Jurnal Gantang IX (2) (2024): 191-204

JURNAL GANTANG

e-ISSN: 2548-5547
p-ISSN: 2503-0671

http://ojs.umrah.ac.id/index.php/gantang/index

Rasch model analysis: Evaluating students' spatial thinking ability in higher-order thinking skills trigonometric comparison assessment (HOTS-TCA)

Dwi Rismi Ocy*, Wardani Rahayu, Riyan Arthur

Universitas Negeri Jakarta, Jakarta, DKI Jakarta, 13220, Indonesia
*Correspondent Author: dwirismiocy@gmail.com

Submission: November 26th, 2024; Accepted: December 24th, 2024; Published: December 31st, 2024
DOI: https://doi.org/10.31629/jg.v9i2.7386

Abstract

This study addresses the critical issue of insufficient spatial thinking abilities among students, which significantly affects their performance in higher-order thinking skills (HOTS) tasks, particularly in trigonometry. Focusing on 11th-grade students in Tanjungpinang, Indonesia, the research investigates how spatial thinking influences the ability to solve trigonometric comparison problems. Employing a mixed-method approach, the study integrates quantitative data from the Higher-Order Thinking Skills in Trigonometric Comparisons Assessment (HOTS-TCA) with qualitative insights from post-test interviews. Rasch Model Analysis evaluates the quality of an assessment instrument by providing measures of respondent abilities and item difficulties, fit statistics to ensure model alignment, reliability, and separation indices for consistency, a Wright Map for visualizing the relationship between skills and difficulties, and checks for unidimensionality and potential item bias to ensure fairness and validity. The Rasch analysis further confirms the reliability and validity of the HOTS-TCA instrument, highlighting its effectiveness in measuring spatial thinking across varying ability levels. The study finds that students with high spatial ability excel in visualizing geometric relationships but struggle with complex three-dimensional tasks. In contrast, medium-ability students have difficulties with mental manipulation and real-world applications, and low-ability students face significant challenges in basic visualization and interpreting geometric structures, leading to frequent misconceptions.

Keywords: Higher-order thinking skills (HOTS); rasch model analysis; spatial thinking; trigonometric comparison; assessment validity

How to cite: Ocy, D. R., Rahayu, W., & Arhur, R. Rasch model analysis: Evaluating students' spatial thinking ability in higher-order thinking skills trigonometric comparison assessment (HOTS-TCA). *Jurnal Gantang, 9*(2), 191 – 204. https://doi.org/10.31629/jg.v9i2.7386

I. Introduction

Spatial thinking is critical in developing pupils' cognitive skills, particularly in mathematics, where perceiving and manipulating geometric relationships is required. Research has consistently demonstrated that spatial thinking is closely linked to mathematical achievement, with students who excel in spatial thinking often outperforming their peers in complex tasks such as geometry and trigonometry (Lakin & Wai,

2020; Uttal & Cohen, 2012). Spatial thinking encompasses the ability to mentally represent, manipulate, and interpret the relationships between objects in both two-dimensional (2D) and three-dimensional (3D) spaces, making it essential for solving higher-order mathematical problems (Gilligan, 2020; Palupi et al., 2023). In trigonometry, these skills are particularly significant as students are required to visualize, analyze, and conceptualize intricate geometric relationships between angles, sides, and trigonometric functions (Ngu & Phan, 2020).

Despite the recognized importance of spatial thinking ability, many students, especially in Indonesia, struggle to develop and apply these skills effectively in solving complex mathematical problems, including higher-order thinking skills (HOTS) tasks in trigonometry (Ocy et al., 2023a). Studies have shown that Indonesian students face significant challenges in visualizing geometric relationships and translating spatial representations into mathematical solutions (Nanmumpuni & Retnawati, 2021; Riastuti et al., 2017). These difficulties are particularly evident in trigonometric comparison problems, which demand advanced spatial thinking and the ability to mentally manipulate and compare geometric figures and relationships. Classroom observations and empirical studies highlight a persistent gap in students' ability to engage with and solve such problems, often due to insufficient exposure to spatial thinking strategies and tools during instruction (Anggraini & Putra, 2020; Rohimah & Prabawanto, 2020).

Several studies have explored the relationship between spatial thinking abilities and mathematical problem-solving, as well as the application of the Rasch model in educational research. Resnick et al. (2020) and Adams et al. (2023) highlighted a strong correlation between spatial skills and success in solving geometric and trigonometric problems, emphasizing the need for teaching strategies that enhance spatial visualization. Similarly, Battista et al. (2018) and Sorby et al. (2022) identified spatial visualization as a key predictor of success in higher-order thinking tasks involving geometry. The use of the Rasch model has also been well-documented; Bond & Fox (2013) demonstrated its effectiveness in analyzing cognitive skill assessments, while Boone et al. (2014) showed its utility in identifying misfit items and ensuring unidimensionality in tests measuring HOTS in science and mathematics. Recent studies, such as Ma et al. (2024) and Xie et al. (2020), further emphasize the importance of spatial reasoning in solving complex trigonometric problems, highlighting the Rasch model's capacity to evaluate the validity and reliability of instruments. Building on these findings, applying the Rasch model provides a robust framework for further understanding and enhancing students' spatial thinking abilities in mathematical problem-solving. By employing this model, researchers can uncover the underlying cognitive factors influencing students' performance and pinpoint specific areas that require targeted interventions. Additionally, the Rasch model facilitates the development of diagnostic tools that not only evaluate students' abilities with greater precision but also assist educators in designing effective strategies to strengthen spatial reasoning and improve overall mathematical proficiency.

This study aims to achieve three primary objectives by analyzing students' spatial thinking abilities in solving higher-order thinking skills (HOTS) trigonometric comparison problems. First, it seeks to provide a detailed understanding of the specific spatial thinking abilities students employ during these tasks. Second, it aims to identify key areas where students encounter difficulties, offering insights that can inform the development of targeted instructional interventions. Finally, this research intends to validate the HOTS Trigonometric Comparison Assessment (HOTS-TCA) instrument as a reliable and effective tool for evaluating spatial thinking within mathematical problem-solving contexts. The study aspires to contribute to theoretical advancements in mathematics

education and the practical design of pedagogical strategies that enhance students' trigonometric competence by addressing these objectives.

