CINEFORUM
ISSN: 0009-7039
Vol. 65. No. 4, 2025

Development and Validation of Mathematical Thinking Test Among Undergraduate Students in The Western Zone Kingdom of Saudi Arabia

Ahmed Algamdi [1]*, Ibnatul Jalilah Yusof[2]
[1]*Faculty of Educational Science and Technology, Universiti Teknologi Malaysia.
[2]Senior Lecturer, Faculty of Educational Science and Technology,
Universiti Teknologi Malaysia. ijalilah@utm.my
Corresponding Author Email ID: **Ahmed-ahm-2@hotmail.com**

Abstract:
This study developed and validated a 50-item instrument to measure MT among undergraduate students in public universities in western KSA (Jeddah, Makkah, and Ta'if), addressing gaps in existing tools that rely on Classical Test Theory and target younger learners. Grounded in Item Response Theory (IRT) using the one-parameter logistic (Rasch) model, the instrument encompasses six dimensions: Generalization, Induction, Deduction, Use of Symbols, Logical Thinking, and Mathematical Proof. A cross-sectional survey design was employed with a stratified random sample of 229 participants. Content validity was established through expert review by six professors, yielding I-CVI scores of 0.833–1.00, S-CVI/Ave = 0.97, S-CVI/UA = 0.82, and excellent Fleiss' kappa agreement. Reliability analyses revealed Cronbach's alpha = 0.95, person reliability = 0.95 (separation = 4.25), and item reliability = 0.96 (separation = 5.14). Model-data fit confirmed unidimensionality (first-contrast eigenvalue = 1.8 < 3.0), local independence (residual correlations < 0.50), and productive item fit (infit/outfit MNSQ = 0.5–1.5). Differential item functioning (DIF) analysis identified minimal bias, with only two items flagged for gender and 4 for university location. The instrument aligns with Saudi Vision 2030 by providing a robust, psychometrically sound tool for enhancing pre-service teacher training, curriculum design, and MT assessment in higher education.
Keywords: Mathematical Thinking, Generalization, Induction, Deduction, Use of Symbols, Logical Thinking, Mathematical Proof, Item Response Theory (IRT).

Introduction:

Thinking is a fundamental component of human behavior and is closely tied to problem-solving and cognitive development, particularly in education (Laland, Brown, & Brown, 2011; Çelik & Özdemir, 2020; McPeck, 2016). Cognitive abilities play a crucial role in students' capacity to learn and retain information, emphasizing the importance of teaching them how to think and study effectively (Schoenfeld & Sloane, 2016). Among the various types of thinking**,** Mathematical Thinking (MT) is particularly significant. MT involves applying mathematical principles, logical reasoning, and problem-solving to address challenges and analyze patterns (Henderson et al., 2002; Parshall, 2006). It encompasses skills such as logical reasoning, pattern recognition, and data

analysis, which are essential for understanding complex information, making informed decisions, and solving real-world problems (Zaman, 2011; Melnik, 2005). MT is a dynamic process that facilitates understanding, critical reasoning, and problem-solving by combining ideas and evaluating arguments systematically (J. Mason, Burton, & Stacey, 2010; Chukwuyenum, 2013).

Numerous studies emphasize the importance of fostering MT from early schooling to enhance cognitive skills (Aizikovitsh-Udi & Cheng, 2015; Goos & Kaya, 2020; Sam & Yong, 2006; Van Oers, 2010). In Kingdome of Saudi Arabia (KSA), the educational philosophy prioritizes thinking skills, as highlighted in initiatives like Vision 2030, which aims to enhance cognitive strategies across all educational levels (Alharti & Evans, 2017; Aljabreen & Lash, 2016; Yaacob et al., 2019). Despite such efforts, KSA's low performance in international assessments, such as TIMSS and GAT, underscores the need to improve students' MT skills and teachers' ability to foster them (Alatawi, 2022; Almazroo, 2024).

TIMSS evaluates students' proficiency in mathematics and science, emphasizing conceptual understanding, logical reasoning, and problem-solving, while GAT assesses analytical and deductive skills (Alatawi, 2022; Alahmadi, 2019). Research has shown that intermediate students in KSA possess low levels of MT (Albaqawi, 2023; Alahmadi, 2019), prompting recommendations to enhance teacher training, develop educational programs, and include classroom activities to foster MT skills (Albaqawi, 2023). Teacher competency in MT is crucial, as teachers with strong MT skills can effectively guide students in developing their problem-solving and reasoning abilities (Schoenfeld & Sloane, 2016).

This study aims to address gaps in the existing literature by developing a valid and reliable instrument to measure MT levels among undergraduate students training to become teachers in KSA. Most existing tools, such as those used by Alatawi (2022) and Albaqawi (2023), focus on younger students and lack constructs tailored to pre-service teachers. Additionally, these tools often rely on Classical Test Theory (CTT), which estimates reliability but may not ensure validity across different populations (Crocker & Algina, 1986). To address these limitations, this study will develop a new MT instrument specifically designed for undergraduate teacher candidates, incorporating theoretical frameworks, literature reviews, and empirical research (Arfat, Shahid et al. 2025)Cowdell & Dyson, 2019; Weintrop et al., 2016).

By providing a reliable tool for assessing MT in pre-service teachers, this research will contribute to teacher training programs, curriculum design, and policy development, aligning with Vision 2030's goals. The findings will help identify areas for improvement in developing essential MT skills, ultimately enhancing the quality of education in KSA and preparing future educators for effective teaching (Bilgiç & Azak, 2019; Caravolas et al., 2019; Schoenfeld & Sloane, 2016).

Background of this Study and Problem Statement:

Equipping students with thinking skills is vital for adapting to rapid changes in society, the economy, science, and technology (Kereluik et al., 2013; Birgili, 2015). Critical and accurate thinking is fundamental for students to acquire and apply knowledge effectively (Halpern, 2013).

Education must focus on teaching students how to think and learn, emphasizing the cultivation of thinking abilities (Entwistle, 2013).

MT is a critical subset of thinking skills, defined as the ability to solve mathematical problems using numbers, symbols, shapes, and concepts (Aizikovitsh-Udi & Cheng, 2015; Abed & Zeinah, 2012). Beyond academia, MT enhances problem-solving, decision-making, and reasoning skills in real-life contexts (Chew et al., 2019). In education, MT plays a pivotal role in nurturing logical reasoning, numeracy, data analysis, and critical thinking across subjects (Chiam et al., 2014; Mitana et al., 2021; Saad, 2020).

MT is multifaceted and involves logical reasoning, artistic sensibility, and iterative refinement of ideas to ensure coherence and validity (Dreyfus & Eisenberg, 2012; Sternberg & Ben-Zeev, 2012). Researchers have categorized MT into two main approaches: mathematical processes and social processes (Isoda & Katagiri, 2012; NCTM, 2000). Psychologists highlight that MT activates complementary brain systems for precise calculations and visual processing (Fischbein, 1999; Roth & Walshaw, 2019).

