Rasch Rating Scale Modelling of the Arabic Version of the Critical Thinking Disposition Scale

Sharif Alsoudi* & Yousef Abu Shindi*[1]
*A'Sharqiyah University, Sultanate of Oman
*[1] Sultan Qaboos University, Sultanate of Oman

Received: 25/11/2022 Modified:13/06/2023 Accepted: 13/06/2023

Abstract: The tendency to think critically is the motivation of an individual for using critical thinking when faced with a problem that requires a solution, making a decision or evaluating an idea. This study used the Rasch Rating Scale Model (RSM) analysis to examine a set of psychometric properties of an Arabic version of the Critical Thinking Disposition Scale (EMI): items fit, unidimensionality, local independence, equal-item-discriminations, gender differential item functioning, reliability and separation indicators and scale calibration. The findings indicated that EMI showed good compatibility with the RSM as all the items matched the model except for item 11. In addition, the assumptions of the Rasch model which were unidimensionality, local independence, and equal-item-discriminations were realized. The scale had excellent reliability for persons and good reliability for items. The scale showed good separation indicators for items, and excellent separation indicators for persons. The items did not show differential gender performance. The distances be-tween the response categories were appropriate, and the category measurements showed a consistent in-crease.

Keywords: critical thinking disposition, item response theory, psychometrics, rasch model.

تطوير نسخة عربية من مقياس النزعة للتفكير الناقد باستخدام نموذج راش للتقدير

شريف السعودي* و يوسف أبو شندي
*جامعة الشرقية، سلطنة عُمان
جامعة السلطان قابوس، سلطنة عُمان

اسُتلم بتاريخ: 2022/11/25 **عُدل بتاريخ: 2023/06/13** **قُبل بتاريخ: 2023/06/13**

الملخص: هدفت الدراسة إلى توظيف نموذج راش لفحص الخصائص السيكومترية لنسخة عربية من مقياس النزعة للتفكير الناقد، من خلال مجموعة من المؤشرات، وهي: مواءمة الفقرات، وأحادية البعد، والاستقلال المحلي، وتساوي القدرة التمييزية للفقرات، والأداء التفاضلي للجنس، ومؤشري الثبات والفصل، ومعايرة المقياس. استخدمت عينة مكونة من 251 من طلبة جامعة الشرقية في سلطنة عمان. أشارت التحليلات إلى أن المقياس أظهر توافقاً جيداً مع نموذج راش. إذا طابقت جميع فقراته للنموذج باستثناء الفقرة 11. بالإضافة إلى تحقق جميع افتراضات نموذج راش، وهي: أحادية البعد، والاستقلال المحلي، وتساوي التمييز بين الفقرات. وكان للمقياس مؤشرات ثبات مرتفعة للأفراد وجيدة للفقرات. وكذلك مؤشرات فصل جيدة للفقرات، وممتازة للأفراد. ولم تظهر الفقرات أداءً تفاضلياً للجنس. وكانت المسافات بين فئات الاستجابة مناسبة، وتتقدم قياسات الفئات بشكل متسق.

الكلمات المفتاحية: الخصائص السيكومترية، النزعة للتفكير الناقد، نظرية الاستجابة للفقرة، نموذج راش.

*Sharif.alsoudi@asu.edu.om

Introduction

Thinking is one of the most important distinguishing properties of human beings. It is also one of the most important reasons for development and improvement in life. Critical thinking represents the most important types and forms of thinking which are necessary to develop knowledge effectively. It is also one of the most important skills of the twenty-first century. It is the right way of thinking (Lyutykh, 2009). It represents the participation of the individual in assuming responsibility for his/her actions in everyday life (Bowell & Kemp, 2005). Page (2007) argues that critical thinking is associated with thinking about higher cognitive levels in the Bloom taxonomy which includes analysis, synthesis and evaluation. Hurst (1999) emphasizes the need to develop critical thinking skills and include them in the curricula of all academic and educational systems. In addition, he believes that students should pass some courses in critical thinking before they graduate.

