PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

Using Machine Learning to Score Multidimensional Assessments of Students' Skill Levels in Mathematics

Doungruethai Chitaree[1], Putcharee Junpeng[1]*, Suphachoke Sonsilphong[2] and Keow Ngang Tang[3]

[1]Faculty of Education, Khon Kaen University, 40002 Khon Kaen, Thailand
[2]Faculty of Medicine, Khon Kaen University, 40002 Khon Kaen, Thailand
[3]Faculty of Business, Hospitality and Humanities, Nilai University, 71800 Nilai, Negeri Sembilan, Malaysia

ABSTRACT

This research aims to establish a mathematical skill measurement model to examine seventh-grade students' mathematical skills in two aspects: their understanding of mathematical processes and the concept and structure. The researchers surveyed the mathematical skills of 521 seventh-grade students from the northeastern province of Thailand. Their test results were used to prototype a mathematical skill measurement model using machine learning. It involved a design-based approach that included four stages: a construct map, item design, a Wright Map, and outcome space, the so-called Multidimensional Random Coefficient Multinomial Logit Model, to verify its quality. The initial findings revealed the creation of a construct map consisting of five levels. The researchers determined the cut-off point in the form of the threshold level after considering the Wright Map criteria area for each aspect. Lastly, the measurement model was examined to provide adequate evidence of the internal structure's validity and reliability. In conclusion, students' skill levels can be measured accurately using multidimensional assessments, even though the levels of mathematical capabilities of the students varied from low to moderate to high. Therefore, it provides significant evidence of the mathematical skill measurement model to diagnose seventh-grade students' learning. The significant implications contributed to educational measurement and evaluation are that machine learning algorithms can provide more accurate and consistent scoring of assessments compared to human graders. With accurate assessment using machine learning, teachers can gain deeper insights into individual students' mathematical skills across multiple dimensions.

Keywords: Construct modeling approach, machine learning, mathematical skill measurement model, Rasch model analysis, seventh-grade students

ARTICLE INFO

Article history:
Received: 18 August 2023
Accepted: 09 November 2023
Published: 19 March 2024

DOI: https://doi.org/10.47836/pjssh.32.1.10

E-mail addresses:
Doungruethai_c@kkumail.com (Doungruethai Chitaree)
jputcha@kku.ac.th (Putcharee Junpeng)
suphason@kku.ac.th (Suphachoke Sonsilphong)
kn.tang@nilai.edu.my (Keow Ngang Tang)
*Corresponding author

ISSN: 0128-7702
e-ISSN: 2231-8534

INTRODUCTION

The results of the 2018 Program for International Student Assessment (PISA) examinations showed that Thai students consistently perform below the international average in core subjects (Organization for Economic Cooperation and Development [OECD], 2019). A total of 47% of Thai students achieved a Degree 2 or greater in mathematics compared with an OECD average of 76%. It means that Thai students are capable of understanding and knowing in what way an easy condition can be embodied mathematically without direct instructions (OECD, 2019). Moreover, only 2% of students in Thailand achieved Degree 5 or greater in mathematics against an OECD average of 11%. If we compared the six Asian countries and economies in the form of Korea (21%), Chinese Taipei (23%), Macao, China (28%), Hong Kong, China (29%), Singapore (37%), and Beijing, Shanghai, Jiangsu, and Zhejiang, China (44%), this result shows that comparatively few Thai students are able to choose, compare, and assess problem-solving approaches for handling complex situations mathematically (OECD, 2019). The reasons behind the low PISA scores of Thai students are complex and multifaceted, involving various social, economic, and educational factors. However, machine learning can assist in identifying correlations between various factors and educational outcomes, even though machine learning techniques do not directly explain the low performance in terms of PISA scores.

Zhai et al. (2020) emphasized the importance of machine learning, which increases the possibility of transforming assessment and achieving educational aims, particularly in the promising field of artificial intelligence, which indicates a substantial technological advancement. Furthermore, Howell and Walkington (2020) suggested the use of three-dimensional learning to assist students' understanding and skills development to achieve the requirements of a contemporary STEM (science, technology, engineering, mathematics) workforce. The concept of three-dimensional learning relates to when students apply corrective fundamental thoughts in a cycle with mathematical practices and crosscut the concepts to solve mathematical problems (Harris et al., 2019; Wilson et al., 2024). However, there are comparatively limited studies employing machine learning to measure items involving students executing assignments that relate to applying various components of mathematical knowledge and skills to make sense of specific occurrences (Zhai et al., 2020). As the field of machine learning continues to evolve, it is important for us to stay updated on the latest research to gain insights and make informed decisions when implementing machine learning in mathematics curricula specifically and STEM curricula generally.

The STEM curriculum blends those four subjects to teach 21st-century skills or meet students' needs if they wish to possess the skills needed to compete in the future workplace (Craig, 2021). Therefore, Leyva et al. (2022) emphasized the importance of mathematics skills, which are not only primarily significant for ordinary career roles but also are a sure sign of wider intellectual skills, particularly numeracy and numerical problem-solving. Since intellectual skill is one of the key features of career accomplishment, assessing our students' mathematical skills is an excellent way to determine their capacity to flourish on the job (Leyva et al., 2022). In other words, mathematical skills such as problem-solving abilities, quantitative reasoning, and analyzing and modeling real-world situations are particularly crucial in STEM careers.