II. Research Method

This study employs a mixed-method research approach with an explanatory design to investigate students' spatial thinking abilities in solving Higher-Order Thinking Skills (HOTS) problems related to trigonometric comparisons. The research methodology integrates test administration, data analysis using the Rasch model, and interpretation of findings to achieve its objectives. Participants were 11th-grade senior high school students from Tanjungpinang who had completed basic trigonometry topics in their mathematics curriculum. A purposive sampling technique was used to select 515 students from various private and public schools in Tanjungpinang to ensure diversity in academic achievement.

The primary instrument developed for this study was the Higher-Order Thinking Skills in Trigonometric Comparisons Assessment (HOTS-TCA). The assessment consisted of ten carefully designed problems measuring students' ability to visualize and conceptualize geometric relationships, manipulate trigonometric figures, and solve trigonometric comparison problems involving angles and sides. These problems aligned with Bloom's taxonomy (C4-Analyzing; C5-Evaluating; C6-Creating) (Anderson et al., 2001) and emphasized the assessment of spatial thinking abilities in trigonometric comparisons. The reliability and validity of the instrument were tested using Item Response Theory through Rasch Model Analysis.

Data collection occurred in two stages. First, the HOTS-TCA was administered in a controlled environment to minimize external influences. Students were allotted 90 minutes to complete the test, and all responses were collected systematically for analysis. Second, post-test interviews were conducted with a subset of 20 students to obtain qualitative insights into their problem-solving strategies and challenges related to spatial thinking. These interviews provided a deeper understanding of the spatial thinking processes used to solve the assessment tasks.

The data were analyzed using the Rasch measurement model, which enabled a detailed examination of students' abilities, the difficulty levels of the test items (Ocy et al., 2023b; Riani Siregar et al., 2021), and the overall validity of the assessment. This analysis was conducted using the WINSTEPS software, which provided robust insights into the instrument's alignment with the study objectives. The integration of quantitative and qualitative data strengthened the conclusions, offering a comprehensive perspective on the spatial thinking abilities of students in solving HOTS problems related to trigonometry.

III. Results and Discussion

Assumptions of the Rasch Model Analysis

Meeting both assumptions of unidimensionality and local independence is a prerequisite for performing a valid Rasch model analysis and ensures the instrument measures a single, dominant latent construct.

1. Unidimensionality

Unidimensionality is a fundamental assumption of the Rasch model, requiring that the assessment instrument measures a single latent trait or construct. This assumption is typically evaluated through a principal components analysis (PCA) of the residuals. According to the recommended criteria, the variance explained by the first principal component should be at least 20%, while the unexplained variance in the first contrast should be less than 15% (Everitt & Dunn, 2001; Rios, 2013).

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Empirical --		Modeled
Total raw variance in observations	=	17.6	100.0%	100.0%
Raw variance explained by measures	=	7.6	43.0%	44.0%
Raw variance explained by persons	=	3.6	20.5%	21.0%
Raw Variance explained by items	=	3.9	22.5%	23.0%
Raw unexplained variance (total)	=	10.0	57.0% 100.0%	56.0%
Unexplned variance in 1st contrast	=	1.5	8.7% 15.3%	
Unexplned variance in 2nd contrast	=	1.3	7.3% 12.8%	
Unexplned variance in 3rd contrast	=	1.2	7.0% 12.3%	
Unexplned variance in 4th contrast	=	1.1	6.5% 11.3%	
Unexplned variance in 5th contrast	=	1.1	6.5% 11.3%	

Figure 1. Output standardized residual variance from PCA

The variance analysis in Figure 1 shows that the total raw variance in the observations is 17.6, which is the total variance that can be analyzed from the data. The variance explained by the measure is 7.6 with 43% (>20%), indicating that most of the variance can be explained by the measured construct. The variance explained by persons is 3.6, and by items is 3.9, suggesting that individual differences and item differences contribute to the variance but do not indicate the presence of different dimensions.

The HOTS-TCA instrument's eigenvalue values for unexplained variance indicate a unidimensional structure. Specifically, the eigenvalues for the first to fifth contrasts (1.5, 1.3, 1.2, 1.1, and 1.1, respectively) reveal that a significant portion of the total variance remains unexplained, with the first contrast accounting for 8.7% and subsequent contrasts showing diminishing contributions. Despite the presence of unexplained variance, the majority of the explained variance (43.0% by measures and 20.5% by persons) suggests that the instrument effectively captures a primary construct related to higher-order thinking skills. The consistent eigenvalue values across contrasts further reinforce the reliability of the HOTS-TCA instrument in measuring spatial thinking abilities, supporting its classification as a unidimensional assessment tool.

2. Local Independence

This assumption ensures that the measurement focuses solely on the intended latent trait without interference from inter-item dependencies. Standardized residual correlations provided by Winsteps software are analyzed to assess local independence. As Christensen et al. (2017) recommended, these correlations should have absolute values below 0.20.