Assessments like TIMSS, PISA, and NAEP evaluate MT globally, focusing on concepts such as problem-solving, reasoning, and procedural fluency (Crowe & Sheppard, 2011). In KSA, national tools like the Secondary School Certificate Examination and University Admission Tests address MT within the local curriculum, emphasizing procedural fluency, conceptual understanding, and problem-solving. However, these tools often overlook regional and cultural nuances, leaving significant gaps in the comprehensive evaluation of MT (Altannir et al., 2019).

In KSA, particularly in the Western region, several factors affect the development of MT:

- Teaching Methods: Traditional rote learning limits critical thinking and problem-solving, but a shift toward student-centered approaches shows promise (Alrashdi & Almutawa, 2022).
- Socioeconomic Disparities: Variations in resources and access to educational opportunities create inequalities in MT skills (Hein et al., 2015).
- Cultural and Regional Identity: Cultural attitudes toward mathematics, STEM fields, and language skills influence students' engagement and success in MT (BinAli, 2014).

This study redefines MT for Saudi undergraduate students, particularly in teacher-training programs, through a meta-analysis of the literature. The proposed framework highlights key dimensions, including:

- Generalization: Recognizing patterns and applying mathematical rules to broader contexts.
- Induction and Deduction: Using logical reasoning to form and validate conclusions.
- Symbol Understanding: Comprehending and applying mathematical symbols.
- Logical Thinking and Proofs: Constructing systematic arguments to validate mathematical statements (Baker, 2010; Devlin, 2012; Dreyfus & Eisenberg, 2012).

Focusing on undergraduate students preparing to become teachers is essential for understanding the development of MT. Unlike current teachers, who may have rigid practices, undergraduate students are still shaping their skills, making them more adaptable to new approaches. This focus

can also identify systemic weaknesses in education early, allowing for targeted improvements (Ball, Lubienski, & Mewborn, 2001; Wahlstrom & Louis, 2008).

Despite the Saudi government's efforts to enhance education, students' proficiency in MT remains below acceptable standards, as reported by the Ministry of Education and other studies (Al-Sadan, 2000; Alamri, 2011; Battal, 2016). Many students lack foundational mathematical skills, such as addition, subtraction, multiplication, algebra, and geometry (P. Brown, 2016; Hestenes & Sobczyk, 2012). This issue is compounded by traditional rote learning methods and insufficient emphasis on problem-solving and critical thinking in schools, colleges, and universities (Altakhayneh, 2022; Allmnakrah & Evers, 2020).

MT issues are not unique to KSA; similar challenges have been identified worldwide. For instance, studies in Malaysia, the USA, and Australia show that both students and teachers need stronger MT skills to meet modern educational demands (Beng & Yunus, 2015; Jacobs, Lamb, & Philipp, 2010; Stacey, 2006). These findings highlight the global relevance of improving MT education and its assessment.

Current MT assessments in Saudi Arabia face several challenges, including:

1. Insufficient Validity and Reliability: Existing tests lack robust evidence of validity and reliability (Kargar, Tarmizi, & Bayat, 2010; Onal, Inan, & Bozkurt, 2017).
2. Limited Standardization: Test items often fail to meet psychometric standards, making assessments less effective (Khoshaim & Rashid, 2016).
3. Focus on Procedural Fluency Over Conceptual Understanding: Many assessments prioritize rote memorization and computational speed, neglecting critical problem-solving and reasoning skills (Alam, 2022; Guggemos, Seufert, & Román-González, 2023).
4. Lack of Creativity Opportunities: Over-reliance on algorithms stifles students' ability to explore multiple approaches and think creatively (Marrone, Cropley, & Wang, 2023; Miranda & Mamede, 2022).

A new approach to MT assessment in KSA emphasizes meaningful, criterion-referenced evaluations that enhance learning rather than merely measuring performance. Continuous assessment methods, including diagnostic, formative, and summative evaluations, can help students identify areas for improvement and foster deeper engagement with mathematics (Khormi & Woolner, 2019; Alafaleq & Fan, 2014).

To improve MT assessments, comprehensive strategies should:

- Focus on problem-solving skills, enabling students to apply mathematical concepts in real-world scenarios (Özpinar & Arslan, 2023).
- Balance assessments between computational speed and reasoning abilities to ensure a fair evaluation of students' potential (Tang et al., 2020).
- Encourage creative thinking by integrating problem-solving and problem-posing tasks into assessments (Munakata et al., 2023).

- Prioritize conceptual understanding alongside procedural skills to strengthen students' ability to apply mathematics in practical contexts (De Zeeuw et al., 2013; Sebsibe & Feza, 2019).

Materials and Method:

Research Design:

A quantitative research approach was employed using a cross-sectional survey design, which allowed for data collection from a population or representative subset at a specific point in time. This design facilitated the analysis of relationships among variables by obtaining all sample measurements simultaneously, thus indicating temporal relationships (Sedgwick, 2014). Cross-sectional studies are effective in examining health outcome prevalence, investigating influencing factors, and describing population characteristics. Unlike longitudinal studies, they do not track individuals over time, making them cost-effective and easy to implement. These studies provide initial evidence that can guide the design of more complex future research, which aligns with the goals of the current study (Wang & Cheng, 2020).

Population, Sample, and Sampling Techniques:

The study targets undergraduate students in education programs at public universities in western Saudi Arabia—specifically Jeddah, Makkah al-Mukarramah, and Ta'if. This focus is supported by previous research indicating a need for enhanced mathematical preparation and engaging teaching methods (Fnais et al., 2015; Mania & Alam, 2021; Sithole et al., 2017). The study emphasizes the importance of developing students' mathematical thinking, particularly higher-order skills such as logical reasoning and problem-solving (Stylianides et al., 2024). A representative sample is crucial for research, and sample size is defined as the number of respondents surveyed (Krejcie & Morgan, 1970). While simpler models may require a sample size of 100 (Linacre, 1994), more complex models suggest sizes between 200 and 500 (Orlando & Marshall, 2002; Thissen et al., 1986; Tsutakawa & Johnson, 1990). This study calculated a sample size of 229 using the Thompson formula, aligning with Rasch measurement standards for reliability (Thompson, 1933, 1935). Stratified random sampling will be used to ensure fair representation of male and female students and to address potential differential item functioning (Aoyama, 1954; Nistor, 2013). This study uses a sample of 229 participants, a size calculated using the Thompson formula for finite populations. This number is justified as it meets the Rasch measurement standard of 200-500 cases for reliable results. To ensure the sample is representative, a stratified random sampling technique will be used, grouping the population by gender and university location to account for potential differential item functioning.

Results:

After analysing the instrument data, the results were presented in a table aligned with the research questions.

First Research Question:

What are Rater Agreements Index using Content Validity Index (CVI) and Fleiss' Kappa? To assess and estimate the content validity of the test items, the content validity index (I-CVI) and Fleiss's kappa were used (Yusoff 2019). The experts' ratings of the items were used to compute the ratio. Table 1 showed the deleted items after doing the Content Validity Index and Fleiss' Kappa and the content validity index is presented in Table 2.

