Item Analysis of Measuring STEM Teachers' Practice in Higher-Order Thinking Skills Questioning Using Rasch Model

Nor'Aida Khairuddin[1*] and Rohaya Talib[2]

[1,2] School of Education, Faculty of Social Science and Humanities, Universiti Teknologi Malaysia, Malaysia

Abstract.

For interdisciplinary approach in STEM integration, HOTS questioning application is part of inquiry-based learning which helps learners to construct knowledge by discovering new concepts and developing design thinking. In Malaysia, STEM teachers still insufficient to construct HOTS questions in teaching and learning session. This study aims to analyse HOTS questioning practice items using the Rasch model in terms of validity and reliability. The inventory was developed to measure STEM teachers HOTS questioning practice from five dimensions such as external world view, classroom interactions, deep and analytical thinking and complex thinking strategies using Likert type scale. The pilot study was conducted in eight secondary schools which involved 86 upper secondary STEM teachers. WINSTEPS version 3.72.3 was applied to analyse the pilot study outcomes. The results showed that the reliability based on Cronbach Alpha is 0.97. The construct validity was determined by positive Point Measure Correlation (PMC) value, infit and outfit MSNQ between 0.4 to 1.5 and ZSTD range from -2.0 to 2.0. The summary statistics showed that person and item reliability were 0.97 and 0.95, respectively. The person separation index with the value of 5.39 and item separation index, 4.59 were considered as excellent separation. Finally, for unidimensionality, raw variance explained by measure was 49.1%, and the unexplained variance in the first factor was 6.7% meet the criteria Therefore, after eliminating the unfit items, the results demonstrated that this inventory was valid and reliable to measure STEM teachers HOTS questioning practice in future.

Keywords: STEM Integration, HOTS Questioning, Item Analysis, Rasch Model, Interdisciplinary STEM

1. Introduction

The integration of STEM education encourages students to understand multiple STEM disciplines' concepts to generate innovative solutions for complex real-life and non-routine problems (Kelley & Knowles, 2016).The interdisciplinary approach is part of STEM integration which would concentrate on evolving higher-order thinking such as critical thinking and problem-solving for the real-world problem or issue rather than understanding STEM subject-specific content (Deprez et al., 2018). Inquiry-based learning is an interdisciplinary approach that would develop students to construct their knowledge by discovering new concepts from scientific investigation and enhancing design thinking (Zhang, 2016). Based on the constructivism theory, inquiry-based learning could develop active, authentic and open-ended learning in the classroom (Irekpita et al., 2020). Students are encouraged to elaborate on their ideas, making predictions and observation through experimental or hands-on activities. In inquiry-based learning, HOTS questioning application in one of the instructional strategies could evolve students' reasoning skills when discovering the real-world issues and develop their design skills to solve real-world problems (Aziza, 2018). Several studies have discussed the importance of STEM teachers to generate HOTS questions by formulating the connection between multiple subjects' content area and real-world contextualization (Gardner & Tillotson, 2019; Ramdiah et al., 2019; Ritz & Fan, 2015; Rosidin et al., 2019).

However, STEM teachers' issues in practicing HOTS questioning were well documented in most educational literature(Abdullah et al., 2017; Ramdiah et al., 2019; Tajudin et al., 2018; Zeegers & Elliott, 2019). Several studies from the previous literature had mentioned that most teaching training programmes failed to equip them with STEM teachers' skills in HOTS questioning from inquiry teaching (Çalik et al., 2015; Saribas, 2015; Yeung, 2015). These STEM teachers have poor understanding and competence to generate HOTS questions in inquiry-based teaching either from scientific investigation, problem-solving in mathematics or interdisciplinary engineering design. Several studies have proved the adverse effects of frequent application of low-level or convergent questions towards students' learning achievement and higher-order thinking in STEM (Purdum-Cassidy et al., 2015; Wahono et al., 2020).

Most secondary schools, especially in Malaysia, still employed low-level and closed-ended questions in the classroom, for instance, by the recalling questions to confirm the learning facts or identify glitch in STEM concepts(Iksan & Daniel, 2015; Perera & Asadullah, 2019). For example, most STEM teachers struggled to master subject content knowledge in each discipline, latest curriculum and pedagogical skills due to time constraint and restricted assessment procedures (Mahmud et al., 2018). STEM teachers' attitude, motivation and inner belief towards practising HOTS questions also cause this issue to exist (Tengku Ariffin et al., 2018). Also, they even did not have an opportunity to verbalize open-ended questions with accurate scientific language based on the STEM content (Ping et al., 2019). Some of them refused to pursue HOTS learning outcomes but more concern in completing certain learning content specific goals for high-stake examinations (Perera & Asadullah, 2019; Yen & Halili, 2015). The paucity of STEM pedagogical training, especially in HOTS questioning, cause these STEM teachers insufficient in producing high quality of STEM teaching (ASM, 2018; MOE, 2018).

For this study's purpose, was to design a valid and reliable inventory to measure STEM teachers' HOTS questioning practice in the context of formative assessment from an integrated STEM interdisciplinary approach. Therefore, this study focused on analysing HOTS questioning items using the Rasch model in terms of validity and reliability.

2. Methods

2.1 Study sample

This study was a cross-sectional survey design that measured STEM teachers practice in HOTS questioning when performing a teaching and learning session. This pilot study was conducted using convenience sampling, which represented from the sample of the population. This pilot study sample involved 86 upper secondary STEM teachers as the respondents in eight secondary schools around Johor Bahru, Johor Malaysia. The respondents were selected from the school administrators based on their teaching experienced and academic background in STEM disciplines. Ethical approval for this pilot study was authorized by the Educational Planning and Research Division, Ministry of Education and Johor State Office of Education.