Based on the above problems, the researchers found a need for more research into the mathematical skill measurement model pertaining to its diagnostic assessment and development. The mathematical skill measurement model is a potential tool for developing mathematical capabilities instead of just a tool to report students' mathematical skills at a certain level (Alfayez, 2022). In this sense, this research is of relevance, given that it utilizes the concept of machine learning (Maestrales et al., 2021; Zhai et al., 2020) to score multidimensional assessments of students' skill levels in mathematics that will not only holistically assess students' mathematical learning, but will also inform mathematics teachers' practices to support them in further diagnosing and developing their students' mathematical skills. As such, in this study, the researchers posed the following research aims: (i) to analyze seventh-grade students' test results for a holistic understanding of their mathematical skills in two aspects, namely mathematical processes as well as the concepts and structure of mathematics, and (ii) to design an automated interactive feedback system using machine learning for students operating at three different mathematical skill levels as a measurement model, and ultimately use it to lead to the development of students' mathematical capabilities.

Literature Review

Past researchers have explored machine learning and utilized it in scoring students' mathematical skills. Maestrales et al. (2021) developed measurement items that exceed routine memorization assignments with regard to scientific tasks that need rich knowledge and the application of thinking to enhance students' science learning. These measurement items are generally performance-based constructed answers and require machinery association to alleviate teachers' responsibility for scoring. The researchers examined the precision of a machine learning content assessment procedure as an option to the human marking of composed answers. A total of 6,700 students were selected as test-takers of the over 26,000 responses in chemistry and physics. The researchers instructed human raters and undertook a vigorous coaching session to build machine algorithmic models and cross-validate the machine results. Their results indicated that human raters generated great (Cohen's k = .40-.75) to outstanding (Cohen's k >.75) interrater consistency on the measurement elements with a diverse number of aspects. A parallel result revealed that the machine scoring algorithms accomplished equivalent marking precision to the human raters on the identical elements. Furthermore, their results showed that replies involving conventional terminology, for example, 'velocity,' were likely to generate fewer machine-human settlements, which can be related to the fact that not many students made use of conventional words matched with the casual options.

Phaniew et al. (2021) successfully developed a mathematical skill measurement model to establish the specifications for determining the skill levels in Measurement and Geometry. They used a design-based research approach to collect data from 517 Secondary Year 1 students involved as test-takers. They then used the Multidimensional Random Coefficient Multinomial Logit Model (MRCMLM) to analyze the specifications set of the mathematical skill measurement model. Their results revealed that the standard-setting of the mathematical skill measurement model can deliver significant evidence for individual students who have higher than the lowest degree of mathematical skill.

Chinjunthuk et al. (2022) designed and inspected a digital learning platform's quality in detecting seventh-grade students' mathematical skill stages associated with the subjects of Measurement and Geometry. They used an MRCMLM to inspect the quality of the measurement model. Their results indicated that there are three characteristics of proof to maintain the quality of the mathematical capabilities measurement model they created. Their empirical results revealed that the digital learning platform was extremely fitting in its usefulness, suitability, and accuracy apart from feasibility, which was determined to be only fairly suitable.

Junpeng et al. (2020) employed a multidimensional approach to creating a diagnostic framework consisting of 11 and 7 tasks, respectively, for mathematical skills in understanding mathematical processes (SMP) and mathematical skills in understanding the concept and structure of mathematics (CSM). They selected 1,504 Thai seventh-grade students to take the test on Numbers and Algebra. Their results showed internal structural evidence of validity based on the comparison of model fit and the Wright Map. Moreover, their results showed that reliability evidence and item fit are compliant with the quality of the digital instrument as indicated in the analysis of standard error of measurement and the infit-and outfit of the items. Therefore, they have created a digital instrument capable of delivering effective data, especially to students with average and superior mathematical skills.

Machine learning diagnostic assessment is essential for mathematics teachers, in particular, because it permits them to assess their students' understanding of a subject topic or to decide on the basis of their mathematics skills (Chinjunthuk et al., 2022). Moreover, mathematics teachers can also use the diagnostic results of mathematical skills to offer corrective teaching or to allocate students to lessons suited to their capability. According to Corrêa and Haslam (2021), mathematics classes have to allow for a broader focus involving conceptual discussion and exploration. Therefore, it is indisputable that mathematics involves thinking, reasoning, evaluating, and inferring. That is, the processes involved in doing mathematics are not straightforward and are not exempt from errors or misleading paths. As emphasized by Corrêa and Haslam, the essence of mathematics involves students' mathematical skills in understanding the mathematical processes, concepts, and structures, which are far from memorizing procedures and formulas. These studies highlight the potential of machine learning when it comes to automating the scoring process for assessing students' mathematical skills. By analyzing various features and patterns within students' responses, machine learning algorithms can provide efficient and objective scoring, freeing up valuable time for teachers.

The above literature review indicated that existing mathematical skill measurement methods face several issues and challenges that can impact their accuracy and effectiveness. These issues include limited assessment scope, one-size-fits-all approach, cultural and language bias, lack of real-world context, ignoring process over outcomes, and technology-driven changes. Maestrales et al. (2021) commented that many traditional assessments focus on a narrow range of mathematical skills, often emphasizing rote memorization and computation over problem-solving, critical thinking, and real-world applications. This limited scope may not adequately capture a student's overall mathematical proficiency. Moreover, standardized tests use a uniform set of questions for all students, regardless of their individual learning needs and abilities (Phaniew et al., 2021). According to Chinjunthuk et al. (2022), some assessments focus on abstract mathematical concepts divorced from real-world applications. It can make it difficult to assess a student's ability to apply mathematical skills in practical applications. Corrêa and Haslam (2021) stated that traditional assessments often prioritize getting the correct answer over understanding the problem-solving process. It can fail to recognize the importance of developing problem-solving strategies and critical thinking skills. As technology advances, this study uses machine learning to score multidimensional assessments of students' skill levels in mathematics, which need to evolve because traditional paper-and-pencil tests may not effectively assess digital literacy and the ability to work with mathematical tools.

