(0.94, 1.06)	-1.02	1.26		
15	0.67	0.03	1.14	(0.93, 1.07)	1.08	(0.93, 1.07)	-0.08	0.51	1.54	
16	-0.06	0.05	0.95	(0.93, 1.07)	0.95	(0.93, 1.07)	-2.03	1.90		
17	0.12	0.04	0.88	(0.93, 1.07)	0.90	(0.94, 1.06)	-0.74	0.98		
18	0.94	0.03	1.08	(0.93, 1.07)	1.06	(0.93, 1.07)	0.21	0.79	1.77	

3.4. Overall results of MP

The overall results of the 1,504 Isan's students' MP level for each dimension are presented in Figure 7. Students' MP level was measured in five and four levels of MAP and SLO dimensions respectively. The criteria for diagnosing MP in each dimension were following the cut-off score to classify student's MP levels according to [11]'s classification in their initial study. Results of the MAP dimension showed that a total of 24.14 percent of students were at the lowest level because their logits were below -0.24. This is followed by 44.95 percent of students who were found at a moderately high level whereby their logits were ranged from -0.23 to +0.49. There were 30.91 percent of students who were having a high MAP as their logits were above 0.50.

On the other hand, the results of the SLO dimension showed that there was 24.14 percent of students are at the lowest level as their logits were lower than -0.63. The majority of the students, that was 62.57 percent, were regarded as having a moderate level because their logits are from -0.62 to +1.18. However, only 12.10 percent of students were found to be at the high level of MP as their logits were above 1.19 in the SLO dimension. Researchers would like to add more proficiency levels in the range between low and intermediate for the SLO dimension. This is because 1.18 logit should be at a high level of proficiency, thus the range of logits at the moderate level seemed to be inappropriate and not logical for interpretation.



(a) Students' MP level in MAP dimension

(b) Students' MP level in SLO dimension

Figure 7. Overall results of students' MP in the two dimensions

4. CONCLUSION

The key result of this study is a digital diagnostic tool so-called ITAR to assess the seventh-grade students' MP in the Isan region of Thailand. This ITAR tool has been validated using three arguments, namely validity, reliability, and item fit. Overall results revealed that ITAR was found to be an assessment tool to diagnose students' MP in both MAP and SLO dimensions in terms of its accuracy, consistency, and stability. On top of that, results also showed that MP was better measured using a multidimensional model rather than a unidimensional model.

The main implication of this study is that ITAR can provide rich information of those students who are at the intermediate and high level of MP. This is reflected in the results of SEM □ value for estimating latent ability in MAP and SLO dimensions were at the lowest range of logits (between 0.0 to 0.5). In this line reasoning, researchers would like to add more easy items in the MAP dimension as well as more proficiency levels of MAP, from four levels to five levels in the lowest level of the SLO dimension. Moreover, researchers found that the scoring system of some tasks in ITAR seemed to overestimate MP. Consequently, the scoring guide should be revised in the input data component and process system component of ITAR.

The main contribution of this study is the ITAR has been successfully provided formative feedback for both teachers and students to enhance their MP so they can know what students learn, what students need to learn, and how well they understand it. As a result, ITAR can be utilized to guide the learning and instruction based on multiple proficiencies. Although ITAR can be accessed freely as web-based resources and compatibility with multiple sequences and approaches, the results revealed that the process system component still has problems estimating students' MP because of the complexity of the algorithm which is a psychometric model. Future researchers should be concerned about the negative and positive consequences of using ITAR in the actual mathematics classrooms. The following consequences of using a digital diagnostic tool should be taken into consideration, for example, (i) how students improve the MP when using ITAR; (ii) how much is the students' growth rate between before and after using ITAR; (iii) how to use ITAR simply for both students and teachers, and (iv) how to integrate curriculum, instruction, and diagnostic through ITAR and link the students' progression learning whenever they upgrade in the higher level as well as in a cohort study.

ACKNOWLEDGEMENTS

The research and development activities that discussed in this article were supported by the grant from Thailand Research Fund (TRF) Advanced Research Scholar and Khon Kaen University, Thailand (Grant No: RSA6080074). The researchers would like to thank the Center for Research in Mathematics Education (CRME), Khon Kaen University for providing the supports to make the research a success. Special thanks to Faculty of Education, Khon Kaen University, Thailand for providing financial support to present our research results at the 3rd World Conference on Education 2019 (WCEDU) at Kualar Lumpur, Malaysia.

REFERENCES

[1] P. Junpeng, M. Inprasitha and M. Wilson, "Modeling of the open-ended items for assessing multiple proficiencies in mathematical problem solving," *The Turkish Online Journal of Educational Technology*, vol. 2, Special Issue for INTE-ITICAM-IDEC, pp. 142-149, 2018.

[2] J. Mensah and D. Dake, "Test, measurement, and evaluation: understanding and use of the concepts in education," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 9, no. 1, pp. 109-119, 2020.

[3] U. Zainiyah and Marsigit, "Improving mathematical literacy of problem solving at the 5th grade of primary students," *Journal of Education and Learning (EduLearn)*, vol. 13, no. 1, pp. 98-103, 2017.