2.2 Instrument

The STEM experts and psychometricians validated the content of the inventory. The inventory was a self-administered and contains 60 items which measuring STEM teachers HOTS questioning practice based on four dimensions: External World View (6 items), Classroom Interactions (12 items), Deep and Analytical Thinking (30 items) and Complex Thinking Strategies (12 items). In the external world view dimension, six items measured STEM teachers practice in constructing HOTS questions by relating subject content matter to a real-life situation through authentic learning. In the classroom interaction dimension, 12 items separated into two elements: Wait-time in HOTS questioning session (6 items) and recognition of STEM teachers toward students' response (6 items).

In a deep and analytical thinking dimension, these 30 items from this dimension measured STEM teachers' application in generating analytical questions to enhance students' thinking and complex reasoning skills. These 30 items were divided into five elements: Inductive reasoning (6 items), Deductive reasoning (6 items), Error analysis (6 items), Constructing support (6 items) and analysing perspectives (6 items). Finally, in the dimension of complex thinking strategies, 12 items measured STEM teachers' practice in generating HOTS question to develop students' design thinking skills by solving the problem and making decision through learning activities. The dimension of complex thinking strategies was divided into two elements, problem-solving (6 items) and making decisions (6 items). All items in all dimensions used polytomous response options of five points Likert type scale. The scale based on the frequency of practices using five options rating scales; (1) Never, (2) Rarely (3) Frequently, (4) Usually and (5) Every time. Table 1 presented a list of items in each dimension.

Table 1: List of items in each dimension.

Dimensions	Elements	Items	Item description
External world view	Authenticity	AT41	Real-life situations
		AT42	STEM basic concepts
		AT43	Multiple STEM concepts
		AT44	Prepare teaching aid and material
		AT45	Relate with current issues
		AT46	Group activities
Classroom interactions	Wait-time	WT47	Time to response
		WT48	Longer wait-time
		WT49	Various cognitive development

		WT50	Rephrase the questions
		WT51	Listens to students' response
		WT52	Eye contacts
	Recognition	CR53	Scrutinise student's response
		CR54	Positive feedbacks
		CR55	Balancing students' participation
		CR56	Encourage response based on stimulus
		CR57	Positive body language
		CR58	Scaffold students' response
Deep and analytical thinking	Inductive reasoning	IR59	Identify a concept
		IR60	Generate a hypothesis
		IR61	Obtain relevant data
		IR62	Test the hypothesis
		IR63	Relate the hypothesis with real-life situations
		IR64	Affirm the hypothesis
	Deductive reasoning	DR65	Relate the real-life situations with concepts
		DR66	List the related theories
		DR67	Apply *if* to predict the expected results
		DR68	Predict the experiment outcomes
		DR69	State the conclusions
		DR70	Implications from experiment results
	Error analysis	EA71	Identify any misconceptions
		EA72	Clarify the misconceptions
		EA73	Identify any errors of the concepts
		EA74	Opportunities to change the response
		EA75	Modify the existing answer
		EA76	Reflect the given answer
	Constructing support	CS77	Information from the reliable sources
		CS78	Elaborate answer based on evidence
		CS79	Create strong argument
		CS80	Relate the evidence with theories
		CS81	Elaborate the concepts clearly
		CS82	Develop a new prototype model from STEM concepts
	Analysing perspective	AP83	Opinion in current issues
		AP84	Relate STEM concepts to current issues
		AP85	Terminology in STEM
		AP86	Describe STEM concepts in detail
		AP87	Implications of the STEM concepts by analogy
		AP88	Evaluate others' response
Complex thinking strategies	Problem-solving	PS89	Generate numerous ideas
		PS90	Deep understanding of the concepts for design task
		PS91	Recognise the future obstacles
		PS92	Prioritise the action plan
		PS93	Delve into various strategies
		PS94	Determine the best action plan
	Making decisions	MD95	Specify the STEM concepts to integrate
		MD96	Investigate the suitable STEM concepts
		MD97	Determine the STEM concepts accurately
		MD98	Design a prototype model
		MD99	Evaluate the outcomes of the design
		MD100	Justify the decisions

2.3 Rasch model analysis

All of the items were analysed using Rasch model by the Winstep 3.27.3 software. The software was used to assess the inventory's validity and reliability. The validity and reliability were evaluated in a five-stepped approach. First, the person and item polarity were analysed. Secondly, the fits statistics were applied to identify problematic items and persons to remove or modify. Thirdly, the person and item reliability, including person and item separation index, were evaluated. Subsequently, unidimensionality and local independency for the inventory was determined. Finally, item difficulty was investigated from item-person map.

2.3.1 Person and Item Polarity

The positive value for Point Measure Correlation (PtMea Corr) required to ensure that all of the items measured the dimensions being measured (Bond & Fox, 2015). According to Bond & Fox (2015) and Linacre (2015), if PtMea Corr's value exceeds more than 0.30 with a positive value, indicated that the items could measure the proposed dimension. The item's high polarity means that the item could differentiate between the respondents (Rahman et al., 2020). If PtMea Corr's value was negative, the item would be removed because it did not measure the dimension. For person polarity, if the PtMea Corr value was negative, the person was considered an outlier that must be removed (Norhayati et al., 2020).

2.3.2 Person and item fit

Item fit was analysed to identify if the items could measure the latent trait or dimension being measured (Koskey et al., 2016). The person fit was analysed to identify any person who contributed towards aberrant from the normal response(Bond & Fox, 2015). In this study, according to Linacre (2015), the range of 0.4 to 1.5 for MNSQ (Mean-square) outfit and infit was acceptable for the person and item fit of polytomous data (Likert type scale) if the instrument was self-developed. This study accepted the z-standard value from the range of -2 to 2.

2.3.3 Person and item reliability

The Alpha Cronbach (KR-20) value of 0.60 was accepted if the instrument was newly self-developed (Pallant, 2020). The higher value of Cronbach Alpha (KR-20) means the inventory has excellent internal consistency among all items within the same dimension (Souza et al., 2017). Based on the Rasch model, the value which exceeds more than 0.80 indicated high reliability for item and person (Bond & Fox, 2015). High person reliability in the Rasch model is defined as the items can differentiate between respondents for the variables or construct being measured (Bond, 2003).