MATERIALS AND METHODS

The researchers employed construct modeling to introduce a real-time digital learning platform involving machine learning, which was embedded in the subject of Numbers and Algebra. It is done by designing diagnostic tasks to measure the level of mathematical skills of seventh-grade students (Wilson, 2005). In addition, a design-based research approach was embraced by Vongvanich (2020). It involved four phases using machine learning to score multidimensional assessments of students' skill levels in mathematics. Finally, the MRCMLM was applied to confirm the quality of the mathematical skill measurement model (Adams et al., 1997).

Population and Sample

A total of 521 seventh-grade students in the 2017 academic year from schools under the administration of the Office of Secondary Education Service Area in the northeastern province of Thailand were purposively designated to be the test-takers. Seventh-grade students were chosen because they faced various challenges when it came to mathematics. Curriculum, teaching methods, and individual student characteristics could influence these challenges. Some common problems they encountered in mathematics include the complexity of topics, lack of foundation, fear of mistakes, mathematics anxiety, and resource constraints (Inprasitha, 2022). The key purpose of employing the purposive sampling method was to confirm that the selected test-takers were reasonably presumed to be demonstrative of a cross-section of the population in that they vary in the levels of mathematical skills from low to moderate to high, as indicated in their assessments. As emphasized by Wright and Stone (1979), a sample size of more than 500 test-takers is required to acquire adequate quality and data for analysis when consuming the multidimensional test response theory, As emphasized by Wright and Stone (1979), a sample size of more than 500 test-takers is required to ensure adequate quality and acquire sufficient data for analysis when the multidimensional test response theory is used in practice.

Research Procedure

The researchers utilized the four building blocks introduced by Wilson and Sloane (2000) as the segments of the measurement scheme. These four building blocks were: (i) a progressive standpoint with regard to students' understanding-construct map, (ii) the connotation between teaching and measurement-item design, (iii) teacher managing classroom teaching and measurement-Wright Map, and (iv) the construction of high-quality indicators-outcome space.

The first step in creating a construct map was defining a measurement construction by the researchers. A construct is an abstract concept or characteristic the researchers intend to assess, such as mathematical skills. To clearly articulate the two aspects of the construct, the researchers undertook a survey to explore the mathematical skills of seventh-grade students based on their test results to create a diagnostic digital tool as a prototype of a mathematical skill measurement tool in the first phase. It was followed by the second step of identifying the key concepts and sub-constructs. The researchers broke down the construct into its constituent parts. Since the construct in this study is "mathematical skills," sub-constructs might include algebraic skills, geometry skills, and problem-solving abilities. In the third step, a hierarchical structure of the construct and its sub-constructs is established to determine the relationship of the sub-constructs. Some sub-constructs might be more fundamental and serve as building blocks for higher-level constructs. The researchers created this visual representation to illustrate the hierarchical relationships. In the final step, the researchers define the relationships between the construct, its sub-constructs, and related constructs. This step identifies the contribution of the sub-constructs to the overall construct and their interaction. These relationships helped design a real-time automatic digital platform to accurately report their machine learning and interpret results.

In the second phase, the researchers piloted an advanced prototype to generate a construct map for each mathematical skill to fit the real setting. In the third phase, a sequence of interactive testing sets and fine-tuning of resolutions in practice was executed. The researchers used MRCMLM, a form of the Rasch Model (Rasch, 1960), to generate quality evidence of the value of the real-time digital learning platform for diagnosing levels of mathematical skill on the part of seventh-grade students.

In the final phase, researchers reported validity and reliability indications to improve the advanced prototype of the diagnostic digital tool for seventh-grade students' mathematical skills. The researchers chose an appropriate assessment method to measure the construct for item design. These diagnostic digital tools took the form of a multiple-choice test with four options as part of an online testing system to assess the Number and Algebra content elements, which consisted of 13 items in the form of so-called "e-Mat-Testing." The e-Mat-Testing was chosen considering the nature of the construct and the educational context. All test tasks were created in line with the Revised Edition of the Core Curriculum of Basic Education (Thailand Ministry of Education, 2017).

In addition, the researchers developed specific items found in the construct map and aligned with the construct and sub-constructs. The researchers also ensured the clarity, no ambiguity, and biases of each item. Item content reflected the intended learning outcomes and cognitive complexity associated with the construct. The mathematical skills diagnostic digital tool focuses on mathematical skills in understanding mathematical processes (SMP) and mathematical skills in understanding concepts and the structure of mathematics (CSM). The e-Mat Testing consists of 13 SMP and CSM aspects, with nine and six items, respectively. However, items 8 and 9 included mathematical skills for the SMP and CSM aspects. The internal structure of the e-Mat Testing approach was then designed and assessed by undertaking model fit and interdimensional correlation analysis.