Critical thinking allows people to use their mental energy and interact effectively and strongly with the environment in which they live. It also enables them to face the complexities of life (Profetto, 2003). Although critical thinking is important in people's life and society, the possession of critical thinking skills alone is not enough, as one must have the desire or disposition to use and employ these skills (Stedman & Andenoro, 2007). The concept of critical thinking disposition refers to an individual's motivation to use critical thinking when faced with a problem that requires a solution, making a decision or evaluating an idea (Paul & Elder, 2014). In addition to focusing on the affective aspects of thinking, which appear in the form of tendencies, trends and mental habits responsible for activating the process of acquiring knowledge, the mind is directed towards good thinking through a set of behaviors that achieve these tendencies and desires (Kwon et al., 2007). People with a tendency to think critically use critical thinking skills more efficiently. They also seem more willing and desirable to practice critical thinking (Yüksel & Alci, 2012).

Although educational institutions include programs designed to develop students' critical thinking capacity, critical thinking education has not been offered on a systematic basis in many of these institutions. A 2005 report by the Association of American Colleges and Universities indicated that only 6% of the universities performed satisfactorily in teaching critical thinking (Paul et al., 1997). This is because teaching critical thinking involves many difficulties, one of which is the lack of an objective and effective assessment tool to measure whether students' critical thinking is weak or strong (Ennis, 2003; Halpern, 2003; Norris, 2003). Research reports students' critical thinking as a crucial factor in problem-solving and decision-making (Heidari & Ebrahimi, 2016; Lismayani et al., 2017). Other researchers also reported that undergraduate students who were high in critical thinking were found to be better in stress management and academic achievement (Mahal et al., 2015; Taghva et al., 2014), and lower in academic procrastination (Goroshit, 2018). Critical thinking was also found to be an effective predictor of social adjustment (Hashemiannejad et al., 2016). It is, therefore, necessary to provide scales for critical thinking with good psychometric properties, the results of which can be relied upon in assessing this trait in individuals.

The Critical Thinking Disposition Scale (EMI) developed by Ricketts & Rudd (2005) is one of the key tools for assessing the level of critical thinking in students, especially university students. It aims to measure the tendencies and motivation of individuals to practice critical thinking in different situations that require solutions or decision-making (Lai, 2011). The psychometric properties of the scale have been verified by numerous studies (e.g., Demircioğlu & Kilmen, 2015; Irani et al., 2007; Karami & Shakurnia, 2020; Lee, 2009; Rincker, 2014; SK & Halder, 2020; Stedman & Andenoro, 2006) and in different societies and countries (e.g., USA, Turkey, India, Malaysia, and Iran). These studies aimed to verify the validity, reliability and factorial structure of the scale, and all their results showed that the scale has good psychometric properties.

The previous studies relied on the Classical Test Theory (CTT) to verify the properties of the scale, but none of them verified the properties of the scale according to Item Response Theory (IRT). IRT is a new and good approach to the development of measuring instruments (Wilson, 2005) which came as an improvement on the CTT approach. IRT provides rich information about the properties of the scale and has many advantages compared to CTT (Embretson, 1996). IRT can be used to evaluate the psychometric properties of items in an existing scale to optimize the scale when necessary and to evaluate the performance of the short scale. When used appropriately, IRT modeling can produce accurate, valid and relatively concise scales (Edelen & Reeve, 2007). An additional advantage of IRT is that item parameters such as difficulty and discrimination estimated in one sample of the population can be linearly converted into estimates of those parameters in another sample of the same population. IRT is different from CTT in which the estimation of section parameters depends on the characteristics of the group in which they are estimated (Baker, 2001).

IRT has three main types of models: One-parameter (1-PL) which is called Rasch analysis and estimates only the item which is difficulty parameter, Two-parameter (2-PL) which estimates the item difficulty and discriminant parameters, and Three-parameter (3-PL) which calculates the item difficulty, item discriminant, and pseudo-guessing parameters (Baker, 2001). The purpose of this study is to employ one model of IRT which is the Rasch Rating Scale Model (RSM).

It became clear to the researcher through the review of the previous studies that there has been no such study in the Arab context. In fact, there has been no study that measured the critical thinking inclinations of Omani undergraduate students. Therefore, this study aims to measure the critical thinking inclinations of undergraduate students in Oman and provide a new scale that can measure critical thinking disposition based on IRT, specifically the Rasch model.