2.3.4 Person and item separation index

Person separation index defined as an efficiency of the items to differentiate and classify the respondents' traits or characteristics from the measuring variables (Teman, 2018). Item separation index is used to verify item hierarchy by distributing the items based on item difficulty (Geldenhuys et al., 2019). The higher value of both person and item separation index indicated excellent separation. The values between 1.5 and 2.0 are considered acceptable (Lo et al., 2015). They suggested the value must exceed 3.0 for an excellent person and item separation.

2.3.5 Unidimensionality and Local Independency

In the Rasch model, unidimensionality and local independence criteria provide the items' performance to contribute to a single construct being measured (Lo et al., 2015). Unidimensionality is defined that all items could relate to the same latent variable (Bond & Fox, 2015; Linacre, 2015). A minimum value of *Raw variance explained by the measure, 40%* is required for better measurement of unidimensionality. The *unexplained variance in the 1st contrast* value should not exceed 15%. The *eigenvalue of 1st contrast* should in the range from 1.4 to 3.0 (Linacre, 2015). Local dependence in Rasch analysis identifies the correlation between the items (Linacre, 2015). The Standard Residual Correlations value was analysed to determine either these items are dependent on each other (Bond & Fox, 2015). If the standard residual correlation between the items' value exceeds more than 0.70, these items are dependent and share the same traits (Linacre, 2015; Norhayati et al., 2020). One of these items has to be removed from the inventory.

2.3.6 Item-person map (*Wright map*)

Item-person map in Rasch analysis represents interval-level measurement scale using a logic scale to identify the location of respondents' ability on item difficulty (Bond & Fox, 2015). The valid inventory should balance the items' distribution on various levels of item difficulty (Boone, 2016). The gap between the items, the fit of respondents 'measure, item difficulty and test-item targeting could be investigated (Petrillo et al., 2015). If the targeting index higher than 0 indicates that the respondent contributes to positive responses and lower than 0 signify as negative responses (Lo et al., 2015; Zanon et al., 2016).

3. Results

The total sample of 86 respondents from all STEM disciplines in Johor Bahru for this pilot study were analysed to investigate this inventory validity and reliability. The demographic information for 86 respondents differed in terms of gender, academic background, teaching experience, and schools type. After performing a person measure analysis, 10 respondents were eliminated and detected as outliers or misfit person. Only 76 respondents were analysed for further psychometric analysis. The result of the analysed items' general description using the Rasch model, which includes fit statistics and point measure correlation (PtMea Corr) values, was presented in Table 2.

Table 2: General description of analysed items based on dimensions.

Dimensions	Item	Measure	Model S.E	Infit		Outfit		PtMea Corr
				MNSQ	ZSTD	MNSQ	ZSTD	
<u>External world view</u>								
Authenticity	AT41	0.17	0.20	1.12	0.7	1.23	1.1	0.55
	AT42	-0.74	0.23	2.54	5.5	2.75	5.9	0.27
	AT43	0.32	0.19	0.72	-1.5	0.79	-1.1	0.66
	AT44	1.04	0.17	1.43	2.2	1.78	3.5	0.39
	AT45	0.17	0.20	1.30	1.5	1.35	1.7	0.62
	AT46	0.43	0.19	1.43	2.1	1.64	2.8	0.55

Classroom interactions								
Wait-time	WT47	0.98	0.17	2.05	4.6	2.51	5.8	0.24
	WT48	0.24	0.20	1.55	2.5	1.65	2.9	0.36
	WT49	0.35	0.19	0.80	-1.1	0.84	-0.8	0.50
	WT50	-0.69	0.23	0.82	-0.9	0.82	-0.9	0.52
	WT51	-1.17	0.24	1.18	0.9	1.15	0.8	0.44
	WT52	-2.36	0.25	1.39	2.2	1.39	1.6	0.39
Recognition	CR53	-1.99	0.25	1.36	1.9	1.35	1.6	0.34
	CR54	-2.48	0.25	1.36	2.1	1.33	1.4	0.43
	CR55	-2.11	0.25	1.05	0.4	1.03	0.2	0.49
	CR56	-0.54	0.22	1.02	0.2	1.14	0.7	0.47
	CR57	-2.36	0.25	0.90	-0.5	0.89	-0.4	0.47
	CR58	-1.75	0.24	1.12	0.7	1.04	0.2	0.47
Deep and analytical thinking								
Inductive reasoning	IR59	-1.06	0.23	0.71	-1.5	0.89	-0.5	0.58
	IR60	0.43	0.19	1.07	0.4	1.06	0.4	0.62
	IR61	-0.39	0.22	1.03	0.2	0.92	-0.4	0.63
	IR62	-0.17	0.21	1.30	1.4	1.16	0.9	0.61
	IR63	-0.49	0.22	1.35	1.6	1.12	0.6	0.59
	IR64	-0.25	0.21	1.60	2.6	1.36	1.7	0.51
Deductive reasoning	DR65	-0.54	0.22	0.99	0.0	1.13	0.7	0.59
	DR66	-0.17	0.21	0.98	0.0	0.94	-0.2	0.68
	DR67	-0.08	0.21	0.82	-0.9	0.75	-1.3	0.70
	DR68	-0.39	0.22	0.83	-0.8	0.68	-1.7	0.60
	DR69	-0.35	0.22	0.98	0.0	0.86	-0.7	0.54
	DR70	-0.12	0.21	0.74	-1.3	0.67	-1.8	0.65
Error analysis	EA71	-0.12	0.21	0.95	-0.2	0.97	-0.1	0.54
	EA72	-0.12	0.21	0.81	-1.0	0.78	-1.1	0.66
	EA73	-1.11	0.24	0.63	-2.1	0.60	-2.2	0.70
	EA74	-0.84	0.23	0.88	-0.5	0.86	-0.6	0.56
	EA75	-0.64	0.22	0.77	-1.2	0.75	-1.3	0.63
	EA76	-0.89	0.23	0.60	-2.3	0.56	-2.5	0.72
Constructing support	CS77	-0.12	0.21	0.71	-1.6	0.72	-1.5	0.62
	CS78	-0.84	0.23	0.78	-1.1	0.72	-1.5	0.58
	CS79	-0.04	0.20	0.83	-0.8	0.67	-1.8	0.65
	CS80	0.09	0.20	0.68	-1.8	0.65	-2.0	0.69
	CS81	-0.35	0.22	0.78	-1.1	0.79	-1.1	0.61
	CS82	1.55	0.16	0.92	-0.4	1.16	0.9	0.67
Analysing perspective	AP83	0.86	0.18	0.69	-1.9	0.69	-1.8	0.60
	AP84	0.89	0.17	0.66	-2.1	0.77	-1.3	0.65
	AP85	1.04	0.17	0.66	-2.1	0.82	-1.0	0.74
	AP86	1.24	0.17	0.39	-4.6	0.43	-3.9	0.77
	AP87	1.37	0.16	0.49	-3.6	0.52	-3.1	0.74
	AP88	0.39	0.19	0.87	-0.6	0.90	-0.5	0.66
Complex thinking strategies								
Problem-solving	PS89	1.55	0.16	0.81	-1.2	0.89	-0.6	0.73
	PS90	0.92	0.17	0.88	-0.6	0.83	-0.9	0.73