The researchers analyzed the item fit of the mathematical skill measurement model. The researchers utilized the Wright Map to calibrate its difficulty and discrimination parameters. Calibration involved determining the complexity of each item and students positioned accordingly based on their estimated abilities and item difficulties. The Wright Map also helped researchers to understand the contribution of different items to the measurement of the construct. The researchers examined the Wright Map to identify anomalies, such as ill-fitted items and unresponsive students. This evaluation was crucial to ensure the validity and reliability of the measurement scale. These four building blocks, which consist of construct mapping, item design, Wright Map, and outcome space analysis, ensured that the assessments were valid, reliable, and aligned with the educational objectives. These steps helped teachers to assess and measure complex constructs accurately, make data-driven decisions, and enhance the overall quality of education.

It was followed by an assessment of the level of student mathematical skill by means of the Wright Map and the Multidimensional Test Response Theory (MTRT). The Wright Map was used to validate evidence of the quality of the progress map in each aspect with regard to the subject of Numbers and Algebra. On the other hand, the MTRT is a framework that extends traditional Item Response Theory (IRT) to account for the multidimensional nature of assessments. This MTRT is suitable when assessments simultaneously measure multiple underlying traits or dimensions. The MTRT was used to develop items through the following steps: (i) identify the multidimensional construct, (ii) define the latent traits, (iii) design item for each latent trait, (iv) calibrate items separately, (v) establish the Multidimensional Model, (vi) assess cross-loading, (vii) assembly test, (viii) administer assessment and collect data, (ix) analysis and estimate data, (x) validate model, and (xi) interpret results. Therefore, the MTRT was used to guide the development of items by considering the distinct latent traits or aspects of a multidimensional construct. Items were designed to measure these aspects separately, calibrated individually, and integrated into a multidimensional model to comprehensively assess the construct. The MTRT enables researchers to accurately assess complex and multidimensional constructs to make data-driven decisions in educational contexts.

Ultimately, it helps the researchers better understand how seventh-grade students' latent traits in multiple dimensions affect their responses to test items, allowing for more accurate and informative assessment in a wide range of fields (Embretson, 2015). As a result, the MRCMLM was used to correlate the modeling of items in multidimensional skills, assess the student's skill parameters, and place the intersection points for diagnosing the level of the student's mathematical skills in each aspect. The ACER ConQuest Version 2.0 program (Wu et al., 2007) was applied to estimate the mathematical skill measurement model parameters. Figure 1 illuminates the whole research procedure.

RESULTS

The Results of the Construct Map for Each Aspect of Mathematical Skills

Phaniew et al. (2021) indicated that the most

First Phase: A test to diagnose the mathematical skill levels (Phaniew et al., 2021): A survey was conducted to explore seventh-grade students' mathematical skills based on their test results.	Second Phase: Construct Modelling (Wilson, 2005): An advanced prototype was piloted to generate a construct map for each skill to fit the real setting.	Third Phase: Multidimensional test response model (MRCML) (Wilson et al., 2006): MRCML was used to generate quality evidence of the value of the real-time digital leaning platform.	Final Phase: Machine Learning (Lee et al., 2017; Maestrales et al., 2021; Zhai et al., 2020): Pieces of validity and reliability indications were reported to improve the advanced prototype.

Figure 1. Research procedure

effective tool that can be used when it comes to evaluating the quantity and quality of junior high school students' mathematical skills is a test.

Moreover, the researchers adapted the mathematical skills measurement model in line with recommendations by previous researchers that encompass the SMP aspect (Junpeng et al., 2018) and CSM aspect (Briggs & Collis, 1982). The selection of SMP and CSM aspects as focus points in mathematical education was based on recognizing that these two aspects were important for developing well-rounded, proficient, and mathematically literate individuals. By prioritizing these areas, teachers aim to foster deeper learning, critical thinking, and problem-solving abilities in students, which are valuable not only within mathematics but also in various aspects of life and future academic pursuits (Inprasitha, 2022).

Phaniew et al. (2021) indicated that the most effective tool that can be used when it comes to evaluating the quantity and quality of junior high school students' mathematical skills is a test. Moreover, the researchers adapted the mathematical skills measurement model in line with recommendations by previous researchers that encompass the SMP aspect (Junpeng et al., 2018) and the CSM aspect (Biggs & Collis, 1982). Therefore, the researchers created a construct map comprising five skill levels for each aspect of mathematical skills. The results of the first phase demonstrated five levels of the SMP aspect: irrelevance, unrecalled, fundamental, modest, and strategic skills. In contrast, the researchers utilized the SOLO taxonomy as a model to identify, describe, and inform the five levels of the CSM aspect, namely irrelevance, pre-structural skills, uni-structural skills, multi-structural skills, and prolonged structural skills. The results are equivalent to previous research (Chinjunthuk et al., 2022; Junpeng et al., 2020; Phaniew et al., 2021). Figure 2 illustrates the two construct maps relating to mathematical skill levels for seventh-grade students' test results, as established by the researchers.