Rasch Rating Scale Model (RSM)

The Rasch Rating Scale Model (RSM) is a model based on IRT. It is one of the simplest of these models and is widely used in evaluating item quality (Magno, 2009). Under the umbrella of Rasch analysis, four models can be used, depending on the items' response pattern. These are (1) Rasch dichotomous model (i.e. for items with two answers), (2) Andrich Rating scale model (i.e., for polytomous Likert scale type where the responses have the same response weights), (3) Masters Partial Credit Model (i.e., polytomous items that can be partially correct and the items have different response weights; e.g., math problems where several operations can have different marks), and (4) Grouped Model (i.e., for polytomous different questions with different response weights) (Von Davier, 2016). RSM postulates a set of assumptions, the most important of which are unidimensionality, equal-item-discriminations and low guess. RSM estimates only one parameter which is the difficulty parameter (Sick, 2009). RSM uses the raw score to estimate trait ability, and places trait ability on the same scale (i.e., logit scale) with item difficulty estimates. The overlap between ability distributions and item difficulty on the logarithm scale can then be examined to determine whether or not the instrument is suitable for the selected sample. If the measurement tool is working correctly (i.e., the IRT model fits the data), the estimation of its item parameters does not depend on the specific sample used, and unbiased estimates of item parameters can be obtained from unrepresented samples (Embretson & Reise, 2000).

RSM assumes that the responses on the measuring instrument are at the ordinal level and therefore do not require the normal distribution of data. In addition, with small sample sizes, the Rasch model provides more consistent estimates of parameters when compared to 2 PL or 3PL (Kim & Kyllonen, 2006). RSM can convert nonlinear raw data into a linear scale once the instrument items fit the model well (Boone, 2016). Accordingly, researchers can provide a meaningful explanation of their instrument scores.

RSM is suitable for polytomous data collected from the Likert scale (Andrich, 2005). RSM describes the probability of a person n in the rating-scale category x on a particular item i through the following equation:

$$P(X_{ni} = x) = \frac{exp \sum_{k=0}^{x}[\beta_n - (\delta_i + \tau_k)]}{\sum_{x=0}^{m} exp \sum_{k=0}^{x}[\beta_n - (\delta_i + \tau_k)]}, x = 0,1,\dots,m$$

Where p is the probability that person n is observed in the rating-scale category x on item i, which has m + 1 rating-scale categories, and (βn): raters' perception of their ratees, (δi): the item's endorsability, and (τk): a set of threshold parameters. The RSM assumes that the threshold structure is fixed across items (Andrich, 2010).

Methods

Participants

The participants in the present study were 324 students from the Department of Education enrolled in the Classroom Measurement and Evaluation Course from the bachelor's and Diploma in Educational Qualification at A'Sharqiyah University, Sultanate of Oman. Approval was obtained to apply the scale to the participating students from the Unit of Research Ethics and Biosafety Committee (UREBC) at A'Sharqiyah University. The link to the scale was sent to all the participating students via e-mail who were asked to respond to it optionally. 251 students responded (29% male, 71% female), (54% bachelor's, 46% Diploma in Educational Qualification).

Instrument

The original Critical Thinking Disposition Assessment (EMI) scale was developed by Ricketts and

Rudd in 2005. The responses were assessed via a five-point Likert scale (Strongly Agree5, Strongly Disagree1), consisting of 26 positive items. Sub-dimensions of the scale were Engagement 11 items (e.g., "I enjoy finding answers to challenging questions"), Cognitive Maturity 7 items (e.g., "I enjoy learning about many topics"), and Innovativeness 8 items (e.g., "I consider how my own biases affect my opinions"). The overall score on the scale ranges between (26-130). A higher score on the scale indicated a higher disposition to critical thinking. Individuals who were disposed to engagement believed that one should always think well and seek opportunities to use their thinking skills in reasoning, problem-solving, and decision-making. Individuals who were cognitively mature were aware that many problems they encountered were more complicated than they initially seemed. Individuals who were innovative were described as being “hungry to learn”. By examining the guidebook of the scale, the reliability coefficients of the sub-dimensions were 0.90, 0.78 and 0.79, respectively. The Cronbach-alpha reliability coefficient for the total scale was 0.93.