ITEM	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q1	1	-0.1975	-0.0252	-0.0543	-0.1928	-0.1855	-0.0827	-0.0863	-0.0581	-0.0433
Q2	-0.1975	1	-0.1111	-0.0188	-0.1442	-0.2111	-0.1043	-0.0181	0.0528	-0.2054
Q3	-0.0252	-0.1111	1	-0.097	-0.0005	-0.1069	-0.156	-0.152	-0.1861	-0.0014
Q4	-0.0543	-0.0188	-0.097	1	-0.0723	-0.0776	-0.166	-0.1277	-0.0843	-0.2038
Q5	-0.1928	-0.1442	-0.0005	-0.0723	1	-0.1885	-0.0649	-0.1222	-0.1897	0.0092
Q6	-0.1855	-0.2111	-0.1069	-0.0776	-0.1885	1	-0.0076	-0.0961	-0.0851	-0.0155
Q7	-0.0827	-0.1043	-0.156	-0.166	-0.0649	-0.0076	1	-0.1341	-0.2055	-0.106
Q8	-0.0863	-0.0181	-0.152	-0.1277	-0.1222	-0.0961	-0.1341	1	-0.0281	-0.2067
Q9	-0.0581	0.0528	-0.1861	-0.0843	-0.1897	-0.0851	-0.2055	-0.0281	1	-0.1555
Q10	-0.0433	-0.2054	-0.0014	-0.2038	0.0092	-0.0155	-0.106	-0.2067	-0.1555	1

Figure 2. Item residual correlations

Residual correlations in Figure 2 represent the difference between the empirical and model-predicted correlations. Significant residual correlations indicate a violation of the local independence assumption, while small values suggest fulfilling this assumption. Based on the presented residual correlation data, the residual correlations range from -0.2111 to 0.0528. Residual correlations should be close to zero. If the values exceed a certain threshold, usually around ±0.2, it may indicate a violation of the local independence assumption.

In this data, some residual correlation values approach or slightly exceed the threshold, such as the residual correlations between Q2 and Q6 (-0.2111), Q2 and Q10 (-0.2054), and Q7 and Q9 (-0.2055). This suggests potential local dependence among these item pairs. If some item pairs have significant residual correlations, there may be undesirable relationships between those items, such as content redundancy, similar response patterns, or context effects influencing the responses. However, the degree to which the observed residual correlations exceed the threshold is relatively minor, and the overall values remain within an acceptable range. Therefore, the potential local dependence issues can be negligible and do not significantly compromise the validity of the HOTS-TCA instrument.

Rasch Model Analysis Outcomes

Rasch Model Analysis provides essential outcomes for assessing the quality of an instrument, including person and item measures to evaluate respondent abilities and item difficulty levels, as well as fit statistics (infit and outfit) to ensure alignment with the model (Sumintono et al., 2015). The analysis generates reliability and separation indices to assess consistency and discriminative power, while the Wright Map visualizes the alignment between respondent abilities and item difficulties (Linacre, 2019; Wright, 1994). It verifies unidimensionality to confirm that the instrument measures a single construct and identifies

potential item bias through Differential Item Functioning (DIF) analysis, ensuring fairness and validity (Yang & Kao, 2014).

1. Item Fit Order (Infit and Outfit)

The fit statistics provide insights into how well each item in the HOTS-TCA instrument aligns with the expected model. The INFIT MNSQ measures how well an item functions within the context of individuals with similar ability levels, and values close to 1 indicate a good fit (Huei et al., 2020; Linacre, 2002, 2024; Wang & Chen, 2005). The OUTFIT MNSQ measures how well an item functions overall, including individuals with very high or low abilities, and values near one also suggest a good fit (Linacre, 2002; Sumintono, 2018; Wright, 1994).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
9	98	515	1.93	.13	.93	-1.0	.63	-1.9	.53	.48	83.7	83.4	Q9
10	107	515	1.77	.13	.90	-1.6	.60	-2.2	.56	.49	84.3	82.7	Q10
2	170	515	.85	.11	.98	-.4	.77	-2.0	.58	.56	76.3	78.3	Q2
1	174	515	.80	.11	.98	-.4	.77	-2.0	.59	.56	75.5	78.1	Q1
6	174	515	.80	.11	.97	-.5	.82	-1.6	.58	.56	78.3	78.1	Q6
5	186	515	.65	.11	.94	-1.1	.79	-2.0	.60	.57	77.5	77.4	Q5
8	357	515	-1.47	.12	1.14	2.4	1.20	1.4	.48	.55	76.9	78.9	Q8
7	359	515	-1.50	.12	1.23	3.8	1.47	3.0	.43	.55	76.1	79.1	Q7
4	381	515	-1.81	.12	1.05	.9	.90	-.6	.51	.53	79.3	80.8	Q4
3	394	515	-2.01	.13	1.03	.5	.92	-.4	.51	.52	82.5	81.7	Q3
MEAN	240.0	515.0	.00	.12	1.01	.2	.89	-.8			79.1	79.8	
S.D.	112.2	.0	1.45	.01	.10	1.6	.25	1.7			3.1	2.0	

Figure 3. Fit statistics

Based on Figure 3, represented by the INFIT and OUTFIT column, the item fit order analysis, items Q4, and Q3 have the best fit, with INFIT MNSQ values of 1.05 and 1.03, respectively, and OUTFIT MNSQ values of 0.90 and 0.92, indicating they function well both within the context of similar ability levels and overall. Items Q1, Q2, Q6, Q5, and Q9 are also considered well-fitting, with INFIT MNSQ values close to 1 and OUTFIT MNSQ values near or below 1. In contrast, items Q8 and Q7 have higher INFIT and OUTFIT MNSQ values, suggesting potential fit issues. Items Q9 and Q10 have the lowest OUTFIT MNSQ values, which may indicate over-fit or issues with unexpected responses from individuals with very high or low abilities.

However, items Q8 and Q7 have higher INFIT and OUTFIT MNSQ values, which could suggest they are well-aligned with the expected model rather than indicating potential fit issues. Conversely, the low OUTFIT MNSQ values for items Q9 and Q10 may point to these items being overly predictable or failing to adequately capture responses from individuals at the extreme ends of the ability spectrum. Overall, the item fit order analysis suggests that the HOTS-TCA instrument largely adheres to the Rasch model expectations, with a few items potentially requiring further investigation or refinement to optimize the instrument's psychometric properties (Boone, 2016; Karabatsos, 2003; Wright, 1994).