Table 1 The Deleted Items After Doing the Content Validity Index and Fleiss' Kappa

	Items	**Number giving rating of 3 or 4 to relevancy of item**	**I-CVI***	**Pc****	**K*****	**Interpretation**
5	5	4	0.66	0.23	0.56	Fair Agreement
12	2	4	0.66	0.23	0.56	Fair Agreement
14	4	4	0.66	0.23	0.56	Fair Agreement
16	6	4	0.66	0.23	0.56	Fair Agreement
26	6	4	0.66	0.23	0.56	Fair Agreement
28	8	4	0.66	0.23	0.56	Fair Agreement
29	9	4	0.66	0.23	0.56	Fair Agreement
32	2	4	0.66	0.23	0.56	Fair Agreement
33	3	4	0.66	0.23	0.56	Fair Agreement
34	4	4	0.66	0.23	0.56	Fair Agreement
S-CVI/Ave = (sum of I-CVI scores)/ (number of item) = 0.97 **S-CVI/UA** = (sum of UA scores)/ (number of item) = 0.82						

Six professional professors and educationists were invited to participate as expert panelists to validate the test items. Eventually, all six experts agreed to participate and completed the content validation sheet, allowing them to make necessary modifications. The deleting content ratings were used at the initial test development and later to compile content validation (I-CVI) for the final round (Appendix 'K').

These final fifty (50) items were selected as perfect or excellent and spread into six dimensions. Again, all the item's scores were in excellent agreement, with an I-CVI (Cicchetti and Sparrow

1981) score between 0.833 and 1 (Generalization, Induction, Deduction, Logical Thinking, Symbols, and A Mathematical Proof) sub-dimension, respectively.

Second Research Question:

What is The Reliability of the Evidence of Mathematical Thinking Test among Undergraduate Students in the Kingdome of Saudi Arabia?

Four analyses were performed to establish the reliability evidence in IRT: A Cronbach's Alpha for reliability, test statistics, item reliability, and person reliability and separation indices were obtained. The results were presented in Tables 3, and 4.

a. A Cronbach Alpha

A Cronbach Alpha coefficient was evaluated using the guidelines suggested by George and Mallery (2016) where $\alpha \geq 0.9$ Excellent, $0.9 > \alpha \geq 0.8$ Good, $0.8 > \alpha \geq 0.7$ Acceptable, $0.7 > \alpha \geq 0.6$, Questionable, $0.6 > \alpha \geq 0.5$ Poor, and $0.5 > \alpha$ unacceptable as shown in Table 3.

Table 3 Item Analysis for Reliability and Test Statistics

Reliability Statistics	**Cronbach's Alpha**		**Cronbach's Alpha Based on Standardized Items**	**N of Items**
	.95		.95	50
Test Statistics	Mean	Variance	SD	N of Items
	119.45	961.476	31.008	50

Table 3 presents the descriptive statistics for items, including mean, standard deviation, and variance, as summarized in the item statistics. For example, the "Reliability Statistics" presented two different values for Cronbach's Alpha (Kiliç 2016). The first Alpha value is 0.95, which is the raw or unstandardized value of Alpha based on item covariance, measuring the distributions of two variables (Henrysson and Wedman 1972). The second value of Alpha is 0.95, which was treated as the standardized value of Alpha based on item correlation. The stronger the items are interrelated, the greater the test consistency. Therefore, the Alpha selection must be based on a statistical tool, such as covariance or correlation, but not to show the most significant value. The item and item-total statistics were explained in (Appendix C and Appendix D). For this study, a

value of 0.95 for the item correlation was used as evidence of the reliability index for the mathematical thinking test.

b. Person Reliability

Table 4 presented the person reliability, and separation indices obtained from the analysis were 0.95 and 4.25, respectively. The reliability value was significant and excellent, with a value of 0.95. According to Perera, Sumintono et al. (2018), the reliability coefficient should be within the range of 0.91 to 0.94. A reliability value of 0.95 and above indicates excellent reliability, and the separation index was more than 0.2, which was 4.25 (Boone, Staver et al. 2014). Additionally, this implies that the variability in the student's abilities in this study was adequate, as it indicates that people were differentiated (Abedalaziz 2010).

Table 4 Summary of Person Reliability Index Based on RMM

	Total Score	**Count**	**Measure**	**Model Error**	**INFIT**		**OUTFIT**	
					MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
Mean	120.1	50.0	-.13	.16	1.00	.0	1.00	.0
S.D.	30.4	.0	.76	.05	.16	1.0	.19	1.0
Max.	187.0	50.0	1.77	.39	1.42	1.9	1.62	2.3
Min.	56.0	50.0	-2.45	.14	.62	-2.8	.62	-2.6
Real RMSE .17		True SD .74		Separation 4.25		Person Reliability 0.95		
Model RMSE .17		True SD .74		Separation 4.41		Person Reliability 0.95		
S.E. of Person Mean = .06								
Person Raw Score-to-Measure Correlation = .92								
Cronbach Alpha (KR-20) Person Raw Score "Test" Reliability = .95								

c. Item Reliability

To assess item reliability, Table 5 showed that the reliability and separation indices obtained from the analysis were 0.96 and 5.14, respectively. The item reliability and item separation values indicated that the item's reliability in this developed mathematical thinking test was excellent. In addition, the person sample was large enough to confirm the item difficulty hierarchy of the test

items (Abedalaziz 2010). Similarly, the Cronbach Alpha (KR-20) test reliability score of 0.98 further confirmed that the test had adequate internal consistency reliability.

Table 5 Summary of Item Reliability Index Based on RMM

	Total Score	**Count**	**Measure**	**Model Error**	**INFIT**		**OUTFIT**	
					MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
Mean	547.1	229.0	.00	.07	1.00	.0	1.00	-.1
S.D.	75.3	.0	.38	.00	.11	1.3	.15	1.4
Max.	728.0	229.0	.84	.08	1.26	2.7	1.50	3.4
Min.	392.0	229.0	-.95	.07	.75	-3.5	.73	-3.0
Real RMSE .07		True SD .38		Separation 5.14		Item Reliability 0.96		
Model RMSE .07		True SD .38		Separation 5.26		Item Reliability 0.97		
S.E. of Item Mean = .05								
Item Raw Score to Measure Correlation= -1.00								

Third Research Question:

Does the Development of the Mathematical Thinking Test among Undergraduate Students in Kingdom of Saudi Arabia fit into One Parameter (1-PL) Logistics Model?

The analysis of model-data fit was checked on internal validity (Maydeu-Olivares 2014). Within the latent trait test model, the internal validity of test was assessed in terms of the model that best fits the data (Andrade 2018). If the data is compatible with the model, the item is valid. The IRT has three models: one-parameter, two-parameter, and three-parameter models. In IRT analysis (Wise and DeMars 2006) , a likelihood-ratio (Chi-square) test was considered for polytomous test data, and a standardized residual (z Resid) test was considered for dichotomous test data (Mislevy and Bock 1990, Guyer and Thompson 2014).

The item fit was reported in many places in Winsteps 3.73. The output here was generated by the software version 3.73 of Winsteps. Table 6 displayed the point-measure correlation (PTMZ) (Boone and Staver 2020), the Pearson correlation between the individual items, and the measures of the respondents who received those scores (Benesty, Chen et al. 2009). These point-measure correlations should be positive; however, the correlation size was less indicative of fit with the

model than the fit statistics, which average the squared standardized residuals between actual and modeled data.