The linguistic equivalence of the scale was examined by translating the items of the scale into Arabic, then translating them back into English to verify the validity of the translation which was pr sented to 8 arbitrators specializing in educational psychology measurement and evaluation to ensure the accuracy and clarity of the items. The corrected Pearson correlation coefficients between the items and the total score on the scale ranged between (0.31 - 0.67), except for item 11 which had a low correlation coefficient 0.14. The reliability coefficient of the sub-dimensions for the Arabic version were 0.84, 0.75 and 0.77, respectively. The Cronbach-alpha reliability coefficient was 0.89. The intercorrelations among subscales were high and homogeneous (0.54 to 0.67) and they indicated the existence of a general factor across dimensions.

Data Analysis

To verify the characteristics of the Arabic version of the EMI scale, Rasch analysis was used according to the Andrich Rating scale model for the polytomous responses. Winsteps program version 5.3.1.0 with the Joint Maximum Likelihood Estimation (JMLE) method was used to test the scale properties: item fit, unidimensionality, local independence, equal-item-discriminations, gender differential item functioning, reliability and separation indicators and scale calibration. The SPSS (version 28) was used to calculate the descriptive statistics of the participants' data and to conduct the exploratory factor analysis (EFA).

Results

Descriptive Statistics

Table 1 contains the descriptive statistics of the overall score on the EMI scale. The mean values indicate that the students' performance was close according to their gender and academic program. There were no significant differences between both males and females and their academic qualifications (bachelor’s and diploma). The skewness and kurtosis values were all close to zero, and the Shapiro test values were non-significant. This indicated that the distribution of the students' scores was normal

Table 1. Descriptive statistic

	Gender		Program		Total
	Male	Female	BA	High diploma	
N	73	178	136	115	251
Mean	3.98	3.94	3.92	3.97	3.96
SD	0.54	0.58	0.58	0.53	0.57
t (Sig)	0.51 (0.31)		-0.71 (0.24)		
Skewness	0.10	0.03	0.11	0.12	0.06
Kurtosis	-0.22	-0.09	-0.12	-0.21	-0.16
Shapiro (Sig)	0.93 (0.36)	0.99 (0.40)	0.99 (0.67)	0.98 (0.52)	0.99 (0.44)

The results of the EFA revealed that there were 7 factors which had an eigenvalue higher than 1, respectively (6.28, 1.65, 1.35, 1.22, 1.20, 1.07, and 1.04) which represented (25.11%, 6.60%, 5.42%, 4.89%, 4.78%, 4.26%, and 4.18%) of the overall variance in the scale, respectively. All the items were correlated with the first factor by a loading coefficient greater than (0.30). This indicates that the scale was one-dimensional (Domain factor) according to one-dimensional criteria (i.e., the difference in the eigenvalue between the first and second factors was more than double, the proportion of variance explained by the first factor was greater than 20%, and all the items loaded on the first factor) (Harlow, 2005). Therefore, a Rasch analysis was conducted for the whole scale, rather than at the dimensional level.

Item Fit

Item fit refers to the analysis of the suitability of the Rasch measurement models for each item of the scale (Ariffin, 2008). The fit criteria are as follows:

Mean square (MNSQ). Infit and outfit are used to determine the discrepancy between the statistical model and the observed data (Gustafson, 1980). MNSQ values range from zero to infinity, and 1 is the ideal value. The value is considered acceptable and appropriate if it falls between 0.5-1.5 (Linacre, 2012). According to the data used in the current study and shown in Table 2, all MNSQ values ranged from 0.5 to 1.5, except for item 11 "I am likely to change my opinion when I am given new information that conflicts with my current opinion". It exceeded the range set for this test with MNSQ of 1.59 and 1.61 for infit and outfit respectively.