[4] Y-M. Huang, S-H. Huang and T-T. Wu, "Embedding diagnostic mechanisms in a digital game for learning mathematics," *Education Technology Research and Development*, vol. 62, no. 2, pp. 187-207, 2014.

[5] B. M. N. B. Bakar, "The process of thinking among junior high school students in solving HOTS question," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 4, no. 3, pp. 138-145, 2015.

[6] M. Marwiang, P. Junpeng, and N. Nakorn, "The development of a model for mathematics classroom assessment: Collaborative assessment pyramid," *Procedia Social and Behavioral Sciences,* vol. 143, pp. 764-768, 2014.

[7] P. Junpeng, "The development of classroom assessment system in mathematics for basic education of Thailand," *Procedia Social and Behavioral Sciences*, vol. 69, pp. 1965–1972, 2012.

[8] D. Fouryza, S. M. Amin, and R. Ekawati, "Designing lesson plan of integer number operation based on fun and easy math (FEM) approach," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 8, no. 1, pp. 103-109, 2019.

[9] C. Redecker and O. Johannessen, "Changing assessment – towards a new assessment paradigm using ICT," *European Journal of Education*, vol. 48, no. 1, pp. 79-96, 2014.

[10] P. Patel and A. Thakkar, "The upsurge of deep learning for computer vision applications," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 538-548, 2020.

[11] 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.

[12] M. Marwiang, J. Klaharn, L. Saree, and P. Junpeng, “Assessing students’ mathematical problem solving skill through the innovative lesson study and open approach,” *Turkish Online Journal of Educational Technology*, vol. 16, Special Issue, pp. 385-395, 2017.

[13] M. Marwiang, *et al.*, “Assessment of learning progression on mathematical problem solving of students using open approach,” *Journal of Physics: Conference Series*, vol. 1340, no. 1, pp 1-7, 2019

[14] J. B. Briggs and K. Collis, *Evaluating the quality of learning: The SOLO taxonomy*, New York: Academic Press, 1982.

[15] M. R. Wilson, *Constructing Measures: An Item Response Modeling Approach*, Mahwah, NJ: Lawrence Erlbaum Assoc., 2005.

[16] R. J. Adams, “Reliability as a measurement design effect,” *Studies In Educational Evaluation*, vol. 31, no. 2-3, pp. 162-172, 2005.

[17] R. J. Adams, M. R. Wilson and W. Wang, “The multidimensional random coefficient multinomial logit model,” *Applied Psychological Measurement,* vol. 21, no. 1, pp. 1-23, 1997.

[18] Thailand Ministry of Education, *Learning Standards and Indicators Learning of Mathematics (revised edition 2017) according to the Core Curriculum of Basic Education, B.E. 2551*. Bangkok: Printing House, Agricultural Cooperative of Thailand, 2017.

[19] M. Custer, “Sample size and item parameter estimation precision when utilizing the one-parameter ‘rasch’ model,” *The annual meeting of the mid-western Educational Research Association*, Evanston, Illinois, 21-24 October, 2015.

[20] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, “Supervised data mining approach for predicting student performance,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 16, no. 3, pp. 1548-1592, 2019.

[21] M. Kumar and A. J. Singh, “Evaluation of data mining techniques for predicting student’s performance,” *Modern Education and Computer Science*, vol. 9, no. 8, pp. 25-31, 2017.

[22] K. E. Merraoui, A. Ferdjouni, and M. Bounekhla, “Real time observer-based stator fault diagnosis for IM,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 210-222, 2020.

[23] S. Jiang, C. Wang and D. J. Weiss, “Sample size requirements for estimation of item parameters in the multidimensional graded response model,” *Frontiers in Psychology*, vol. 7, Article 109, 2016.

[24] M. R. Wilson and P. De Boeck, “Descriptive and explanatory item response models,” In P. De Boeck & M. Wilson (Eds.), *Explanatory Item Models: A Generalized Linear and Nonlinear*. New York: Springer-Verlag, 2004.

[25] L. Yao and R. D. Schwarz, “A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests,” *Applied Psychological Measurement*, vol. 30, no. 6, pp. 469-492, 2006.

[26] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461-464, 1978.

[27] B. Duckor, K. E. Castellano, K. Těllez, D.Wihardini and M. R. Wilson, “Examining the internal structure evidence for the performance assessment for California teachers: A validation study of the elementary literacy teaching event for tier 1 teacher licensure,” *Journal of Teacher Education*, vol. 65, no. 5, pp. 402-420, 2014.

[28] American Educational Research Association, American Psychological Association and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (6th ed.). Washington, DC: American Educational Research Association, 2014.

[29] R. J. Adams and S. T. Khoo, *Quest: The Interactive Test Analysis System,* Melbourne, Vic: Australian Council for Educational Research, 1996.

[30] Y-J. I. Chen, M. R. Wilson, R. C. Irey and M. K. Requa, “An Innovative measure of orthographic processing: Development and initial validation,” *Language Testing*, vol. 37, no. 3, pp. 1-18, 2020.