Table 6 displayed the mean-squared fit statistics (MNSQ) for infit and outfit. These indicated showed how well the item responses fit the model. Infit is more sensitive to unexpected responses to items sufficiently targeted to the distribution of persons. In contrast, the outfit is more susceptible to unexpected responses on items that fall at the extremes. Moreover, it displays the standardized weighted (infit) and unweighted (outfit) mean-squared fit statistics (ZSTD), respectively, which convert the mean-square values into approximate t-statistics that are less sensitive to sample size than MNSQ values. The interpretation of parameter level means square fit statistics was < 2.0, which distorts or degrades the measurement system. Therefore, 1.5-2.0 could be more productive for the construction of measurement but not degrading. Therefore, the range of 0.5-1.5 is productive for measurement. The last one is greater than 0.5, which is less productive for measurement but not degrading and may produce misleadingly good reliabilities and separation. Based on this explanation, these results have no misfit and outfit items because all items were productive for measurement. Again, all the items fell within the range of 0.5-1.5, making them suitable for measurement, and ZSTD values exceeded 2.0 (Wright and Linacre 1994, Smith, Rush et al. 2008, Linacre 2012). Finally, the summary of item fit for the 1PL, as presented in Table 6, revealed that the Infit MNSQ thresholds did not identify any items as misfitting, nor did the outfit MNSQ thresholds. Finally, the first pilot study was isolated for the analysis procedure of the 1PL model.

Table 6 Model Fit Statistics

INFIT		**OUTFIT**		**Model S.E.**	**Measure**	**PT-MEASURE**	**Total Score**	**Item**
MNSQ	**ZSTD**	**MNSQ**	**ZSTD**			**CORR**		
1.13	1.3	1.50	3.3	.08	.84	.43	392	44
1.11	1.4	1.38	3.4	.07	.22	.45	501	3
1.22	2.6	1.28	2.6	.07	-.42	.40	632	39
1.26	2.7	1.25	1.9	.08	.64	.42	424	26
1.17	2.0	1.25	2.4	.07	.16	.39	514	36

INFIT		OUTFIT		Model S.E.	Measure	PT-MEASURE	Total Score	Item
MNSQ	ZSTD	MNSQ	ZSTD			CORR		
1.06	.8	1.19	2.0	.07	-.01	.51	548	5
1.15	1.6	1.19	1.5	.08	-.95	.47	728	19
1.11	1.3	1.15	1.4	.07	-.59	.47	665	20
1.14	1.8	1.07	.7	07	-.14	.50	575	1
1.14	1.8	1.03	.3	07	-.02	.52	551	12
1.13	1.6	1.13	1.3	07	.10	.45	526	30
1.13	1.4	1.02	.2	.08	.58	.48	434	50
1.11	1.5	1.05	.6	.07	-.14	.48	575	6
1.04	.5	1.09	.9	.07	.23	.47	499	16
1.08	1.0	1.00	.1	.07	-.34	.55	615	38
1.07	.9	1.06	.7	.07	.11	.47	524	43
1.02	.2	1.04	.4	.07	.36	.46	474	47
1.04	.5	98	-.2	.07	.07	.51	532	15
1.03	.4	.93	-.7	.07	.44	.50	460	37
1.03	.4	1.00	.1	.07	-.21	.53	590	32
1.03	.4	.95	-.5	.07	.08	.54	529	23
1.02	.3	.98	-.1	.07	-.48	.52	643	31
1.02	.3	1.01	.2	.07	-.26	.51	600	13
.97	-.3	1.02	.2	.07	-.03	.48	552	8
.92	-.9	1.01	.2	.07	.28	.50	490	45
1.00	.1	.92	-.8	.07	.22	.53	501	25
1.00	.0	.94	-.6	.07	-.35	.55	617	14
1.00	.0	.96	-.4	.07	-.03	.50	553	33
.91	-1.2	1.00	.0	.07	-.26	.48	599	7
.99	.0	.94	-.6	.07	-.11	.53	568	2
.90	-1.4	.98	-.1	.07	.11	.52	524	41
.98	-.2	.92	-.8	.07	-.08	.57	563	9
.98	-.2	.92	-.8	.07	.36	.49	475	35
.98	-.2	.92	-.7	.07	.32	.52	482	49
.97	-.3	.98	-.2	.08	-.79	.56	702	22
.98	-.3	.90	-.9	.07	.49	.50	451	27
.95	-.6	.92	-.8	.07	-.09	.51	564	4

INFIT		OUTFIT		Model S.E.	Measure	PT-MEASURE	Total Score	Item
MNSQ	ZSTD	MNSQ	ZSTD			CORR		
.94	-.6	.94	-.6	.07	-.61	.56	669	17
.94	-.8	.90	-1.0	.07	-.32	.56	611	48
.94	-.8	.88	-1.3	.07	.06	.56	534	40
.93	-.8	.91	-.8	.07	.35	.57	477	46
.92	-.9	.91	-.8	.07	-.53	.52	653	21
.89	-1.4	.88	-1.3	.07	-.37	.53	623	24
.89	-1.5	.83	-1.9	.07	-.22	.59	591	18
.85	-1.9	.88	-1.2	.07	.21	.53	503	42
.87	-1.8	.83	-1.9	.07	-.08	.52	562	34
.85	-2.1	.80	-2.2	.07	-.24	.57	596	11
.81	-2.1	.83	-1.4	.08	.73	.51	410	29
.80	-2.4	.75	-2.3	.08	.61	.52	429	28
.75	-3.5	.73	-3.0	.07	.11	.61	523	10
MEAN 1.00	**.0**	**1.00**	**- .1**	**.07**	**.00**		**547.1**	
S.D .11	**1.3**	**.15**	**1.4**	**.00**	**.38**		**75.3**	

Figure 1 shows the item characteristic curves. All the items have the same level of discrimination but differ in difficulty. The lefthand curve represents an easy item because the probability of a correct response is high for low-ability examinees and approaches. The center curve represents an item of medium difficulty because the probability of a correct response is low at the lowest ability levels, around 0.5 at the middle of the ability test, and near one at the highest ability levels. The righthand curve represents a hard item. The probability of correct response is low for most ability tests and increases only when higher ability levels are reached, even at the highest ability level.