2. Item Measure (Difficulty Estimates)

The item difficulty levels for the HOTS-TCA instrument, as shown in Figure 3, are represented by the MEASURE column, which uses a logit scale. Higher MEASURE values indicate more complicated items, while lower values correspond to easier items. The difficulty levels of the items range from -2.01 logits to 1.93 logits, covering approximately 4.0 logits. The most challenging item is Q9 (1.93 logits), while the easiest is Q3 (-2.01 logits). This range suggests that the instrument includes items of varying difficulty, which is beneficial for assessing participants with a wide range of abilities.

Difficult items, such as items Q9 (1.93 logits) and Q10 (1.77 logits), are the most difficult in the instrument. These items are likely correctly answered only by participants with high ability levels, as they fall at the upper end of the difficulty scale. Moderately complex items**,** including items such as Q1 (0.80 logits), Q2 (0.85 logits), Q5 (0.65 logits), and Q6 (0.80 logits) are closer to the average difficulty level (MEASURE ≈ 0 logits). These items are appropriate for participants with medium ability levels and provide a good balance for the instrument. Easy items, such as items Q8 (-1.47 logits), Q7 (-1.50 logits), Q4 (-1.81 logits), and Q3 (-2.01 logits) cater to lower-ability participants and are expected to be answered

correctly with minimal difficulty, supporting broader accessibility.

In Rasch Model Analysis, item measures can be examined through Item Characteristic Curves (ICCs).

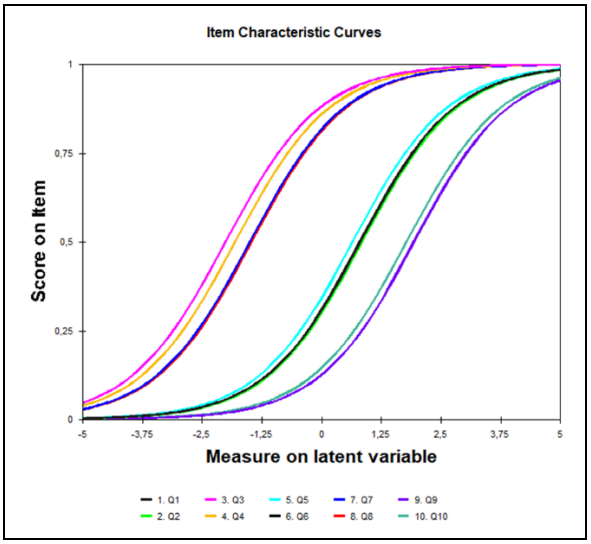


Figure 4. Item characteristic curves (ICCs)

The ICCs, as shown in Figure 4, visually represent the probability of a correct response to an item based on the respondent's ability level. In the ICCs graph, the curves moving further to the left along the x-axis indicate that the corresponding items are easier. This means that individuals with lower levels of the latent trait (the underlying ability or skill being measured) have a higher probability of answering these items correctly. Conversely, items positioned further to the right require a higher level of the latent trait for respondents to achieve a higher probability of a correct response, indicating more incredible difficulty.

This distribution of item difficulties indicates that the HOTS-TCA instrument includes a well-spread range of items, catering to participants across varying ability levels. However, further refinement may be required to ensure optimal coverage of the ability continuum, mainly if the target population exhibits specific ability clusters.

3. Item Separation and Reliability

Item separation measures the instrument's precision in differentiating item difficulty levels (Boone et al., 2014; Fisher, 2024; Sumintono & Widhiarso, 2013). On the other hand, reliability refers to the consistency and stability of the measurement results. It indicates how reliably the test measures what it is intended to measure (Boone, 2016; Sumintono et al., 2015; Sumintono, 2018).

SUMMARY OF 10 MEASURED (NON-EXTREME) Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	240.0	515.0	.00	.12	1.01	.2	.89	-.8
S.D.	112.2	.0	1.45	.01	.10	1.6	.25	1.7
MAX.	394.0	515.0	1.93	.13	1.23	3.8	1.47	3.0
MIN.	98.0	515.0	-2.01	.11	.90	-1.6	.60	-2.2

REAL RMSE .12 TRUE SD 1.45 SEPARATION 11.78 Item RELIABILITY .99
MODEL RMSE .12 TRUE SD 1.45 SEPARATION 12.04 Item RELIABILITY .99
S.E. OF Item MEAN = .48

UMEAN=.0000 USCALE=1.0000
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00
5030 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 4325.76 with 4518 d.f. p=.9796
Global Root-Mean-Square Residual (excluding extreme scores): .3766
Capped Binomial Deviance = .1825 for 5150.0 dichotomous observations

Figure 5. Summary statistics (Item)

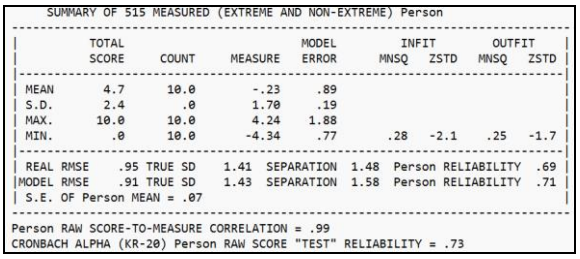
Item separation measures, as shown in Figure 5, are exceptionally high. The real separation value of 11.78 and model separation value of 12.04 demonstrate the instrument's effectiveness in distinguishing items based on their difficulty. This indicates a broad range of item difficulties and suggests that the instrument can categorize items into multiple difficulty levels. Such a high separation value reflects a well-constructed test design that covers a wide spectrum of abilities and provides comprehensive coverage of the targeted construct.

Item reliability reflects the stability and consistency of the item difficulty hierarchy within the instrument. Real and model reliability values of 0.99 indicate excellent reliability, demonstrating that variations in item measures result from actual differences in difficulty rather than measurement error. This high level of reliability ensures the instrument's ability to consistently rank item difficulties across various contexts by affirming its validity and robustness.

The combined item separation and reliability values confirm that the HOTS-TCA instrument is highly reliable and well-designed. Its high separation and reliability suggest a diverse range of well-distributed items, making it an effective tool for assessing participants with varying abilities.