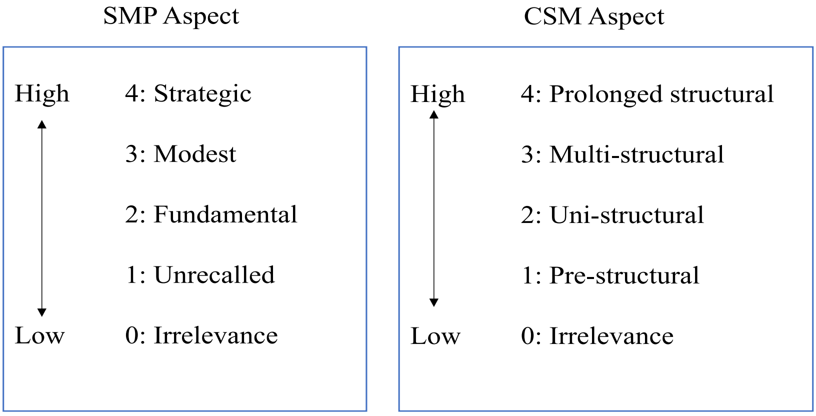

Figure 2. Construct a map of SMP and CSM aspects of 13 items in Number and Algebra

The Results of the Determination of the Intersection in Assessing Mathematical Skills

All the 13-item problems in the mathematical skill measurement model were studied, as illustrated in Table 1. The intersections were determined by the threshold level divided by the number of tests at the same level for the two aspects before considering each aspect's Wright map criteria area. In this line of reasoning, the results showed that the item difficulties ranged from -0.651 logits to +2.491 logits and -0.178 logits to +4.423 logits in the SMP aspect and CSM aspect, respectively, in the subject of Number and Algebra.

The mean threshold of mathematical skill level for each aspect was used to identify the level of mathematical skill in the subject of Numbers and Algebra to create a standardized mathematical skill measurement model. Hence, the transition point was calculated from the mean of the item thresholds in each aspect level, as shown in Table 1.

Table 1
Results of determination of the intersection in assessing mathematical skills for seventh-grade students

Mathematical Skills	Item	Difficulty	Threshold			
			1	2	3	4
Understanding Mathematical Process (SMP)	1	0.929			0.93	
	2	0.476		0.48		
	3	0.325		0.33		
	4	0.696		0.70		
	5	-0.651	-0.65			

Table 1 *(Continue)*

Mathematical Skills	Item	Difficulty	Threshold			
			1	2	3	4
	6	1.500				1.50
	7	1.084			1.08	
	8	-0.178	-0.18			
	9	2.491				2.49
	Mean		-0.651	0.499	1.007	1.500
Understanding Concept & Structure of Mathematics (CSM)	8	-0.178	-0.18			
	9	2.491				2.49
	10	4.277	2.36	4.59	5.11	5.20
	11	4.423	3.03	4.81	5.47	-
	12	1.956	0.77	2.26	2.41	2.52
	13	2.366	1.28	2.52	2.86	2.96
	Mean		1.45	3.29	3.55	3.96

The Results of Scoring Criteria and Determination of Seventh-grade Students' Mathematical Skill Levels

Based on the determination of the cut-off point in measuring the students' mathematical skills in the subject of Numbers and Algebra, the results indicated that the transition point can be allocated using four cut-off points of five levels, each in ascending order. The SMP aspect's transition point intersection results from levels 1 to 5 were reported from the lowest to the highest level at -0.65, 0.50, 1.01, and 1.50, respectively. In contrast, the intersections of the CSM aspect were recognized from Level 1 to 2, Level 2 to 3, Level 3 to 4, and Level 4 to 5 as 1.45, 3.29, 3.55, and 3.96, respectively. Figure 3 elucidates the transition point in every aspect of the Wright map with regard to the subject of Numbers and Algebra.

The Results of Multidimensional Test Response on the Part of the Seventh-grade Students

The researchers used the measurement criterion results from the Wright map to propose the specification adjustments for a mathematical skill measurement model in Numbers and Algebra. Consequently, the researchers determined five score ranges, which were transformed from assessment mathematical skill parameters into scale scores and raw scores in that order. Table 2 shows that those students who achieve logits smaller than -0.65 and 1.45 in the SMP and CSM aspects, respectively, are identified as possessing the lowest level of mathematical skill. Similarly, if their logits are greater than 1.50 and 3.96 with regard to the SMP and CSM aspects, respectively, those students are identified as having achieved the highest level of mathematical skill. The results indicate that none of the test-takers had the lowest mathematical skill level in the SMP or CSM aspects. It implies they all had a higher mathematical skill level than the minimum.

As a result, the researchers set the mathematical skill requirements of the measurement model after corroborating the criterion area discovered from the Wright map. Following this line of reasoning, the researchers established five score ranges. These were converted into the scale and raw scores from evaluating the mathematical skill parameters. Table 2 shows the results of the 517 seventh-grade students' mathematical skill levels in terms of the SMP and CSM aspects.