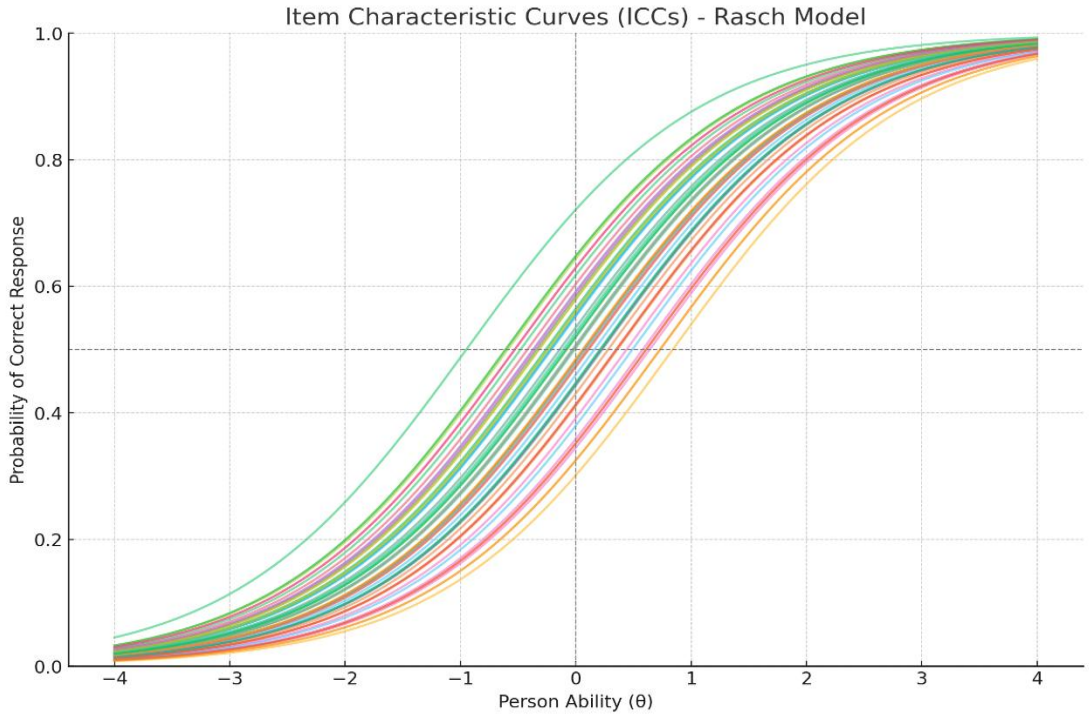

Figure 1 1-PL, Item Characteristic Curves (ICC) for the Items

Fourth Research Question:
Does the Developed Mathematical Thinking Test Among Undergraduate Students in Kingdome of Saudi Arabia for Satisfy the Assumptions of IRT Uni-dimensionality and Local Independence?

Table 7 explains the uni-dimensionality, where detecting uni-dimensionality can be done using Principal Component Analysis PCA (Linacre, Bisaccio et al. 2019). Besides, there were many indicators to decide the uni-dimensionality. First, the variance in the data is explained by the measures (RVEM), which were 36.2%. If the data fit the model perfectly, 36.3% of the variance would be explained (Linacre, Bisaccio et al. 2019). These percentages were close, indicating that the estimation of Rasch measures has been successful. This indicated that test data satisfied the assumption of uni-dimensionality and that the test assessed a unidimensional construct. Second, the variance in the data explained by the difficulties of the items (25.0 %) is larger than the variance explained by a person's abilities (11.2%). Third, Linacre (2015) suggested examining unexplained variance in the first contrast (UVFC) to examine uni-dimensionality. The eigenvalue should not exceed 3, which is the recommended value. Additionally, according to the current results, the first

contrast, 1.8, was lower than the suggested or recommended value, achieving conditional uni-dimensionality and indicating the non-existence of a second dimension (Ishak, Osman et al. 2018, Maiwada, Khan et al. 2019). The results of the (PCA) residuals to check the unidimensional characteristics of the test items (Bro and Smilde 2014) were presented in Table 8.

The result of the PCA, presented in Table 9, showed the unidimensional nature of the test. It further confirmed that the test items measured perceptions of mathematical thinking. Furthermore, the analysis revealed no substantial or meaningful secondary dimension within the data (Abedalaziz and Leng 2018). Thus, the results supported the uni-dimensionality of the test items and strongly confirmed that the test items were related only to the content of the mathematical thinking test.

Table 8 Standardized Residual Variance (in Eigenvalue Units)

	Eigenvalue	**Observed**	**Expected**	
Total raw variance in observations	78.3	100.0%	100.0%	
Raw variance explained by measures	28.3	36.2%	36.3%	
Raw variance explained by persons	8.7	11.2%	11.2%	
Raw variance explained by items	19.6	25.0%	25.1%	
Raw unexplained variance (total)	50.0	63.8%	100.0%	63.7%
Contrasts	**Eigenvalue**	**Observed**	**Expected**	
Unexplained variance in 1st contrast	2.7	3.4%	5.3%	
Unexplained variance in 2nd contrast	2.6	3.3%	5.1%	
Unexplained variance in 3rd contrast	2.4	3.0%	4.7%	
Unexplained variance in 4th contrast	2.3	2.9%	4.5%	
Unexplained variance in 5th contrast	2.2	2.8%	4.3%	

To investigate Local Independence (LI), there should be no correlation between the residuals of a pair of items or between such items. LI assumption should be violated when the residual correlations are high (Mesbah, Kreiner et al. 2013). The standardized residual correlations determined whether the unexplained variance in the responses was random or correlated with other items. The items with the most significant standardized residual correlations were to identify the dependent items in the test (Shi, Maydeu-Olivares et al. 2018). As presented in Table 9, the standardized residual correlation showed a low correlation among the items in the test. Since the standardized residual correlations showed no sign of a strong correlation of at least 0.50 (Shi et al., 2018), it can be concluded that there was no evidence of a violation of the local independence assumption on this test (Vermunt and Magidson 2004).

Table 9 Items with Largest Standardized Residual Correlations Standardized

Correlation	Entry Number	Item	Entry Number	Item
.37	10	Item10	41	Item 41
.28	8	Item8	48	Item 48
.26	5	Item 5	25	Item 25
.26	31	Item 31	40	Item 40
-.36	31	Item 31	32	Item 32
-.30	14	Item 14	41	Item 41
-.29	7	Item 7	50	Item 50
-.28	9	Item 9	20	Item 20
-.28	14	Item 14	16	Item 16
-.27	6	Item 6	17	Item 17

Based on the results of unidimensionality and local independence, there is no evidence to suggest that the data violated any of the conditions or assumptions of IRT, either statistically or substantially. Hence, the test data satisfied the basic assumptions of IRT (Van der Linden and Hambleton 2013).

Fifth Research Question:

Does the Mathematical Thinking Test Develop Among Undergraduate Students in Kingdome of Saudi Arabia Items Free from Bias Based on Demographic Characteristics (gender and university location)?

The DIF analysis detected biased items based on demographic characteristics, gender, and university location using WINSTEPS 3.73 software (Zenisky, Hambleton et al. 2003). The size or magnitude of DIF is referred to as the differential item functioning contrast. DIF (0.5) are considered negligible, with contrasts 0.5 to 1 as moderate and one as substantial, provided that the DIF contrasts are statistically significant at $p < 0.05$ (Bond, Fox et al. 2007). The analysis results were presented in Table 10 for DIF analysis on gender. Due to space limitations, only DIF-flagged items appeared on the gender and university location-based DIF Tables (Appendix 'I'). Additionally, the DIF analysis presented in Table 10 revealed that out of the 50 items, only two (items 43 and 48) exhibited DIF for males and females participating in the study. Therefore, the DIF contrast estimation revealed differential performance by gender, and all contrasts are statistically significant ($p < 0.05$). The method used here was the Mantel-Haenszel approach,

which did not distinguish between uniform and non-uniform DIF. Also, Figure 2 shows the person DIF plot (Gender).