4. Person Separation and Reliability

Person separation and reliability statistics evaluate the instrument's ability to differentiate between respondents based on their levels of the measured trait. The person separation index indicates the number of statistically distinct levels of the trait identified by the instrument (Kimberlin & Winterstein, 2008). The person reliability coefficient, analogous to Cronbach's alpha, reflects the proportion of actual variance in the observed person measures (Greco et al., 2018; Peterson, 1994).



SUMMARY OF 515 MEASURED (EXTREME AND NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	4.7	10.0	-.23	.89				
S.D.	2.4	.0	1.70	.19				
MAX.	10.0	10.0	4.24	1.88				
MIN.	.0	10.0	-4.34	.77	.28	-2.1	.25	-1.7

REAL RMSE	.95	TRUE SD	1.41	SEPARATION	1.48	Person RELIABILITY .69
MODEL RMSE	.91	TRUE SD	1.43	SEPARATION	1.58	Person RELIABILITY .71
S.E. OF Person MEAN = .07						

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .73

Figure 6. Summary Statistics (Person)

The person separation values in this analysis, as shown in Figure 6, are relatively low, with an absolute separation of 1.48 and a model separation of 1.58. While this may suggest that the HOTS-TCA instrument is not optimally differentiating individuals based on their ability levels, an argument can be made that the observed separation values are still within an acceptable range. Person separation values of 2.0 or greater are often considered ideal, as they indicate the ability to effectively divide individuals into at least three distinct ability groups (Boone et al., 2014; Linacre, 2019). However, values as low as 1.5 can still provide helpful information about the examinee's performance and allow for the identification of at least two distinct ability groups (e.g., "low" and "high" ability) (Linacre, 2019).

Given the inherent challenges in developing assessment instruments that precisely measure a wide range of abilities, the HOTS-TCA separation values, while not optimal, may still offer meaningful insights into the diverse skill levels of the examined population. These relatively low person separation values imply that the test items may not be sufficiently varied in their difficulty level or too easy or difficult for the examinees, failing to adequately reflect the full range of examinee abilities. This could limit the instrument's capacity to provide comprehensive and meaningful insights into the diverse skill levels of the examined population.

The person reliability values in this analysis are moderately adequate, with an absolute reliability of 0.69 and a model reliability of 0.71. Reliability values range from 0 to 1, where higher values indicate better consistency. Reliability values in the 0.69-0.71 range are considered relatively good but not optimal. This instrument is reliable for identifying differences in participant abilities, but the results may be less accurate in differentiating very fine-grained ability groups. This relatively low reliability implies that the instrument may not have a sufficiently varied range of item difficulties to reflect the full spectrum of participant abilities precisely.

The raw score-to-measure correlation of 0.99 demonstrates a nearly perfect relationship between participants' raw scores (total correct answers) and their ability estimates in logits, as determined by the Rasch model. This high correlation suggests that raw scores are a very accurate representation of participants' latent abilities. This finding highlights the instrument's effectiveness in converting raw scores into precise ability estimates. Although raw scores and Rasch analysis yield similar results, the Rasch model provides additional benefits by accounting for item difficulty and generating interval-level measurements.

The Cronbach's Alpha (or KR-20 for dichotomous items) value closer to 1 would indicate higher internal consistency, while the KR-20 value of 0.73 indicates the test's internal consistency. While this value is categorized as acceptable, it is not optimal. This means that the test items are sufficiently consistent in measuring the same underlying ability, though there is room for improvement.

5. Person Measure (Ability Estimates)

JURNAL GANTANG. December 2024; IX(2): 191 – 204
p-ISSN. 2503-0671
e-ISSN. 2548-5547

The analysis of the Person Measure offers valuable insights into the relationship between test difficulty and respondent abilities, highlighting how well the assessment aligns with the skills of the participants (Boone, 2016; Boone & Noltemeyer, 2017; Sumintono et al., 2015).

Person STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
104	10	10	4.24	1.88		MAXIMUM MEASURE			.00	.00	100.0	100.0	104S
375	10	10	4.24	1.88		MAXIMUM MEASURE			.00	.00	100.0	100.0	375S
403	10	10	4.24	1.88		MAXIMUM MEASURE			.00	.00	100.0	100.0	403S
449	10	10	4.24	1.88		MAXIMUM MEASURE			.00	.00	100.0	100.0	449S
508	10	10	4.24	1.88		MAXIMUM MEASURE			.00	.00	100.0	100.0	508S
60	9	10	2.88	1.11	.77	-.1	.35	-.2	.44	.29	90.0	90.0	060S
...													
241	0	10	-4.34	1.88		MINIMUM MEASURE			.00	.00	100.0	100.0	241N
257	0	10	-4.34	1.88		MINIMUM MEASURE			.00	.00	100.0	100.0	257S
381	0	10	-4.34	1.88		MINIMUM MEASURE			.00	.00	100.0	100.0	381N
459	0	10	-4.34	1.88		MINIMUM MEASURE			.00	.00	100.0	100.0	459N
MEAN	4.7	10.0	-.23	.89	1.00	.1	.89	.1			79.1	79.8	
S.D.	2.4	.0	1.70	.19	.36	.8	.77	.7			12.1	4.6	

Figure 7. Person statistics

The analysis of the Person Measure, as shown in Figure 7, provides insights into the alignment between the test difficulty and respondent abilities. The mean person measure of -0.23 indicates that, on average, respondents have slightly lower ability than the difficulty of the test items, suggesting the test may be moderately challenging for most participants. Ideally, the mean measure should align closely with zero to ensure the test is well-targeted to the population. The standard deviation (SD) of 1.70 reflects a wide range of respondent abilities, demonstrating that the test effectively captures individual differences, which is beneficial for identifying variability in skills.

The fit statistics further confirm the appropriateness of the test. The average infit mean square (MNSQ) of 1.00 with an SD of 0.36 and the outfit MNSQ of 0.89 with an SD of 0.77 indicates that most responses align well with the Rasch model, with minimal unexpected responses. Additionally, the observed exact match percentage (79.1%) closely aligns with the expected percentage (79.8%), reinforcing the test's reliability.