	PS91	1.07	0.17	0.78	-1.3	0.71	-1.7	0.76
	PS92	1.45	0.16	0.69	-2.0	0.70	-1.8	0.75
	PS93	1.34	0.16	0.72	-1.7	0.72	-1.6	0.74
	PS94	1.34	0.16	0.59	-2.8	0.60	-2.5	0.78
Making decisions	MD95	0.80	0.18	0.70	-1.8	0.71	-1.7	0.73
	MD96	0.92	0.17	0.65	-2.2	0.68	-1.8	0.72
	MD97	0.83	0.18	0.71	-1.7	0.66	-2.0	0.72
	MD98	1.50	0.16	1.09	0.6	1.22	1.2	0.68
	MD99	1.18	0.17	1.41	2.1	1.51	2.4	0.67
	MD100	0.83	0.18	1.19	1.1	1.10	0.6	0.67
Mean S.D		0.00 1.05	0.20 0.03	0.98 0.37	-0.3 1.8	1.00 0.42	-0.2 1.9	

Source: Winsteps output

3.1 Item Fit

The fit statistic, including infit and outfit MNSQ, z-standard values and point measure correlation (PMea Corr) values were shown in Table 2. All items in each dimension were in the range of positive PtMea Corr values measured according to the latent variables. From the External World View dimension, item AT42 was removed because MNSQ outfit value exceeds the range. Even item AT44 and item AT46 were out of MSNQ outfit range, and these items were modified by restructuring the sentences without changing its operational definition for each item. Item WT47 and WT48 from the classroom interaction dimension were removed permanently. The value of outfit MNSQ and z-standards for item WT47 and WT48 exceeded the maximum value range. Moreover, all items from Deep and Analytical thinking and Complex Thinking Strategies dimensions fit the Rasch model.

3.2 Reliability and Separation Index

The person reliability for this inventory was 0.97 while item reliability was 0.95. Both person and item reliability exceed 0.90, which showed high reliability and tremendously accepted (Bond & Fox, 2015; Linacre, 2015). High person reliability value proved the respondents' adequacy with various ability levels to respond to the inventory (Souza et al., 2017). The Cronbach Alpha (KR-20) value of 0.97 indicated an excellent internal consistency and extremely qualified for the next data collection (Bond & Fox, 2015). High item reliability proved the items could locate according to the latent variable accurately. Figure 1 presented the summary statistics of reliability and separation index.

Figure 1: The summary statistics of reliability and separation index

```
SUMMARY OF 76 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT    MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     150.2     53.0       1.21     .29       .99    -.2   1.00    -.2 |
| S.D.      20.2       .0       1.69     .02       .38    1.9    .42    1.9 |
| MAX.     203.0     53.0       6.14     .39      1.98    4.2   2.03    4.2 |
| MIN.     106.0     53.0      -2.08     .26       .28   -4.5    .24   -4.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .31 TRUE SD    1.67  SEPARATION  5.39  Person RELIABILITY  .97 |
```

```
|MODEL RMSE    .29 TRUE SD    1.67  SEPARATION  5.79  Person RELIABILITY  .97 |
| S.E. OF Person MEAN = .20                                                    |
-------------------------------------------------------------------------------
                DELETED:     10 Person
Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .97

     SUMMARY OF 53 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     215.5      76.0        .00     .24        .99    -.1   1.00    -.1 |
| S.D.      20.4        .0       1.18     .01        .24    1.4    .26    1.4 |
| MAX.     259.0      76.0       1.87     .26       1.66    3.4   1.80    3.8 |
| MIN.     181.0      76.0      -2.69     .22        .60   -2.9    .58   -2.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .25 TRUE SD    1.15  SEPARATION  4.59  Item   RELIABILITY  .95 |
|MODEL RMSE    .24 TRUE SD    1.15  SEPARATION  4.81  Item   RELIABILITY  .96 |
| S.E. OF Item MEAN = .16                                                      |
-------------------------------------------------------------------------------
                DELETED:      7 Item
UMEAN=.0000 USCALE=1.0000
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00
4028 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 5465.90 with 3898 d.f. p=.0000
Global Root-Mean-Square Residual (excluding extreme scores): .4781
```

Source: Winsteps output

The value of person separation index, 5.39 indicated that a set of items could discriminate the respondents into various measured of ability level. In terms of person separation index, the value exceeding more than 3.00 is considered excellent separation (Boone & Noltemeyer, 2017). Besides, the item separation index value of 4.59 more significant than 3.0 is desirable (Boone & Noltemeyer, 2017). The respondents were sufficient to affirm the hierarchy and distribution of items (Linacre, 2015).