Table 10 Differential Item Functioning Analysis on Gender

	Person	DIF	DIF	Person	DIF	DIF	DIF	Welch		
Item	Class	Measure	S.E.	Class	Measure	S.E.	Contrast	t	d.f.	Prob.
Item 43	M	.26	.10	F	-.03	0.10	.29	2.07	222	.0395
Item 48	M	-.16	.10	F	-.46	0.10	.30	2.14	223	.0336

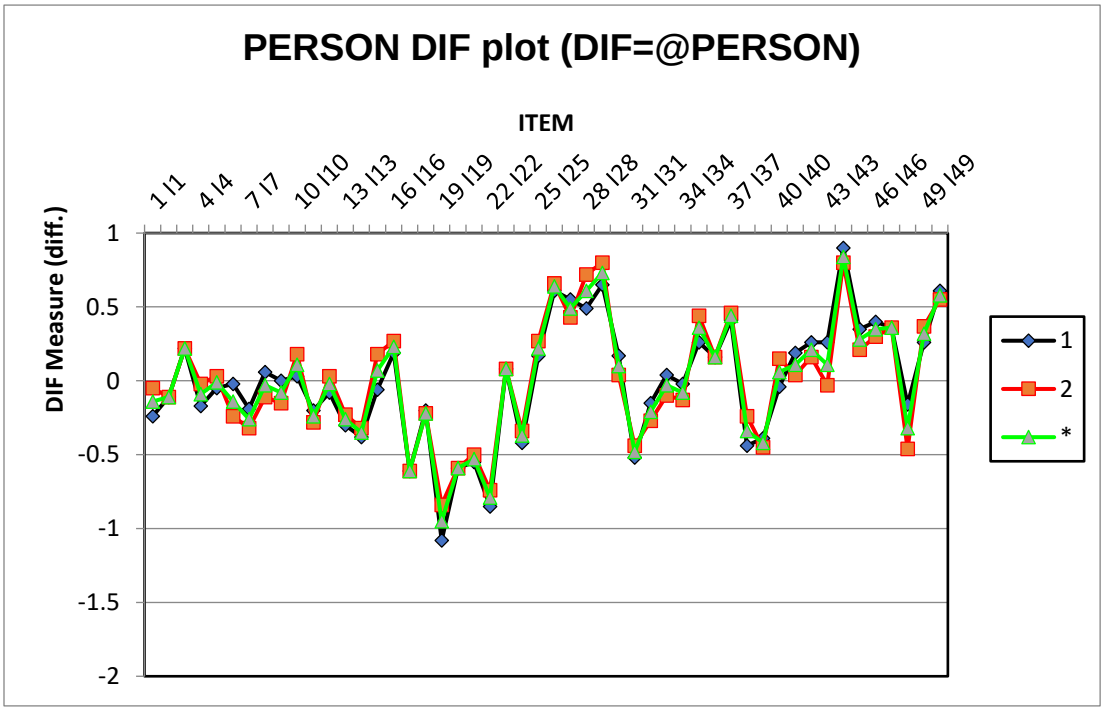

Figure 2 Person DIF Plot (Gender)

As investigated, the DIF analysis between location of the university showed in Table 11. The result showed that out of the 50 items, 4 items (Item 6, Item12, Item33, and Item36) exhibited DIF or was flagged based on DIF. Also, Figure 3 shows the person DIF plot location of the university (Appendix 'J').

Table 11 Location of the University Based Differential Item Functioning

	Person	DIF	DIF	Person	DIF	DIF	DIF	Welch		
Item	Class	Measure	S.E.	Class	Measure	S.E.	Contrast	t	d.f.	Prob.
Item 6	1.00	-.33	.10	2.00	-.14	.14	-.19	-1.07	133	.2854
	1.00	-.33	.10	3.00	.17	.13	-.50	-3.00	148	.0031
Item12	1.00	.18	.10	3.00	-.15	.14	.27	1.53	135	.1276
	1.00	.18	.10	2.00	-.09	.13	.46	2.76	149	.0066
Item33	1.00	-.28	.10	2.00	-.03	.14	-.25	-1.45	133	.1490
	1.00	-.28	.10	3.00	.38	.13	-.66	-3.96	148	.0001
Item36	1.00	.35	.10	2.00	.16	.14	.19	1.08	136	.2835
	1.00	.35	.10	3.00	-.13	.13	.48	2.84	151	.0052

As investigated, the DIF analysis between the locations of the university is shown in Table 11. The results showed that out of the 50 items, four items (Items 6, 12, 33, and 36) exhibited DIF or were flagged as such based on DIF. Also, Figure 3 shows the personal DIF plot location of the university.

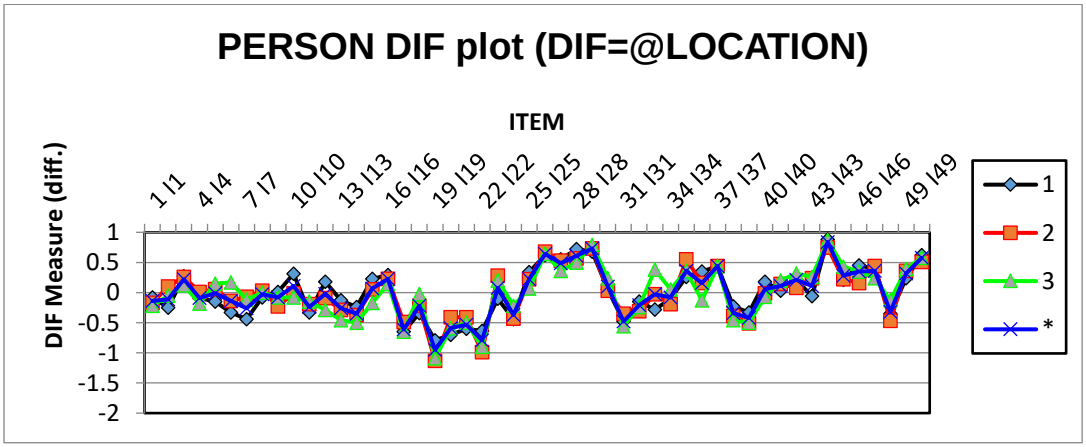

Figure 3 Person DIF Plot Location of the University

Discussions of Findings:

The results for the first study question showed that the MT test, comprising 50 items across six sub-dimensions (Generalization, Induction, Deduction, Use of Symbols, Logical Thinking, Mathematical Proof), underwent rigorous expert review by six lecturers, professors, and educationists. Initially, 60 items were proposed; 10 were removed during preliminary reviews. Experts rated items for relevance and suggested modifications, with ratings compiled across multiple rounds to compute the Item-Content Validity Index (I-CVI) and the Scale-Content Validity Index (S-CVI). Content validity was established using I-CVI and S-CVI, yielding S-CVI/Ave = 0.97 and S-CVI/UA = 0.82—both exceeding recommended thresholds (>0.78; Polit et al., 2007). Inter-rater agreement was assessed via Fleiss' kappa, confirming high item relevance. These indices align with prior studies employing identical methods (Delgado-Rico et al., 2012; Perles-Ribes et al., 2017). Internal consistency was supported by item and person reliability indices. The process adhered to established instrument development protocols: construct identification, item construction, and expert judgment (Messick, 1989). Results affirm that MT items are (i) relevant for measuring mathematical thinking in undergraduate students from Saudi Arabia's Western zone; (ii) comprehensive across all constructs; (iii) sufficient for study objectives and research questions; and (iv) technically sound in clarity, wording, and grammar. While Fleiss' kappa is widely accepted, potential underestimation with few raters or low-prevalence items was noted but deemed minimal here, given high kappa values and six raters (Janssen-Brandt et al., 2017). Content validity alone does not encompass Messick's (1989) multifaceted framework; future studies should explore construct, criterion-related, and consequential validity. Complementary methods, such as cognitive interviews or pilot testing, could enhance rigor. Overall, the MT test demonstrates excellent content validity and reliability, supporting its application in educational assessment.