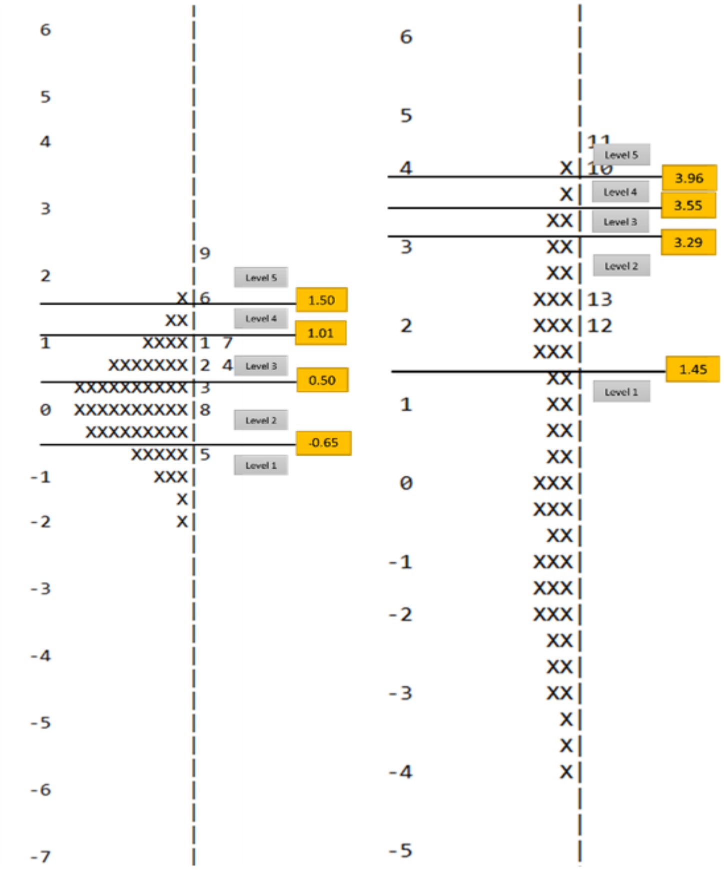

Figure 3. Wright map indicates the transition points of SMP and CSM aspects in the subject of Number and Algebra

The Results of Outcome Space of Each Aspect of the Mathematical Skill Measurement Model

In this final phase, the researchers used the 521 test-takers' test scores to create a scoring guide or outcome space for the five mathematical skill levels, with scores ranging from 0 to 4. The outcome space was identified using the concepts of Webb's depth of knowledge, as proposed by Webb (1997). Table 3 demonstrates the interpretations of the students' mathematical skill capabilities corresponding to each SMP and CSM level.

Table 2
Results of determination of mathematical proficiency standards in the subject of Number and Algebra

Aspect	Intersection θ	θ range	Scale scores	Raw scores
SMP Level 5	1.50	> 1.50	> 65.00	6 – 7
SMP Level 4	1.01	1.01 < θ≤1.50	60.00 – 65.00	5
SMP Level 3	0.50	0.50 < θ<1.01	55.00 – 60.00	4
SMP Level 2	-0.65	-0.65 < θ<0.50	43.50 – 55.00	1 – 3
SMP Level 1	-	< -0.65	< 43.50	0
CSM Level 5	3.96	> 3.96	> 89.60	15 – 16
CSM Level 4	3.55	3.55 < θ≤3.96	85.50 – 89.60	14 – 15
CSM Level 3	3.29	3.29 < θ<3.55	82.90 – 85.50	13 – 14
CSM Level 2	1.45	1.45 < θ<3.29	64.50 – 82.90	10 – 13
CSM Level 1	-	< 1.45	< 64.50	0 – 9

Table 3
The results of outcome space for SMP and CSM aspects

Level	**Score**	**Skills of understanding mathematical process (SMP) aspect**	
		Name of skill	**Proficiency level description**
5	4	Strategic skills	Students can use their strategic method to solve complex problems needing multiple steps leading to the correct answers by demonstrating a wide range of solutions and summarizing mathematical ideas to a higher skill level.
4	3	Modest skills	Students show their understanding of mathematical processes, principles, and theories to solve complex problems with various methods involving more than one step.
3	2	Fundamental skills	Students show their basic understanding of simple mathematical operations and problem-solving steps, but the answers are still incomplete.
2	1	Unrecalled skills	Students cannot remember important and necessary content knowledge or possess partial content knowledge to use as their basic skills to answer open-ended questions.
1	0	Irrelevance	Students do not answer at all.

Table 3 *(Continue)*

Level	**Score**	**Skills of understanding concept and structure of mathematics (CSM) aspect**	
		Name of skill	**Skill level description**
5	4	Prolonged structural skills	Students can link and refer to their content knowledge and draw conclusions based on their conceptual understanding.
4	3	Multi-structural skills	Students can connect, classify, explain, or describe complex concepts and relate their conceptual understanding of the given problem situation. However, students cannot summarize the body of knowledge and are unable to analyze or integrate the concepts.
3	2	Uni-structural skills	Students can connect simple but not complex concepts. They cannot show their conceptual understanding to solve the given problem even if they have completed their answers.
2	1	Pre-structural skills	Students cannot create and connect the related concepts. Therefore, they show a misunderstanding of the studied concepts and a lack of skills in interpreting the information.
1	0	Irrelevance	Students do not answer at all.

DISCUSSION

We aim to develop a typical measurement model for seventh-grade students in Numbers and Algebra. Our research demonstrated the strength of the mathematical skill measurement model created as a real-time digital learning platform for diagnosing levels of mathematical skills. This automated analysis was intended to ease the transition to multidimensional evaluation. Our results show a need for mathematics teachers to respond rapidly to technology adoption by using machine learning with regard to seventh-grade students so that they can overcome the challenges resulting from the COVID-19 pandemic that we have been dealing with for more than two years.

The ultimate results of using machine learning to score multidimensional assessments of students' skill levels in mathematics is a promising approach that can provide several advantages in education. These results parallel past research (Chinjunthuk et al., 2022; Junpeng et al., 2020; Phaniew et al., 2021). In other words, this approach has been proved by current research as well as previous research (Chinjunthuk et al., 2022; Junpeng et al., 2020; Phaniew et al., 2021), enabling us to offer more accurate, efficient, and unbiased assessment results. Besides, the mathematical skill measurement model can minimize the potential for human bias, which can be present in traditional grading systems. This result reflects that machine learning algorithms can evaluate students' performance based on predefined criteria objectively and consistently, thus the result is supported by past researchers (Wilson et al., 2024).