3.3 Unidimensionality and Local Independence

The construct validity was measured by analysing the output items from the domain of HOTS questioning practice dimensionality. Rasch model was performed to examine the efficiency of the obtained data which fit the model. For this study, unidimensionality and local independence provided the analysis of items contribution towards the single domain being measured. Figure 2 showed the results of the analysis for unidimensionality based on the Standardised Residual variance.

Figure 2: The analysis of unidimensionality by the Standardized Residual variance.

```
      Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                -- Empirical --    Modeled
Total raw variance in observations     =       104.2 100.0%         100.0%
  Raw variance explained by measures   =        51.2  49.1%          48.5%
    Raw variance explained by persons  =        28.3  27.1%          26.8%
    Raw Variance explained by items    =        22.9  22.0%          21.7%
  Raw unexplained variance (total)     =        53.0  50.9% 100.0%   51.5%
    Unexplained variance in 1st contrast =       6.9   6.7%  13.1%
    Unexplained variance in 2nd contrast =       4.3   4.2%   8.2%
    Unexplained variance in 3rd contrast =       3.4   3.2%   6.3%
    Unexplained variance in 4th contrast =       2.6   2.5%   4.9%
```

Source: Winsteps output

Figure 2 shows that the raw variance measured obtained 49.1% and the Rasch model predicted was 48.5% for a single domain of measured. Thus, the construct validation empirically was approximately equivalent to the value predicted to the Rasch model. However, the Rasch model required minimum value, 40% for *Raw variance explained by measures* and 60% for excellent unidimensionality (Azrilah Abdul Aziz et al., 2014). The *unexplained variance in the 1st contrast* value should not exceed 15% for ideal measurement (Teman, 2018). Based on the results from Figure 2, the raw variance data of 49.1% and unexplained variance in the 1st contrast of 6.7% meet the criteria of the Rasch model requirement. In Rasch analysis, local dependence is defined to identify any correlation between the items (Bond & Fox, 2015). Item AP86, PS94, MD96 and MD99 were eliminated from the inventory because its standard residual correlations values exceeded more than 0.7. The higher value of standard residual correlations more than 0.70 indicated that these items are dependent and share the same traits (Linacre, 2015). Therefore, after unidimensionality and local independency analysis, all items which meet the Rasch model criteria for measuring HOTS questioning domain were valid.

3.4 Item Difficulty Level Based on the Dimensions

In Rasch analysis, item-person map illustrated the complexity of the items based on the logit scale to identify STEM teachers' HOTS questioning practice in one continuum of measurement (Bond & Fox, 2015).

Figure 3: The analysis of unidimensionality by the Standardized Residual variance.

```
       Person - MAP - Item
           <more>|<rare>
    7            +
                 |
                 |
              X  |
    6         X  +
                 |
                 |
                 |
    5            +
                 |
                T|
                 |
    4        XX  +
            XXX  |
            XXX  |
            XXX  |
    3        XX S+
                 |
              X  |
          XXXXX  |T
    2        XX  +
```

```
          XXXXX   |  AP87    CS82    MD98    PS89    PS92    PS93
            XXX   |
            XXX  M|S AP83    AP84    AP85    AT44    MD97    PS91
 1         XXXX   +  MD100   MD95    PS90
        XXXXXXX   |  AT43    IR60
         XXXXXX   |  AP88    AT46    WT49
           XXXX   |  AT41    AT45    CS80
 0       XXXXXX   +M CS77    CS79    DR67    DR70    EA71    EA72
              X   |  CS81    DR66    DR69    IR62    IR64
           XXXX  S|  CR56    DR65    DR68    EA75    IR61    IR63    WT50
             XX   |  CS78    EA74    EA76
-1            X   +  EA73    IR59
           XXXX   |S WT51
             XX   |
                  |  CR58
-2            X   +  CR53
                 T|T CR55
                  |  CR57    WT52
                  |  CR54
-3                +
           <less>|<frequ>
```

Source: Winsteps output

Figure 3 shows an item-person map which presented the location of respondents' ability and item difficulty for HOTS questioning practice. The symbol of X represents 76 respondents on the left of the item-person map, while the difficulty level of items on the right. The most challenging item located on the top on the map. Thus, item CS82 and PS89 presented the highest logit value of 1.55 which indicated that these respondents were least often practicing in questioning to develop a new prototype model from multiple STEM concepts and generate various ideas to solve daily life problems. Item CR54 showed that these respondents most frequently response to the students after the students answered the questions.

4. Discussion

Rasch model analysis contributes robust psychometric properties to investigate validity and reliability based on respondents' capability and item difficulties. After performing multiple steps in Rasch analysis for the pilot study, only 53 items over a total item of 60 were selected to be administered for the next actual data collection. In revising the scale for modifications of the items from the disordered thresholds, five response choices were collapsed into four response options.

Based on the item-person map, item CR54 was presented as the most frequently applied by most of these respondents. They usually give positive feedbacks towards students‘ response after questioning session. A review of HOTS questioning major studies confirmed that positive feedback towards students‘ response increased students‘ motivation in STEM learning session (McDonald, 2016; Wahono et al., 2020; Yang, 2017). The elements of recognition and wait-time from classroom interaction dimension, which most of the items have the logit values below -1.00, proved these respondents frequently performed. Listen carefully to students‘ response, direct eye contacts including positive body language are important to these STEM teachers to increase students‘ participation in questioning (Gul et al., 2014; Lim et al., 2020). The results show that in HOTS questioning session, the respondents always scrutinise and scaffold students‘ response by regularly rephrasing the question to provide the correct answer. These questioning methods follow the constructivism theory reported in the literature (Biggers, 2018; Calder, 2015).