Table 2. Fit statistics of measurement items

Item	Measure	Model S.E.	INFIT		OUTFIT		PTMEA	
			MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.
11	-0.11	0.09	1.59	5.53	1.61	6.01	0.20	0.44
7	0.08	0.09	1.48	5.12	1.50	4.66	0.49	0.45
16	0.46	0.08	1.48	4.98	1.49	4.84	0.37	0.47
15	0.67	0.08	1.50	5.02	1.47	4.96	0.44	0.48
21	0.01	0.09	1.37	3.49	1.28	2.79	0.46	0.45
26	-1.28	0.11	1.15	1.55	1.09	0.89	0.35	0.36
1	-0.77	0.10	1.02	0.21	1.13	1.37	0.38	0.40
6	1.12	0.08	1.07	0.83	1.12	1.40	0.41	0.51
13	-0.05	0.09	1.06	0.66	1.08	0.88	0.41	0.44
10	-0.69	0.10	1.05	0.51	0.97	-0.30	0.51	0.40
3	0.70	0.08	1.02	0.23	1.02	0.25	0.42	0.49
2	-0.30	0.10	1.01	0.13	1.00	0.06	0.32	0.43
20	0.39	0.09	0.96	-0.44	1.00	0.00	0.37	0.47
25	-0.21	0.09	0.93	-0.66	0.90	-1.08	0.53	0.43
12	0.14	0.09	0.92	-0.84	0.89	-1.18	0.55	0.45
9	-0.16	0.09	0.86	-1.47	0.91	-0.97	0.41	0.44
22	-0.96	0.11	0.88	-1.34	0.83	-1.88	0.52	0.38
8	0.25	0.09	0.83	-1.92	0.85	-1.67	0.39	0.46
14	0.07	0.09	0.84	-1.79	0.84	-1.75	0.50	0.45
23	-0.79	0.10	0.84	-1.79	0.82	-1.99	0.51	0.40
24	-0.29	0.09	0.80	-2.15	0.80	-2.18	0.56	0.43
17	0.24	0.09	0.77	-2.61	0.78	-2.56	0.48	0.46
4	-0.33	0.10	0.72	-3.20	0.72	-3.25	0.60	0.43
18	0.27	0.09	0.67	-4.02	0.67	-4.00	0.54	0.46
5	0.92	0.08	0.60	-5.40	0.61	-5.17	0.49	0.50
19	0.63	0.08	0.61	-4.97	0.60	-5.12	0.57	0.48

The productive ZSTD value. This value ranges from -2.00 to +2.00 (Bond & Fox, 2007). It can be ignored if the MNSQ value is acceptable (Linacre, 2005). In the current analysis, there were a set of items outside the acceptable range of ZSTD values, but all of them met the MNSQ test, except for item 11 which did not meet both criteria.

The standard error value S.E. For the current data, this value ranged between 0.08 and 0.11, indicating the element of accuracy in the estimation (Linacre, 2005). Fisher (2007) considered the range of error values to be excellent.

PTMEA CORR value. This value refers to the point-measure correlation between scored responses and ability measures to determine how well the responses to the item are compatible with the abilities of the participants. A higher category is expected to have a strong positive correlation with ability, and a lower category is expected to have a strong negative correlation with ability (Linacre, 2012). Table 2 shows that all PTMEA CORR values were larger than the acceptable minimum of 0.30 (Wu & Adam, 2007) except for item 11, in which the value of PTMEA CORR was 0.20. It was clear that item 11 did not achieve three criteria of item fit, and was, therefore, omitted from the scale.

Unidimensionality

The concept of unidimensionality refers to scale items that measure only one structure and represent an indication of the validity of scale construction (Hambleton & Swaminathan, 1985). Unidimensionality was detected in the current study using the principal component analysis (PCA). The unidimensionality results shown in Table 3 indicate that the variance explained by the measures was 45.9% which was close to the model's estimates of 46.2%. This value is higher than the minimum value (40%) set by Linacre (2012). In addition, the value of the unexplained variance in the first contrast was 6.1% which is good because it fell within the range of 5-10%, while the rest of the values of the unexplained variance of the structures, from the second to the fifth, within the range were 3-5%, and, therefore, they were very good (Fisher, 2007). The eigenvalue of the first construction was 2.19 which is lower than the value proposed by Linacre (2010) of 3. This means that there was no second dimension in the scale. The ratio of the variance explained by the items 19.8% to the unexplained variance in the first contrast of 6.1% was 3.25 which is at least three times greater (Conrad et al., 2012).