The results corresponding to the second research question established that the reliability of the 50-item MT test was evaluated using the Rasch Measurement Model (RMM) via Winsteps analysis, reporting Cronbach's Alpha (KR-20), person reliability/separation, and item reliability/separation.

Cronbach's Alpha (KR-20) was 0.95, indicating excellent internal consistency. Person reliability (0.95) and separation (4.25) exceeded thresholds (reliability ≥ 0.80; separation > 2.0; Linacre, 2012; Bond et al., 2007), confirming that items effectively distinguished at least two ability levels among students. Item reliability (0.96) and separation (5.14) were also excellent, verifying a stable item difficulty hierarchy and representativeness with a sample of 229 respondents (Abedalaziz, 2010). These indices align with prior studies using similar Rasch-based criteria (Kaur et al., 2012; Saidfudin et al., 2010). Person reliability reflects the range of sample abilities and test length, while item reliability depends on the spread of item difficulties and sample size (Linacre, 2012). High separation indices indicate no redundancy or ability-difficulty gaps, supporting precise measurement. Although some literature permits retaining theoretically critical items with marginal fit (DeVellis, 2017; Boateng et al., 2018), this study adopted a conservative approach, removing 10 misfitting or misaligned items post-pilot to prioritize statistical rigor. The results affirm that the MT test, with 50 items and 229 respondents, yields stable, reliable inferences for assessing mathematical thinking among undergraduate students in Saudi Arabia's Western zone.

The outcomes addressing the third study question confirmed that the internal validity and model-data fit were assessed using the Rasch Measurement Model (RMM; Maydeu-Olivares, 2014; Andrade, 2018). Fit statistics followed Linacre's (2007) sequential protocol: point-measure correlation (PMC; must be positive), outfit mean square (MNSQ), infit MNSQ, outfit Z-standardized (ZSTD), and infit ZSTD. Acceptable ranges were MNSQ 0.50–1.50 and ZSTD −2.00 to +2.00 (Linacre, 2002; Bond et al., 2007). If MNSQ was acceptable, ZSTD deviations were overlooked (Wright & Masters, 1982; Bond & Fox, 2015). Pilot analysis of 60 items identified misfit: one item exceeded MNSQ limits; 10 others (items 5, 12, 14, 16, 26, 28, 29, 32, 33, 34) showed marginal ZSTD (2.4, 2.6, −3.4). Following conservative criteria (Salleh & Ab Aziz, 2014; Iramaneerat et al., 2008; Schumacher et al., 2004), four items were deleted for statistical misfit. An expert panel of six lecturers and professors then reviewed content, leading to the deletion of 10 additional items for poor alignment or redundancy. Final 50 items—distributed across six dimensions (Generalization, Induction, Deduction, Use of Symbols, Logical Thinking, Mathematical Proof)—achieved I-CVI 0.82–1.00, indicating excellent content validity (Cicchetti & Sparrow, 1981; Polit et al., 2007; Lynn, 1986). This rigorous, dual-criteria approach (statistical fit + expert judgment) exceeds recommendations in the literature that prioritize MNSQ over ZSTD

or retain theoretically vital items despite marginal fit (DeVellis, 2017; Boateng et al., 2018; Bond & Fox, 2015). The final 50-item MT test demonstrates strong unidimensionality, item representativeness, and construct coverage, supporting the validity of the measurement of mathematical thinking among undergraduate students in Saudi Arabia's Western zone.

The results addressing the fourth research question revealed that Unidimensionality was assessed via Rasch principal component analysis (PCA) of residuals, item fit statistics (PMC, outfit/infit MNSQ), and reliability (Bond et al., 2007; Sick et al., 2010). The first contrast eigenvalue was 1.8 (< 3.0 threshold), indicating no meaningful secondary dimension (Ishak et al., 2018; Maiwada et al., 2019; Abedalaziz & Leng, 2018). High reliability (0.95) supported but did not confirm unidimensionality (Ayearst & Bagby, 2010). Local independence (LI) was evaluated through standardized residual correlations. The highest was 0.32 (< 0.50), showing no significant item dependency (Shi et al., 2018; Mesbah et al., 2013). Responses to one item did not cue another, satisfying LI (Linacre et al., 2019). These results align with prior Rasch applications, confirming unidimensional IRT models when first-contrast eigenvalues are low and residual correlations minimal (Guyer & Thompson, 2014b). However, literature notes potential multidimensionality in broad or complex constructs (Reise et al., 2003; Baghaei, 2008) and LI violations with residual correlations > 0.20–0.30 (Yen, 1993; Christensen et al., 2017), which may inflate reliability and bias estimates (Marais & Andrich, 2008). Despite covering six sub-domains, the MT test exhibited strong unidimensionality and LI, validating the use of the 1PL Rasch model. Future work could explore multidimensional models for enhanced precision.

Analysis of the fifth research question yielded evidence that DIF was examined across gender and university location using Rasch analysis, with DIF contrast thresholds of <0.50 (negligible), 0.50–1.00 (moderate), and >1.00 (substantial) at $p < .05$ (Bond et al., 2007). Of 50 items, only two (Items 43, 48) showed significant gender DIF, and four (Items 6, 12, 33, 36) exhibited location DIF. Thus, 88% of items functioned in an invariant manner, indicating minimal bias. Low DIF prevalence aligns with studies reporting <10% biased items in well-designed instruments (Bond et al., 2007; Smith & Rush, 2010; Lee et al., 2015). In contrast, higher DIF rates (>20%) have been observed in gender or location comparisons (Garcia & O'Neill, 2013; Khan et al., 2018), possibly due to less rigorous item development, population heterogeneity, or differing DIF criteria (Zumbo, 1999). The MT test's low DIF reflects effective item design, expert validation, and pilot refinement,

supporting fair measurement across gender and university location among undergraduate students in Saudi Arabia's Western zone.

Recommendations:

Based on the findings of this study and considering the significant place of MT in our educational system, the study made the following recommendations:

Expand expert panels to include diverse educational backgrounds and conduct periodic item reviews to maintain alignment with evolving standards. Provide training for instructors on item development to standardize content validation.