On top of that, the measurement model can provide detailed feedback based on individual strengths and weaknesses. This personalized feedback can help students understand their areas of improvement, which is essential for targeting learning as customized feedback (Junpeng et al., 2018). Moreover, automated scoring is another strength of using this mathematical skills measurement model, as teachers can save a significant amount of time. It allows them to focus more on teaching and providing valuable insights to students rather than spending excessive time on grading assignments and assessments, in accordance with the idea of Howell and Walkington (2020).

CONCLUSION

Implications for Practice

The major contribution of this study is using machine learning to score multidimensional assessments of students' skill levels in mathematics. This mathematical skill measurement model is particularly important for multidimensional assessments of different skills. The algorithms can process a large amount of data and identify subtle patterns that human graders might miss. Moreover, machine learning is found in this study that can significantly speed up the assessment process. Generally, grading multidimensional assessments can be time-consuming, especially when assessing various mathematical skills. Therefore, automation through machine learning can lead to quicker turnaround times, allowing teachers to provide timely feedback to students. For example, in terms of the 13 items established to appraise the subject of Numbers and Algebra with regard to seventh-grade students, they all meet the principles of a relevant item in terms of having a respectable difficulty level. The distinguishing power of the item functions well and has good validity and reliability. Amidst the increasing use of online learning during the unpredictable new normal educational transformation, this study evaluates the efficacy of a real-time digital learning platform for diagnosing levels of mathematical skills that can be used to fully support online and self-regulated learning classrooms. Moreover, machine learning can facilitate scoring two-aspect assessments more promptly than human scoring alone, permitting mathematics teachers to accumulate comprehensive evidence on students' mathematical skill-in-use.

Furthermore, teachers can identify common misconceptions and learning gaps within a class of students by analyzing the assessment results through this machine learning model. It contributes to mathematics teachers in addressing these gaps with targeted interventions. This result is consistent with the previous studies (Chinjunthuk et al., 2022; Junpeng et al., 2020; Phaniew et al., 2021); the benefits of applying the two aspects of mathematical skills have been successfully proposed in the form of a quality measurement model which not only concentrates on the mathematical learning process but also on the improvement made by each student by what is compulsory. It takes the form of an automatic feedback system, which can be used to support the development. This concern is especially pronounced because machine scoring, when measured using multi-dimensional assessments, was proficient with regard to categorizing student answers accurately. It implies a potential for machine learning to ease the assessment of comprehensive mathematics understanding. In other words, the data collected from assessments can be used to gain insights into the effectiveness of teaching methods, curriculum design, and educational materials. These insights can inform educational strategies and lead to continuous improvement in mathematics education.

Methodological and Practical Limitations

This study revealed that machine learning algorithms can generate detailed feedback for students, highlighting their strengths and areas that need improvement across various dimensions of mathematics. Although the research results showed that the real-time automatic digital platform using machine learning to score multidimensional assessments of students' skill levels in mathematics offers various advantages, it encompasses methodologies and practical limitations. Methodological limitations include data quality and quantity, curriculum misalignment, extensive and latent traits, and coherence. On the other hand, practical limitations encompass scalability, data privacy and security, real-world skills generalization, and teacher-student acceptance. Methodological limitation models require large and high-quality datasets for training. The researchers were challenged to gather extensive data for multidimensional assessments in mathematics that may alter the data quality. In this line of reasoning, inadequate or biased data can lead to inaccurate model predictions.

Another methodological limitation is the curriculum misalignment. If the assessment does not align with the curriculum, the model's predictions may not accurately reflect students' learning or skills. Moreover, multidimensional assessments involve multiple latent traits or aspects. As a result, designing methodological limitation models that effectively capture and differentiate these dimensions can be complex. It can cause the researcher to face challenges in handling the interplay between different skills and assessments in isolation. On top of that, many methodological limitation models, especially deep learning models, make it difficult to interpret their scoring decisions. This lack of interpretability can be a significant limitation when understanding a student's score.

Practical limitations such as implementing mathematical skills-based scoring for large-scale educational assessments can be challenging. It may require substantial infrastructure and computational resources to process and score assessments efficiently for a large number of students. Besides, handling sensitive student data for scoring purposes raises privacy and security concerns. Ensuring data protection and complying with regulations can be challenging. In addition, applying mathematical concepts or assessing mathematical skills in a controlled assessment environment may not always generalize real-world problem-solving. However, it is an essential skill in mathematics. Finally, teachers and students may be skeptical of mathematical skills-based scoring systems, especially if they do not understand the scoring decisions. Lack of trust in the scoring process can also be a significant barrier to adoption.

Recommendations for Future Study

This feedback can empower students to take ownership of their learning and focus on specific growth areas. Nevertheless, it is important to consider potential biases in the training data used for the machine learning model. If the training data is biased, the algorithm's predictions and scores could also be biased. The researchers suggested that teachers ensure fairness. Bias is crucial to avoid perpetuating educational inequalities. Machine scoring could ease the practice of introducing more vigorous measures when it comes to students' understanding through mathematical skill-in-use assessment. In conclusion, this study utilized machine learning to score multidimensional assessments in mathematics, which promises to enhance accuracy, efficiency, personalization, and insights in education.