Furthermore, inductive and deductive reasoning are parts of critical thinking that involve inquiry in learning STEM (Hill, 2016). From the element of the inductive reasoning process, these respondents invariably generate questions from scientific evidence from the experiment or data to form a conclusion or hypothesis based on items' locations (Abdullah et al., 2020; Lazuardini et al., 2019). Moreover, in the deductive reasoning process, STEM teachers continuously generate question from STEM theory, principles or law to generalise (Lee & Kinzie, 2012). Most of the current STEM education evidence proved that these STEM teachers had generated HOTS questions from inductive and deductive reasoning process (Abdullah et al., 2020; Ferguson, 2019; Kivunja, 2014). Similarly, from the error analysis element, the respondents often apply open questions to identify any misconceptions to the related concepts in STEM. These respondents usually cultivate their students to develop strong justification against the existing statement in STEM learning. These findings confirmed that error analysis in questioning allows students to improve their existing conception and reflect their thinkings (Ernst-slavit & Pratt, 2017; Keong et al., 2016; Wang, 2016). Most of the items from the element of constructing support demonstrated that STEM teachers typically ask students to justify their statement based on strong evidence from reliable sources (Heng et al., 2015). Several works of literature have confirmed that constructing support in STEM learning could enhance students' critical thinking skills (Chen et al., 2019; Wahono et al., 2020).

Subsequently, according to the item-person map, most of the items from the elements of authenticity, analysing perspective, problem-solving, and decision-making are located above the measure of the mean (average) of the item difficulty. From the elements of authenticity, the results illustrated that some of the respondents struggled to construct questions which could relate subject content with real-life situations and other disciplines. A few studies have found that some STEM teachers have difficulty generating authentic questions through interdisciplinary approach (Baharin et al., 2018; Kelley & Knowles, 2016; McDonald, 2016; Ring et al., 2017). Next, from the element of analysing perspective, respondents rarely construct questions for students to evaluate their response and elaborate with good reasoning skills based on STEM concepts. Recent studies have confirmed that STEM teachers infrequently review the STEM concepts with current issues (Baharin et al., 2018; Rinke et al., 2016). Besides, all items from the problem-solving element located exceed the mean value of item measure. It signified that the respondents seldom convey questions to clarify the problem and generate various ideas to solve the problem from multiple STEM concepts. Several studies have claimed that most STEM teachers have deficient experienced in generating questions for students to solve the problem by inventing engineering design from interdisciplinary STEM learning (Chiang et al., 2020; Shernoff et al., 2017). From the results, items from the making decisions element proved that most of the respondents rarely practice generating questions for students to apply scientific reasoning ability in choosing the best solution, especially in the design task. This result is supported by the findings that integrated exact STEM concepts rarely applied in engineering design task from most STEM teachers (Estapa & Tank, 2017; Ring et al., 2017).

Additionally, as predicted, the obtained result of items analysis in term of validity and reliability from the Rasch model for measuring the domain of HOTS questioning practice among the respondents was significant. However, surprisingly, some of the results of analysed items described that most respondents were insufficient to apply HOTS questioning from integrated STEM concepts. This study has its limitation, especially regarding the imbalance of demographic background in teaching experienced and academic qualifications among the respondents. This imbalance issues of demographic information among the respondents should be anticipated and addressed for better justification. The results demonstrated in this study provide a new perspective on the solution to improve HOTS questioning practice in interdisciplinary learning of integrated STEM. Hence, some modification is required for future studies regarding the development of items. These items should be based on the latest issues of STEM teachers' pedagogical skills in delivering HOTS questions, particularly from interdisciplinary STEM integration.

5. Conclusion

In conclusion, this study contributes a successful and robust approach in developing a valid and reliable inventory to measure STEM teachers HOTS questioning practice. This study's findings provide the first preliminary evidence for measuring STEM teachers HOTS questioning application in interdisciplinary STEM integration based on four crucial dimensions such as External World View, Classroom Interaction, Deep and Analytical Thinking, and Complex Thinking Strategies. This study's contribution will provide essential evidence to improve the higher secondary STEM education system in Malaysia, especially in HOTS questioning (MOE, 2018; Schleicher, 2018). This study will help policymakers and school administrators organise and plan better STEM pedagogical teaching training in the future.

Acknowledgment

The authors would like to thank Universiti Teknologi Malaysia and Sponsorship Division, Ministry of Education, Malaysia for providing full sponsor scholarship that support this study to be carried out.

References

Abdullah, A. H., Misrom, N. S., Kohar, U. H. A., Hamzah, M. H., Ashari, Z. M., Ali, D. F., Samah, N. A., Tahir, L. M., & Rahman, S. N. S. A. (2020). The Effects of an Inductive Reasoning Learning Strategy Assisted by the GeoGebra Software on Students' Motivation for the Functional Graph II Topic. *IEEE Access*, *8*(August), 143848–143861.

Abdullah, A. H., Mokhtar, M., Halim, N. D. A., Ali, D. F., Tahir, L. M., & Kohar, U. H. A. (2017). Mathematics teachers' level of knowledge and practice on the implementation of higher-order thinking skills (HOTS). *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(1), 3–17.

ASM. (2018). *Science Outlook*.

Aziza, M. (2018). An Analysis of A Teacher ' s Questioning Related to Students ' Responses and Mathematical Creativity in An Elementary School in The UK. *International Electronic Journal of Elementary Education, 10*(4), pp. 475–487.

Azrilah Abdul Aziz, Muhammad Shahar Jusoh, Omar, A. R., Mohd Haris Amlus, & Tuan

Salwani Awang. (2014). Construct Validity: A Rasch Measurement Model Approaches. *Journal of Applied Science and Agriculture*, *9*(12), pp. 7–12.

Baharin, N., Kamarudin, N., & Manaf, U. K. A. (2018). Integrating STEM Education Approach in Enhancing Higher Order Thinking Skills. *International Journal of Academic Research in Business and Social Sciences*, *8*(7), pp. 810–821.

Biggers, M. (2018). Questioning Questions: Elementary Teachers' Adaptations of Investigation Questions Across the Inquiry Continuum. *Research in Science Education*, *48*(1), pp.1–28.

Bond, T. G. (2003). Validity and assessment : a Rasch measurement perspective.*Metodologia de las Ciencias del Comportamiento 5*(2), pp. 179–194.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model (Fundamental Measurement in the Human Sciences)* (3rd ed.). Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: Why,when,and how? *CBE Life Sciences Education*, *15*(4), pp. 1–7.