Table 3. Standardized residual variance in eigenvalue units

	Eigenvalue	Empirical		Modelled
Total raw variance in observations	46.10	100%		100%
Raw variance explained by measures	21.17	45.9%		46.2%
Raw variance explained by persons	12.02	26.1%		26.2%
Raw Variance explained by items	9.15	19.8%		20.1%
Raw unexplained variance (total)	25.00	54.1%	100%	54.7%
Unexplained variance in 1st contrast	2.19	6.1%	8.9%	
Unexplained variance in 2nd contrast	1.92	4.2%	7.7%	
Unexplained variance in 3rd contrast	1.59	3.5%	6.4%	
Unexplained variance in 4th contrast	1.58	3.4%	6.1%	
Unexplained variance in 5th contrast	1.54	3.3%	5.9%	

Local Independence

The assumption of local independence suggests that the responses of the participants at the same level of ability to different scale items are statistically independent (Hambleton & Swaminathan, 1985). Local independence was verified by the observation residual correlation value (Q3) between item pairs. This-value must not exceed 0.30 (Christensen et al., 2017). Table 4 shows the largest observation residual correlations between item pairs. They ranged from 0.16-0.28 and were all less than 0.30. This means that the assumption of local independence of the items of the scale was fulfilled.

Table 4. Largest observation residual correlations

Items Pair		Correlation	Items Pair		Correlation
22	23	0.28	8	16	-0.29
8	9	0.25	19	26	-0.28
17	18	0.24	10	26	-0.23
18	19	0.19	6	16	-0.23
23	24	0.19	3	28	-0.23
2	16	0.19	5	25	-0.22
3	5	0.18	9	16	-0.20
21	22	0.17	9	21	-0.19
9	17	0.16	6	21	-0.18

Equal-item-discriminations

The assumption of equal item discriminations in RSM indicates that the items have equal discriminating values, and their value is equal to 1, but empirically they are not exactly equal (Linacre, 2012). Item discrimination in the EMI scale was estimated, and it was close to the value of 1 for all items which ranged between (0.91-1.27). The equality of these values was verified using the Van Den Wollenberg Test Q1 (1982) of Parallel ICCs/IRFs whose results were (Q1 = 349.71, d.f.= 315, P= 0.087). These results indicate that the assumption of equal item discrimination in the scale was realized.

Gender Differential Item Functioning (GDIF)

This analysis was performed to determine whether or not there was differential performance on the items of the scale between the genders as an indicator of the validity of the items. Winsteps uses a two-tailed t-test with a 95% confidence level and a critical t value of 2.00 to determine the significance of differences in item difficulty between two groups, as well as the Mantel-Haenszel Method using the Chi-square square test. In addition, the value of the DIF Contrast is determined to illustrate the difference between male and female performance. Lai and Eton (2002) indicate that the value of the DIF Contrast for the Likert scale is no more than 0.50 logits, and the values below that value are considered insignificant.

In the current study, the values of the significance level for the differences between the male and female groups were insignificant in all items. Using the t-test and the Mantel-Haenszel Method indicated that there was no differential performance between males and females on all the items, and the values of the DIF Contrast were less than 0.50 for all items except for items (7, 21, 22) for which the value was slightly above 0.50, and which did not show any significance using the t-test and the Mantel-Haenszel Method.

Reliability & Separation

Table 5 shows the summary of the statistics of the persons and items. For the persons, it is noted from the table that the MNSQ for infit and outfit amounted to (1.01 and 1.00) respectively, and this value is ideal (Linacre, 2012). While the productive ZSTD was equal to zero with a standard deviation (1.30, 1.20) for infit and outfit respectively. These values are also ideal (Bond & Fox, 2007). The value of the reliability for the persons was 0.85, and this value is good according to Fisher (2007). The separation index for the persons was 2.35, and this indicates the ability of the scale to separate persons in a construct. This value indicates that the scale can divide persons into two sets of abilities (high, and low). The separation index is considered acceptable if it exceeds 2.00 (Fox & Jones, 1998). For the items, it is noted from the table that the values of MNSQ for infit and outfit amounted to (1.01 and 1.00) respectively, and these values are ideal (Linacre, 2012). While the productive ZSTD value was equal to (-0.10, -0.19) for infit and outfit respectively, and since they ranged between ±2.00, they are considered acceptable (Bond & Fox, 2007). The value of the reliability of the items was 0.97, and this is excellent according to Fisher (2007). The separation index for the items was 6.12, and this indicates the scale ability to statistically differentiate between 6 levels of difficulty. This value is considered excellent according to Fisher (2007).