6. Person-Item Map (Wright Map)

The Item-Person Wright Map visually represents the alignment between respondent abilities and item difficulties on a shared measurement scale. In Figure 8, respondents are displayed on the left, and items are displayed on the right, both measured in logits.

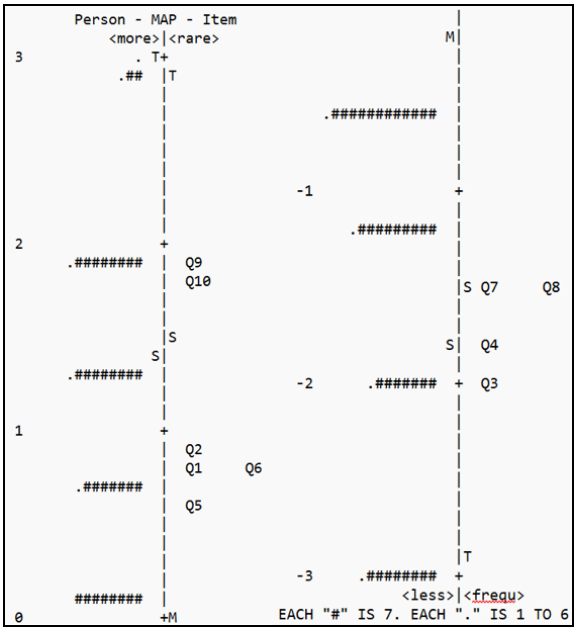


Figure 8. Person-Item Map

The distribution of respondents, as shown in Figure 8, reveals that the majority have abilities ranging from -1 to 2 logits, with a dense cluster around 0 and 1. Only a small number of respondents are positioned at the extremes, with abilities above 2 or below -1. The item distribution ranges from approximately -2 (easiest items, such as Q3 and Q4) to 2 (most difficult items, Q9 and Q10).

The map indicates that the instrument is well-targeted for middle-ability respondents, as most items align closely with their abilities. Items Q1, Q2, Q5, and Q6 are situated near the ability level of most respondents, ensuring accurate measurement. However, there is a mismatch at the extremes: high-ability respondents may find the most difficult items (Q9 and Q10) relatively easy. In contrast, low-ability respondents may struggle with even the most straightforward items (Q3 and Q4).

Despite the overall alignment, improvements could enhance the test's discriminatory power. Adding more challenging items would better assess high-ability respondents (>2 logits) while introducing easier items would provide better coverage for low-

ability respondents (<-2 logits). Additionally, items like Q7 and Q8, located around -1 logits, require further evaluation if their fit statistics indicate suboptimal performance.

7. Differential Item Functioning (DIF)

The provided DIF statistics present each test item's Chi-square values, degrees of freedom (df), p-values, mean squares, and t-values (ZSTD). These statistics help assess whether any item is biased toward one group over the other.

Person CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	BETWEEN-CLASS MEAN-SQUARE	t=ZSTD	Item Number	Name
2	.8171	1	.3660	.3120	-.2112	1	Q1
2	.4088	1	.5226	.1585	-.5019	2	Q2
2	1.2686	1	.2600	.2960	-.2361	3	Q3
2	.0000	1	1.0000	.0005	-1.4823	4	Q4
2	.4367	1	.5087	.1559	-.5081	5	Q5
2	.4098	1	.5221	.1574	-.5046	6	Q6
2	.8296	1	.3624	.2227	-.3640	7	Q7
2	.2344	1	.6283	.0656	-.7945	8	Q8
2	.8629	1	.3529	.3680	-.1298	9	Q9
2	.0909	1	.7631	.0364	-.9471	10	Q10

Figure 9. Item DIF

Item DIF, as shown in Figure 9, a p-value (probability column) greater than 0.05 indicates no significant DIF, suggesting that the item behaves similarly for both groups. Conversely, a p-value less than 0.05 would indicate significant DIF, suggesting potential bias.

Upon reviewing the results, none of the items show significant DIF. For all ten items, the p-values are above 0.05, ranging from 0.0909 (for Q10) to 1.0000 (for Q4), meaning there are no statistically significant differences in how the public and private school groups responded. The Chi-square values for each item indicate that the observed differences between groups are minimal, further supported by the t-values, which do not suggest any meaningful variation in item difficulty. For example, Q4 has a p-value of 1.0000, indicating no DIF at all. At the same time, other items like Q1, Q2, and Q3 show similarly high p-values, reinforcing the conclusion that the items function equivalently for both groups.

In conclusion, the DIF analysis reveals that the HOTS-TCA instrument does not exhibit any bias based on the school background (public vs. private). The test items are equally fair and valid for both groups, suggesting that the instrument measures the intended construct consistently across diverse subgroups.

Identification of Specific Challenges or Misconceptions Observed During the Test and Interviews

Based on the post-test interviews, the students' challenges in applying spatial thinking abilities to solve trigonometric problems varied depending on the questions' complexity and spatial thinking levels. The findings can be categorized based on students' spatial thinking abilities as high, medium, or low, corresponding to the difficulty level of the problems they encountered.

Table 1. Students' challenges in applying spatial thinking to trigonometry

Spatial Thinking Ability	Strengths	Challenges	Misconceptions
High Spatial Thinking Abilities	• Strong visualization skills for complex geometric relationships. • Able to interpret 3D relationships better than lower-ability students.	• Struggles with applying trigonometric formulas to visualized spatial representations. • Difficulty in manipulating visualizations in a mathematical context.	• Disconnect between visualization and mathematical manipulation. • Occasional misapplication of trigonometric formulas in practical problems.
Medium Spatial Thinking Abilities	• Basic understanding of geometric relationships. • Can follow standard trigonometric procedures.	Difficulty mentally manipulating complex shapes and visualizing spatial relationships in 2D and 3D.	• Struggles to connect theoretical formulas with geometric features. • Frequent misinterpretations of diagrams and conceptual errors in applying trigonometry.
Low Spatial Thinking Abilities	Difficulty with most aspects of spatial thinking.	• Struggles to interpret basic 2D representations of 3D objects. • Difficulty in understanding the connection between angles,	• Frequent misinterpretations of diagrams. • Difficulty applying theoretical knowledge to real-world problems. • Weak connection between mathematical

distances, and sides.	symbols and spatial meanings.