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, *4*(1), pp. 1–13.

Calder, N. (2015). Student wonderings: scaffolding student understanding within student-centred inquiry learning. *ZDM - Mathematics Education*, *47*(7), pp.1121–1131.

Çalik, M., Ebenezer, J., Özsevgeç, T., Küçük, Z., & Artun, H. (2015). Improving Science Student Teachers' Self-perceptions of Fluency with Innovative Technologies and Scientific Inquiry Abilities. *Journal of Science Education and Technology*, *24*(4), pp. 448–460.

Chen, Y., Benus, M. J., & Hernandez, J. (2019). Managing uncertainty in scientific argumentation. *Science Education*, *103*(5), pp.1235–1276.

Chiang, F. K., Chang, C. H., Wang, S., Cai, R. H., & Li, L. (2020). The effect of an interdisciplinary STEM course on children's attitudes of learning and engineering design skills. *International Journal of Technology and Design Education*, pp.100.

Deprez, J., Van de Velde, D., De Cock, M., Hellinckx, L., Van Petegem, P., Struyf, A., Boeve-de Pauw, J., Dehaene, W., Struyven, K., Knipprath, H., Goovaerts, L., Depaepe, F., Thibaut, L., De Meester, J., Langie, G., De Loof, H., & Ceuppens, S. (2018). Integrated STEM Education: A Systematic Review of Instructional Practices in Secondary Education. *European Journal of STEM Education*, *3*(1), pp.1–12.

Ernst-slavit, G., & Pratt, K. L. (2017). Teacher questions : Learning the discourse of science in a linguistically diverse elementary classroom. *Linguistics and Education*, 40, pp.1–10.

Estapa, A. T., & Tank, K. M. (2017). Supporting integrated STEM in the elementary classroom : a professional development approach centered on an engineering design challenge. *International Journal of STEM Education*, *4*(6), pp.1-16

Ferguson, J. P. (2019). Students are not inferential-misfits: Naturalising logic in the science classroom. *Educational Philosophy and Theory*, *51*(8),pp. 852–865.

Gardner, M., & Tillotson, J. W. (2019). Interpreting Integrated STEM: Sustaining Pedagogical

Innovation Within a Public Middle School Context. *International Journal of Science and Mathematics Education*, *17*(7), pp.1283–1300.

Geldenhuys, M., Bosch, A., & Geldenhuys, M. (2019). A Rasch Adapted Version of the 30-Item Bem Sex Role Inventory (BSRI) A Rasch Adapted Version of the 30-Item Bem Sex Role Inventory (BSRI). *Journal of Personality Assessment*, *0*(0), pp.1–12.

Gul, R. B., Khan, S., Ahmed, A., Cassum, S., Saeed, T., Parpio, Y., & Profetto-mcgrath, J. (2014). *Enhancing Educators ' Skills for Promoting Critical Thinking in Their Classroom Discourses : A Randomized Control Trial*. *26*(1), pp.37–54.

Heng, L. L., Surif, J., Seng, C. H., & Ibrahim, N. H. (2015). Mastery of scientific argumentation on the concept of neutralization in chemistry: A Malaysian perspective. *Malaysian Journal of Learning and Instruction*, *12*(1), pp. 85–101.

Iksan, Z. H., & Daniel, E. (2015). Emerging Model of Questioning through the Process of Teaching and Learning Electrochemistry. *International Education Studies*, *8*(10).

Irekpita, M., Mohd, S., Samsudin, A., & Yakob, N. (2020). Effect of design thinking approach on students ' achievement in some selected physics concepts in the context of STEM learning. *International Journal of Technology and Design Education*, pp. 1-28.

Kelley, T. R., & Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education*, *3*(11), pp. 1-14.

Keong, K., Ong, A., Hart, C. E., & Chen, P. K. (2016). Promoting Higher-Order Thinking Through Teacher Questioning: a Case Study of a Singapore Science Classroom. *New Waves Educational Research & Development*, *19*(1), pp.1–19.

Kivunja, C. (2014). Innovative Pedagogies in Higher Education to Become Effective Teachers of 21st Century Skills: Unpacking the Learning and Innovations Skills Domain of the New Learning Paradigm. *International Journal of Higher Education*, *3*(4), pp.37–48.

Koskey, K. L. K., Sondergeld, T. A., Stewart, V. C., & Pugh, K. J. (2016). Applying the Mixed Methods Instrument Development and Construct Validation Process : The Transformative Experience Questionnaire. *Journal of Mixed Method Research*, pp. 1-28

Lazuardini, D. A., Wilujeng, I., & Kuswanto, H. (2019). Optimizing Senior High School Students Creative Thinking Skills of Optical Devices through Inductive Learning Models Assisted by e-Mind Map. *Journal of Physics: Conference Series*, *1233*(1).

Lee, Y., & Kinzie, M. B. (2012). Teacher question and student response with regard to cognition and language use. *Instructional Science*, *40*(6), pp. 857–874.

Lim, W., Lee, J. E., Tyson, K., Kim, H. J., & Kim, J. (2020). An Integral Part of Facilitating Mathematical Discussions: Follow-up Questioning. *International Journal of Science and Mathematics Education*, *18*(2), pp. 377–398.

Linacre, J. . (2015). *Winsteps Rasch Measurement Computer Program User's Guide*. Beaverton, US.

Lo, C., Liang, W., Hang, L., Wu, T., Chang, Y., & Chang, C. (2015). A psychometric assessment of the St . George ' s respiratory questionnaire in patients with COPD using rasch model analysis. *Health and Quality of Life Outcomes*, *13*(131), pp. 1–16.

Mahmud, S. N. D., Nasri, N. M., Samsudin, M. A., & Halim, L. (2018). Science teacher education in Malaysia: challenges and way forward. *Asia-Pacific Science Education*, *4*(1), pp.1-12.

McDonald, C. (2016). STEM Education: A Review of the Contribution of the Disciplines of Science, Technology, Engineering and Mathematics. *Science Education International*, *27*(4), pp. 530–569.