Table 5. Statistical summary for persons and items

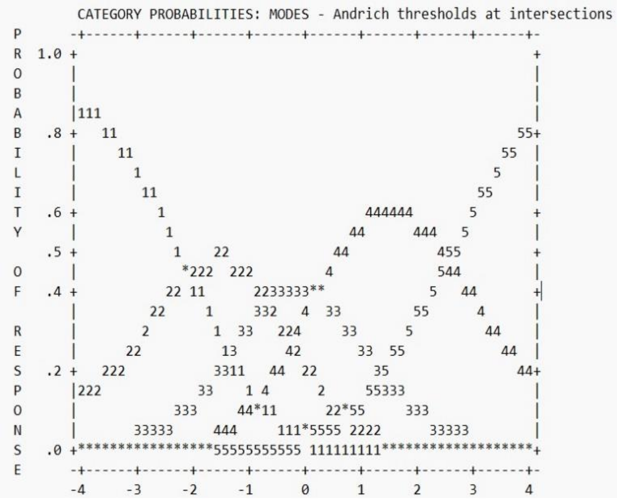
Person	Total	Count	Measure	Realse	INFIT		OUTFIT	
					IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	98.70	25.0	1.52	0.32	1.01	0.00	1.00	0.00
P.SD	9.02	0.00	0.83	0.06	0.38	1.30	0.37	1.20
Real RMSE	0.32	True SD	0.76	Separation	2.35	Reliability		0.85
Item	Total	Count	Measure	Realse	INFIT		OUTFIT	
					IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	991.1	251.0	0.00	0.10	1.01	-0.10	1.00	-0.19
P.SD	71.40	0.00	0.60	0.01	0.27	2.80	0.26	2.78
Real RMSE	0.10	True SD	0.60	Separation	6.12	Reliability		0.97

Rating Scale Calibration

Rasch analysis provides a calibration of scale categories to determine the effectiveness of the scale and to determine if one category should be omitted or included in another. The EMI scale contains five categories (Strongly disagree, Disagree, Neutral, Agree, and Strongly agree). To determine the effectiveness of the scale, five criteria are used (Linacre, 2002). The first criterion is that each category contains 10 observations at the very least which was achieved in the current study as shown in Table 6. The lowest number of observations in the first category was 35. The second criterion suggests that each category must exhibit a probability curve peak, i.e.,

increasing an individual's position on the latent trait (critical thinking) is associated with an increased probability of estimates in higher categories, and this condition is met as shown in Figure 1. The third criterion involves that the average measurement grows in lockstep with the level of measurement. It is noted from Table 6 that this condition has been met. Category measure (-3.32, -1.52, -0.17, 1.84, 3.71) for the categories from the first to the fifth category respectively indicates consistent and escalating response patterns. The fourth criterion, relating to MNSQ values, was also achieved, ranging from 0.91 to 1.36, i.e., within the acceptable range of 0.50-1.50. Finally, the fifth criterion shows that the difference between each category and the next category falls within the range 1.00-5.00. Table 7 shows that all the differences between the categories fell within the specified range.

Figure 1. Category Probability



```
        CATEGORY PROBABILITIES: MODES - Andrich thresholds at intersections
P      -+------+------+------+------+------+------+------+------+-
R  1.0 +                                                         +
O      |                                                         |
B      |                                                         |
A      |111                                                      |
B   .8 +   11                                                  55+
I      |     11                                              55  |
L      |       1                                            5    |
I      |        11                                        55     |
T   .6 +          1                         444444       5       +
Y      |           1                      44      444   5        |
    .5 +            1    22             44           455         +
O      |             *222  222         4             544         |
F   .4 +           22 11      2233333**             5   44       +|
       |         22     1     332   4  33         55      4      |
R      |        2        1  33   224     33      5         44    |
E      |      22          13      42       33  55            44  |
S   .2 +   222           3311   44  22       35                44+
P      |222            33    1 4      2     55333                |
O      |            333     44*11      22*55     333             |
N      |       33333     444     111*5555 2222      33333        |
S   .0 +*****************55555555555 111111111*******************+
E      -+------+------+------+------+------+------+------+------+-
       -4     -3     -2     -1      0      1      2      3      4
```