Students with higher spatial thinking abilities demonstrated stronger visualization skills and could interpret most problems involving complex geometric relationships. However, even these students faced difficulties when problems required a deep understanding of three-dimensional spatial relationships. For example, in problems where a building was represented in two dimensions but required understanding its three-dimensional structure, these students could visualize its components better than those with lower spatial abilities. However, they still struggled to manipulate these visualizations in a mathematical context, particularly in applying trigonometric formulas to the spatial representations. Despite this, their ability to mentally rotate and shift geometric shapes in their minds was generally stronger, allowing them to approach problems involving angles and distances with a clearer understanding. However, errors occasionally occur when applying theoretical knowledge to practical contexts.

Students with medium spatial thinking abilities encountered significant difficulties with tasks requiring them to manipulate objects mentally. These students often showed a basic understanding of geometric relationships but struggled with more complex visualizations and interpretations. For instance, in problems where angles and distances needed to be assessed in relation to three-dimensional objects depicted in two-dimensional diagrams, these students could not mentally manipulate the shapes as quickly as those with high spatial abilities. While they could follow basic procedures with standard trigonometric equations, their ability to interpret diagrams and apply spatial relationships to solve real-world problems was more limited. The struggle to connect the theoretical trigonometric formulas with the geometric features of the problem was more pronounced, leading to frequent misinterpretations and incorrect answers. Their understanding of spatial relationships, such as orientation and distance, was less developed, causing misjudgments in calculating angles and distances, further hindering their problem-solving abilities.

Students with low spatial thinking abilities faced the most significant challenges when solving trigonometric problems. These students had considerable difficulty with almost all aspects of spatial thinking, particularly visualizing the geometric structures involved. They struggled to interpret even basic 2D representations of 3D objects, which led to frequent misinterpretations. For example, in problems where a 2D representation of a building was provided, these students found it almost impossible to visualize or interpret the spatial characteristics of the building as a three-dimensional object, resulting in many incorrect answers. Their difficulty mentally manipulating objects and understanding spatial relationships made it particularly challenging to understand the connections between angles, distances, and sides in geometric shapes. These students also found analyzing and interpreting diagrams challenging, often missing key details or misreading the information provided. The inability to connect mathematical symbols and notations with their spatial meanings further hindered their ability to solve problems correctly. Additionally, applying theoretical knowledge to practical contexts, such as calculating real-world measurements, was especially difficult due to their weak spatial thinking abilities.

IV. Conclusion

This study highlights the significant role of spatial thinking abilities in solving high-order thinking skill (HOTS) trigonometric comparison problems. Students with high spatial abilities perform well in visualizing and interpreting geometric relationships, although challenges remain in tackling complex three-dimensional tasks. Medium-ability students struggle with mental object manipulation and applying spatial thinking to real-world problems. In contrast, low-ability students face difficulties in basic visualization and interpreting geometric structures, leading to frequent misconceptions.

These findings emphasize the need for targeted interventions to enhance students' spatial thinking abilities. Furthermore, the HOTS-TCA instrument has been validated as a reliable and effective tool for assessing spatial thinking within mathematical contexts, providing a robust framework for diagnosing students' abilities and challenges.

To build upon these findings, future research should explore innovative teaching strategies, such as technology-enhanced tools and interactive visualizations, to support the development of spatial thinking abilities. Expanding the study to broader and more diverse populations will help generalize the findings and refine the HOTS-TCA instrument for comprehensive assessment. Longitudinal studies are also recommended to examine the lasting impact of improved spatial thinking on students' mathematical performance and real-world problem-solving. Further refinement of the HOTS-TCA instrument is suggested to address extreme ability levels better and ensure a more balanced assessment. These efforts will contribute to a deeper understanding and effective improvement of spatial thinking in mathematical education.

Limitation

This study's scope is limited by several factors, including its focus on 11th-grade students from senior high schools in Tanjungpinang, Indonesia, which may restrict the generalizability of the findings to other educational contexts. Additionally, while the study examines spatial thinking in higher-order trigonometric comparison problems, it excludes other mathematical domains, such as algebraic reasoning and logical reasoning, which are also important for understanding students' problem-solving skills. Finally, while providing valuable insights, the reliance on a single assessment instrument (HOTS-TCA) and the use of Rasch model analysis suggest that incorporating other psychometric methods or Classical Test Theory (CTT) approaches could enhance the interpretation of the data.

Reference

Adams, J., Resnick, I., & Lowrie, T. (2023). Supporting senior high-school students' measurement and geometry performance: Does spatial training transfer to mathematics achievement? *Mathematics Education Research Journal*, *35*(4), 879–900. https://doi.org/10.1007/s13394-022-00416-y

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. (revised Bloom's taxonomy)*. Longman.