MOE. (2018). *Annual Report 2018 Malaysian Educational Blueprint 2013-2025*.

Norhayati, M. N., Fatin Imtithal, A., & Nor Akma, M. J. (2020). Psychometric properties of the Malay version of the Women's Views of Birth Labour Satisfaction Questionnaire using the Rasch measurement model: a cross sectional study. *BMC Pregnancy and Childbirth*, *20*(1), pp.1–8.

Pallant, J. F. (2020). *SPSS Survival Manual: A step by step guide to data analysis using IBM SPSS* (7th ed.). Routledge.

Perera, L. D. H., & Asadullah, M. N. (2019). Mind the gap: What explains Malaysia's underperformance in Pisa? *International Journal of Educational Development*, *65*(August 2018), pp. 254–263.

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in Health*, *18*(1), pp. 25–34.

Ping, I. L. L., Halim, L., & Osman, K. (2019). Explicit Instruction of Scientific Argumentation in Practical Work: A Feasibility Study. *Creative Education*, *10*(06), pp.1205–1229.

Purdum-Cassidy, B., Nesmith, S., Meyer, R. D., & Cooper, S. (2015). What are they asking? An analysis of the questions planned by prospective teachers when integrating literature in mathematics. *Journal of Mathematics Teacher Education*, *18*(1), pp. 79–99.

Rahman, N. B. A., Yusop, N. A. B. M., Sulaiman, N. F. B., Suhaimi, A. B., Hashim, R. B., Ahamed, A. B. B., Ismail, B. L. B., & Said, A. Bin. (2020). Validity and reliability of measurement instruments using the rasch model: Preschool teacher readiness for STEM implementation in Malaysia. *International Journal of Innovation, Creativity and Change*, *11*(11), pp. 112–131.

Ramdiah, S., Royani, M., Malang, U. M., & Kleij, D. (2019). Understanding , Planning , and Implementation of HOTS by Senior High School Biology Teachers in Banjarmasin-Indonesia. *International Journal of Instruction*, *12*(1), pp. 425–440.

Ring, E. A., Dare, E. A., Crotty, E. A., & Roehrig, G. H. (2017). The evolution of teacher conceptions of STEM education throughout an intensive professional development experience. *Journal of Science Teacher Education*, *28*(5), pp. 444–467.

Rinke, C. R., Gladstone-Brown, W., Kinlaw, C. R., & Cappiello, J. (2016). Characterizing STEM Teacher Education: Affordances and Constraints of Explicit STEM Preparation for Elementary Teachers. *School Science and Mathematics*, *116*(6), pp. 300–309.

Ritz, J. M., & Fan, S. C. (2015). STEM and technology education: international state-of-the-art. *International Journal of Technology and Design Education*, *25*(4), pp. 429–451.

Rosidin, U., Suyatna, A., & Abdurrahman, A. (2019). A combined HOTS-based assessment/STEM learning model to improve secondary students' thinking skills: A development and evaluation study. *Journal for the Education of Gifted Young Scientists*, *7*(3), pp. 435–448.

Saribas, D. (2015). Pre-service elementary teachers' preferences and competencies in relation to inquiry-based instruction and high quality questions. *Anthropologist*, *22*(2), pp.227–236.

Schleicher, A. (2018). *PISA 2018 Insights and Interpretations*.

Shernoff, D. J., Sinha, S., Bressler, D. M., & Ginsburg, L. (2017). Assessing teacher education and professional development needs for the implementation of integrated approaches to STEM education. *International Journal of STEM Education*, *4*(1).

Souza, A. C. de, Alexandre, N. M. C., Guirardello, E. de B., Maria, N., Alexandre, C., Campinas, U. E. De, & Enfermagem, F. De. (2017). Psychometric properties in instruments evaluation of reliability and validity. *Applications of Epidemiology*, *26*(3), pp.1-10.

Tajudin, N. M., Puteh, M., & Adnan, M. (2018). Guiding Principles to Foster Higher Order thinking skills in teaching and learning of mathematics. *International Journal of Engineering and Technology(UAE)*, *7*(4), pp. 195–199.

Teman, E. (2018). A Rasch Analysis of the Statistical Anxiety Rating Scale. *Journal of Applied Measurement, 14*(4), pp. 414-434.

Tengku Ariffin, T. F., Bush, T., & Nordin, H. (2018). Framing the roles and responsibilities of Excellent Teachers: Evidence from Malaysia. *Teaching and Teacher Education*, *73*, pp.14–23.

Wahono, B., Lin, P. L., & Chang, C. Y. (2020). Evidence of STEM enactment effectiveness in Asian student learning outcomes. *International Journal of STEM Education*, *7*(1), 1–18. https://doi.org/10.1186/s40594-020-00236-1

Wang, A. (2016). Exploring the impact of teacher experience on questioning techniques in a Knowledge Building classroom. *Journal of Computers in Education*. pp. 1- 16.

Yang, H. (2017). A Research on the Effective Questioning Strategies in Class. *Science Journal of Education*, *5*(4), pp.158-163.

Yen, T. S., & Halili, S. H. (2015). Effective Teaching of Higher-Order Thinking (Hot) in Education. *The Online Journal of Distance Education and E-Learning (TOJDEL)*, *3*(2), pp. 41–47.

Yeung, S. yin S. (2015). Conception of teaching higher order thinking: perspectives of Chinese teachers in Hong Kong. *Curriculum Journal*, *26*(4), pp.553–578.

Zanon, C., Hutz, C. S., Yoo, H. H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*. 29(18), pp. 1-

18.

Zeegers, Y., & Elliott, K. (2019). Who's asking the questions in classrooms? Exploring teacher practice and student engagement in generating engaging and intellectually challenging questions. *Pedagogies*, *14*(1), pp. 17–32.

Zhang, L. (2016). Is Inquiry-Based Science Teaching Worth the Effort?: Some Thoughts Worth Considering. *Science and Education*, *25*(7–8), pp. 897–915.