1. Reassess reliability with larger, multi-institutional samples and curriculum updates. Offer workshops on IRT and Rasch analysis to strengthen educators' assessment skills.
2. Implement the MT test in undergraduate settings; explore 2PL/3PL models in future adaptations to examine discrimination and guessing. Use this rigorous IRT-based process as a blueprint for other subject tests.
3. Regularly verify unidimensionality and local independence in new cohorts or contexts. Extend to multidimensional models only if assumptions remain satisfied.
4. Revise DIF-flagged items (e.g., 43, 48 for gender; 6, 12, 33, 36 for location), conduct ongoing DIF monitoring, and incorporate diverse pilot groups to minimize bias early.
5. Design targeted interventions for low performers, enrichment for high achievers, and MT-focused instruction for the average group. Use periodic Rasch-based assessments to track progress.
6. Prioritize curriculum emphasis on challenging sub-domains (Induction, Proof, Symbols) via scaffolded practice. Leverage item difficulty hierarchy for balanced assessments; ensure gender-inclusive interventions. Investigate instructional factors influencing sub-domain performance.

Conclusion:

This research successfully developed a psychometrically robust 50-item MT test tailored for undergraduate students in the western zone of the KSA, using the Rasch one-parameter logistic model to overcome the limitations of prior Classical Test Theory-based instruments. Expert validation ensured high content relevance (I-CVI ≥ 0.833; S-CVI/Ave = 0.97), while IRT analyses demonstrated excellent reliability (Cronbach's alpha = 0.95; person/item reliabilities ≥ 0.95), strong model-data fit, unidimensionality, local independence, and minimal DIF (88% items invariant across gender and location). These findings affirm the instrument's validity and fairness

in measuring MT across six key dimensions, enabling precise identification of ability levels and item difficulties. By aligning with Vision 2030's emphasis on cognitive skill development, the tool supports targeted interventions in teacher education programs, such as scaffolded instruction in challenging sub-domains (Generalization, Induction, Deduction, Use of Symbols, Logical Thinking, and Mathematical Proof) and bias mitigation through item revision. Future applications should involve larger multi-regional samples, exploration of multidimensional IRT models, and integration into ongoing assessments to foster critical thinking, problem-solving, and educational reform. Ultimately, this instrument equips educators and policymakers with a reliable means to elevate MT proficiency, preparing future teachers to cultivate these essential skills in KSA's evolving educational landscape.

References

THE DEVELOPMENT AND VALIDATION OF AN INSTRUMENT TO MEASURE VALUES IN MATHEMATICS CLASSROOMS OF MATRICULATION LECTURERS RUZELA TAPSIR THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY FACULTY OF EDUCATION.

Abedalaziz, N. (2010). "A gender-related differential item functioning of mathematics test items." The International Journal of Educational and Psychological Assessment 5(2): 101-116.

Abedalaziz, N. and C. H. Leng (2018). "The relationship between CTT and IRT approaches in Analyzing Item Characteristics." MOJES: Malaysian Online Journal of Educational Sciences 1(1): 64-70.

Andrade, C. (2018). "Internal, external, and ecological validity in research design, conduct, and evaluation." Indian journal of psychological medicine 40(5): 498-499.

Arfat, Y., et al. (2025). "Building inclusive learning environment through hybrid learning system: Role of technology and corresponding engagement." Sustainable Futures 10: 100887.

Benesty, J., et al. (2009). Pearson correlation coefficient. Noise reduction in speech processing, Springer: 1-4.

Bond, T. G., et al. (2007). Applying the Rasch model: Fundamental measurement. in the social sciences (2nd, Citeseer.

Boone, W. J. and J. R. Staver (2020). Point Measure Correlation. Advances in Rasch Analyses in the Human Sciences, Springer: 25-38.

Boone, W. J., et al. (2014). Person Reliability, item reliability, and more. Rasch analysis in the human sciences, Springer: 217-234.

Bro, R. and A. K. Smilde (2014). "Principal component analysis." Analytical methods 6(9): 2812-2831.

Cicchetti, D. V. and S. A. Sparrow (1981). "Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior." American journal of mental deficiency.

George, D. and P. Mallery (2016). Frequencies. IBM SPSS Statistics 23 Step by Step, Routledge: 115-125.

Guyer, R. and N. Thompson (2014). "Xcalibre Item Response Theory Calibration Software User Manual." Assessment System Corporation, Woodbury MN.

Henrysson, S. and I. Wedman (1972). "Analysis of the Inter-Item-Covariance Matrix." Scandinavian Journal of Educational Research 16(1): 25-35.

Ishak, A. H., et al. (2018). "Examining unidimensionality of psychometric properties via rasch model." International Journal of Civil Engineering and Technology (IJCIET) 9(9): 1462-1467.

Kiliç, S. (2016). "Cronbach's alpha reliability coefficient." Psychiatry and Behavioral Sciences 6(1): 47.

Linacre, J. (2012). "Winsteps® Rasch measurement computer program user's guide." Beaverton, Oregon: Winsteps. com.

Linacre, J. M. (2015). "The sin of false precision: Too many rating-scale categories." Rasch Measurement Transactions 28(2): 1463.

Linacre, S., et al. (2019). "Publishing in an environment of predation: The many things you really wanted to know, but did not know how to ask." Journal of Business-to-Business Marketing 26(2): 217-228.

Maiwada, R. M., et al. (2019). "Validation of Anxiety, Life Satisfaction and Social Media Addiction Scales using Rasch Measurement Approach." Indian Journal of Public Health Research & Development 10(9).

Maydeu-Olivares, A. (2014). Evaluating the fit of IRT models. Handbook of item response theory modeling, Routledge: 129-145.

Mesbah, M., et al. (2013). Rasch models in health, John Wiley & Sons Incorporated.

Mislevy, R. J. and R. D. Bock (1990). BILOG 3: Item analysis and test scoring with binary logistic models, Scientific Software Mooresville, IN.

Perera, C. J., et al. (2018). "The psychometric validation of the principal practices questionnaire based on item response theory." International Online Journal of Educational Leadership 2(1): 21-38.

Shi, et al. (2018). "The relationship between the standardized root mean square residual and model misspecification in factor analysis models." Multivariate Behavioral Research 53(5): 676-694.

Smith, A. B., et al. (2008). "Rasch fit statistics and sample size considerations for polytomous data." BMC Medical Research Methodology 8(1): 1-11.

Van der Linden, W. J. and R. K. Hambleton (2013). Handbook of modern item response theory, Springer Science & Business Media.

Vermunt, J. K. and J. Magidson (2004). "Local independence." Encyclopedia of social sciences research methods: 732-733.

Wise, S. L. and C. E. DeMars (2006). "An application of item response time: The effort-moderated IRT model." Journal of Educational Measurement 43(1): 19-38.

Wright, B. D. and J. M. Linacre (1994). "The Rasch model as a foundation for the Lexile Framework." Unpublished manuscript.

Yusoff, M. S. B. (2019). "ABC of content validation and content validity index calculation." Resource 11(2): 49-54.

Zenisky, A. L., et al. (2003). "Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach." Educational and Psychological Measurement 63(1): 51-64.