ACKNOWLEDGMENTS

This research and innovation activity is funded by the National Research Council of Thailand (NRCT) (Research Project No:173806). The authors wish to extend their deepest appreciation to the National Research Council of Thailand (Grant No: N42A650236) for its essential support in providing a dataset critical to the innovation at the heart of this research. This dataset has not only been a cornerstone for the empirical analysis conducted herein but has also significantly contributed to advancing knowledge in the interdisciplinary areas of artificial intelligence and education—a sincere thank you to Mrs. Kalpana Nair a/p Kerupagaran for her attentive proofreading of this article. The authors gratefully acknowledge the use of service and facilities of the Faculty of Education, Khon Kaen University, Khon Kaen 40002, Thailand.

REFERENCES

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23. https://doi.org/10.1177/0146621697211001

Alfayez, M. Q. E. (2022). Mathematical proficiency among female teachers of the first three grades in Jordan and its relationship to their mathematical thinking. *Frontiers in Education, 7*. Article 957923. https://doi.org/10.3389/feduc.2022.957923

Briggs, J. B., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic Press. https://doi.org/10.1016/C2013-0-10375-3

Chinjunthuk, S., Junpeng, P., & Tang, K. N. (2022). Use of digital learning platform in diagnosing seventh grade students' mathematical ability levels. *Journal of Education and Learning, 11*(3), 95-104. https//doi.org/10.5539/jel.v11n3p95

Corrêa, P. D., & Haslam, D. (2021). Mathematical proficiency as the basis for assessment: A literature review and its potentialities. *Mathematics Teaching Research Journal, 12*(4), 3-20.

Craig, O. (2021, June 29). *What is STEM?* https://www.topuniversities.com/courses/engineering/what-stem

Embretson, S. E. (2015). The multicomponent latent trait model for diagnosis: Applications to heterogeneous test domains.*Applied Psychological Measurement, 39*(1), 16-30. https://doi.org/10.1177/0146621614552014

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice, 38*(2), 53-67. https://doi.org/10.1111/emip.12253

Howell, E., & Walkington, C. (2020). Factors associated with completion: Pathways through developmental mathematics. *Journal of College Student Retention: Research, Theory & Practice, 24*(1), 43-78. https://doi.org/10.1177/1521025119900985

Inprasitha, M. (2022). Lesson study and open approach development in Thailand: A longitudinal study. *International Journal for Lesson and Learning Studies, 11*(5), 1-15. https://doi.org/10.1108/IJLLS-04-2021-0029

Junpeng, P., Inprasitha, M., & Wilson, M. (2018). Modeling of the open-ended items for assessing multiple proficiencies in mathematical problem solving. *The Turkish Online Journal of Educational Technology, 2,* 142-149.

Junpeng, P., Marwiang, M., Chiajunthuk, S., Suwannatrai, P., Chanayota, K., Pongboriboon, K., Tang, K. N., & Wilson, M. (2020). Validation of a digital tool for diagnosing mathematical proficiency. *International Journal of Evaluation and Research in Education, 9*(3), 665-674. http://doi.org/10.11591/ijere.v9i3.20503

Leyva, E., Walkington, C., & Perera, H. (2022). Making mathematics relevant: An examination of student interest in mathematics, interest in STEM careers, and perceived relevance. *International Journal of Research in Undergraduate Mathematics Education, 8*, 612-641. https://doi.org/10.1007/s40753-021-00159-4

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using machine learning to score multi-dimensional assessments of Chemistry and Physics. *Journal of Science Education and Technology, 30*, 239-254. https://doi.org/10.1007/s10956-020-09895-9

Organization for Economic Cooperation and Development. (2019). *PISA 2018 results: What students know and can do.* PISA OECD Publishing. https://doi.org/10.1787/5f07c754-en

Phaniew, S., Junpeng, P., & Tang, K.N. (2021). Designing standards-setting for levels of mathematical proficiency in measurement and geometry: Multidimensional item response model. *Journal of Education and Learning, 10*(6), 103-111. https//doi.org/10.5539/jel.v10n6p103

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. The University of Chicago Press. https://doi.org/10.2307/2287805

Thailand Ministry of Education (2017). *Learning standards and indicators learning of mathematics (revised edition 2017) according to the Core Curriculum of Basic Education, B. E. 2551.* Agricultural Cooperative of Thailand. https://drive.google.com/file/d/1F4_wAe-ZF13-WhvnEAupXNiWchvpcQKW/view

Vongvanich, S. (2020). *Design research in education*. Chulalongkorn University Printing House.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers. https://www.researchgate.net/publication/234731918_Criteria_for_Alignment_of_Expectations_and_Assessments_in_Mathematics_and_Science_Education_Research_Monograph_No_6

Wilson, C. D., Haudek, K. C., Osborne, J. F., Bracey, Z. E. B., Cheuk, T., Donovan, B. M., Stuhlsatz, M. A. M., Santiago, M. M., & Zhai. X. (2024). Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching, 61*(1), 38-69. https://doi.org/10.1002/tea.21864

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Assoc. https://doi.org/10.4324/9781410611697

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208. https://doi.org/10.1207/S15324818AME1302_4

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Mesa Press. https://research.acer.edu.au/measurement/1/

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACERConQuest version 2: Generalized item response modeling software.* ACER Press. https://www.researchgate.net/publication/262187496_ConQuest_Version_2_Generalised_Item_Response_Modelling_Software

Zhai, X., Haudek, K. C., Shi, L., Nehm, R., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. J*ournal of Research in Science Teaching, 57*(9), 1430-1459. https://doi.org/10.1002/tea.21658