Anggraini, R., & Putra, E. S. (2020). The ability of cadets to solve trigonometry routine and non-routine problems. *Journal of Physics: Conference Series*, *1480*(1), 012027. https://doi.org/10.1088/1742-6596/1480/1/012027

Battista, M. T., Frazee, L. M., & Winer, M. L. (2018). *Analyzing the relation between spatial and geometric reasoning for elementary and middle school students* (pp. 195–228). https://doi.org/10.1007/978-3-319-98767-5_10

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch Model*. Psychology Press. https://doi.org/10.4324/9781410614575

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education*, *15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, *4*(1), 1416898.

https://doi.org/10.1080/2331186X.2017.1416898

Boone, W. J., Staver, R. J., & Yale, S. M. (2014). *Rasch Analysis in the Human Sciences*. Springer.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q3: identification of local dependence in the rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Everitt, B. S., & Dunn, G. (2001). *Applied multivariate data analysis*. Wiley. https://doi.org/10.1002/9781118887486

Fisher, W. P. Jr. (2024, November 24). *Reliability, Separation, Strata Statistics*. Https://Rasch.Org/Rmt/Rmt63i.Htm. https://rasch.org/rmt/rmt63i.htm

Gilligan, K. A. (2020). Make Space: The importance of spatial thinking for learning mathematics. *Frontiers for Young Minds*, *8*. https://doi.org/10.3389/frym.2020.00050

Greco, L. M., O'Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2018). Meta-analysis of coefficient alpha: a reliability generalization study. *Journal of Management Studies*, *55*(4), 583–618. https://doi.org/10.1111/joms.12328

Huei, O. K., Che' Rus, R., & Kamis, A. (2020). Construct validity and reliability in content knowledge of design and technology subject: a rasch measurement model approaches for pilot study. *International Journal of Academic Research in Business and Social Sciences*, *10*(3). https://doi.org/10.6007/IJARBSS/v10-i3/7075

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298. https://doi.org/10.1207/S15324818AME1604_2

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364

Lakin, J. M., & Wai, J. (2020). Making space for spatial talent. *Phi Delta Kappan*, *102*(4), 36–39. https://doi.org/10.1177/0031721720978061

Linacre, J. M. (2002). What do infit, outfit, mean-square, and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2019). A user's guide to WINSTEPS® Rasch-model computer programs: program manual 4.4. 6. *Mesa-Press, Chicago, IL*.

Linacre, J. M. (2024, November 24). *Fit diagnosis: infit outfit mean-square standardized*.

Ma, W., Chen, H., Zhang, G., de Melo, C. M., Yuille, A., & Chen, J. (2024). *3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark*.

Nanmumpuni, H. P., & Retnawati, H. (2021). Analysis of senior high school student's difficulty in resolving trigonometry conceptual problems. *Journal of Physics: Conference Series*, *1776*(1), 012012. https://doi.org/10.1088/1742-6596/1776/1/012012

Ngu, B. H., & Phan, H. P. (2020). Learning to solve trigonometry problems that involve algebraic transformation skills via learning by analogy and learning by comparison. *Frontiers in Psychology*, *11*.

https://doi.org/10.3389/fpsyg.2020.558773

Ocy, D. R., Rahayu, W., & Makmuri, M. (2023a). Development of a hots-based mathematical abstraction ability instrument in trigonometry using Riau Islands province culture. *Jurnal Gantang*, *8*(1), 37–52. https://doi.org/10.31629/jg.v8i1.5654

Ocy, D. R., Rahayu, W., & Makmuri, M. (2023b). Rasch model analysis: development of hots-based mathematical abstraction ability instrument according to riau islands Culture. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *12*(4), 3542–3560.

Palupi, E. L., Juniati, D., & Khabibah, S. (2023). Research on spatial reasoning in mathematics education: Trend and opportunity. *Jurnal Gantang*, *8*(2), 113–123. https://doi.org/10.31629/jg.v8i2.6619

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, *21*(2), 381. https://doi.org/10.1086/209405

Resnick, I., Harris, D., Logan, T., & Lowrie, T. (2020). The relation between mathematics achievement and spatial reasoning. *Mathematics Education Research Journal*, *32*(2), 171–174. https://doi.org/10.1007/s13394-020-00338-7

Riani Siregar, N. A., Susanti, & Mariyanti Elvi. (2021). analisis model rasch disposisi matematis mahasiswa pada program studi pendidikan matematika UMRAH. *Jurnal Gantang*, *6*(1), 1–10. https://doi.org/10.31629/jg.v6i1.3118

Riastuti, N., Mardiyana, M., & Pramudya, I. (2017). Students' errors in geometry are viewed from spatial intelligence. *Journal of Physics: Conference Series*, *895*, 012029. https://doi.org/10.1088/1742-6596/895/1/012029

Rios, S. G. S. A. N. D. J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Routledge*, *19*(2–3), 170–187. https://doi.org/10.1080/13803611.2013.767621

Rohimah, S. M., & Prabawanto, S. (2020). Students' difficulties in solving trigonometric equations and identities. *Journal of Physics: Conference Series*, *1521*(3), 032002. https://doi.org/10.1088/1742-6596/1521/3/032002

Sorby, S. A., Duffy, G., & Yoon, S. Y. (2022). Math instrument development for examining the relationship between spatial and mathematical problem-solving skills. *Education Sciences*, *12*(11). https://doi.org/10.3390/educsci12110828

Sumintono, B. (2018). Rasch model measurements are tools for the assessment of learning. *Proceedings of the 1st International Conference on Education Innovation (ICEI 2017)*, *173*(Icei, 2017), 38–42. https://doi.org/10.2991/icei-17.2018.11

Sumintono, B., & Widhiarso, W. (2013). *Aplikasi pemodelan rasch pada assessment Pendidikan [applications of rasch modeling in educational assessments]*.

Sumintono, Bambang, & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada asesmen pendidikan*.

Uttal, D. H., & Cohen, C. A. (2012). *Spatial thinking and STEM education* (pp. 147–181). https://doi.org/10.1016/B978-0-12-394293-7.00004-2

Wang, W.-C., & Chen, C.-T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the

winsteps program for the family of rasch models. *Educational and Psychological Measurement*, *65*(3), 376–404. https://doi.org/10.1177/0013164404268673

Wright, B. G. (1994). *Reasonable mean-square fit values*. *8*. https://ci.nii.ac.jp/naid/10017021263/

Xie, F., Zhang, L., Chen, X., & Xin, Z. (2020). Is spatial ability related to mathematical ability: a meta-analysis. *Educational Psychology Review*, *32*(1), 113–155. https://doi.org/10.1007/s10648-019-09496-y

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171–177. https://doi.org/10.3969/j.issn.1002-0829.2014.03.010