Table 6. Summary of category structure

Category		Observed		Observed Average	Sample Expect	MNSQ		Andrich Threshold	Category Measure
Label	Score	Count	%			Infit	Outfit		
Strongly disagree	1	35	1	0.33	-0.03	1.24	1.36	None	-3.32
Disagree	2	326	5	0.44	0.37	1.06	1.10	-2.06	-1.52
Neutral	3	1173	19	0.83	0.87	0.94	0.96	-0.67	-0.17
Agree	4	3134	50	1.48	1.49	0.97	0.91	0.38	1.48
Strongly agree	5	1607	26	2.34	2.31	1.01	0.99	2.55	3.71

Table 7. Calibration Differences

Scale	Gaps calculation	Range of acceptance	Decision
G1-G2	0.00 – (-2.06)	1.00 < 2.06 < 5.00	Accepted
G2-G3	-0.67 – (-2.06)	1.00 < 1.39 < 5.00	Accepted
G3-G4	0.38 – (-0.67)	1.00 < 1.05 < 5.00	Accepted
G4-G5	2.55 – 0.38	1.00 < 2.47 < 5.00	Accepted

Discussion

In this study, the quality of the Arabic version of the Critical Thinking Disposition Scale (EMI) by Ricketts and Rudd (2005) was evaluated using the Rasch Rating Scale Model (RSM). The properties of the items and the scale were verified by a set of analyses: item fit, unidimensionality, local independence, equal-items discrimination, gender differential item functioning, reliability and separation and scale calibration. The results of item fit indicated that all the items met the conformity requirements of MNSQ, ZSTD, standard error, and PTMEA Correlation except for item 11which was, therefore, deleted. These results indicated an appropriate quality of the scale items. Regarding unidimensionality and based on the eigenvalue of the first construction which was 2.19 and less than 3, only one predominant factor was detected on the scale. This factor explained approximately 46% of the variation in performance on the scale. This is an indication of the validity of the construction of the scale (Sumintono & Widhiarso, 2014). By using the observation residual correlation value (Q3) between item pairs, the response to any of the scale items has been shown to be independent of the response to the rest of the items. This fulfilled the assumption of the local independence of the scale items.

The assumption of the equal-item discrimination was also verified using the Van Den Wollenberg Test (Q1) of Parallel ICCs/IRFs. The results of this test indicated equal discrimination of the items. This confirms the fitting of the items to the Rasch model. As an indicator of the validity of the items, it was revealed that there was differential performance between males and females on the scale items. The results of the two-tailed t-test and Mantel-Haenszel Method indicated that there was no DIF in any of the items of the scale. Based on the reliability and the separation index of the items and the persons, the items and the persons had high-reliability coefficients. The items were also sensitive enough to divide respondents into two different levels of ability. The separation index for the persons also indicated the ability of the scale to distinguish between 6 different levels of difficulty.

Finally, based on the scale calibration analysis, and according to five different criteria, the results indicated that all the categories of the scale could be retained and that there was no need to delete any of them. The measurement of the categories increased at a steady pace when moving from one category to another. The distance between the categories was also within the specified range of 1.00-5.00. Thus, an individual's increased ability was associated with a higher probability of choosing higher categories over items. Generally, all the results indicated that the scale had good psychometric properties in the Arab environment. It is worth noting that the scale has been verified in languages other than English such as Turkish (Demircioğlu & Kilmen, 2015), and all its items had good properties except for item 11.

Conclusion and Directions for Future Research

In summary, the results of the analysis indicated that all the items of the Critical Thinking Disposition Scale (EMI), except for item 11, met the assumptions of the Rasch model. This showed good suitability for the analysis of fit, unidimensionality, local independence, equality of discrimination, GDIF, and reliability and separation. The categories of the scale were shown to be steadily increasing. This indicates that the Arabic version of the scale had good psychometric properties. The scale consisted of 25 items in its final form. Researchers and practitioners can be confident in their interpretation of the EMI results when used in an Arabic context, especially in the Omani environment. Besides, the scale's properties were only verified in one Arab country, Oman, and the scale was applied to university students. The percentage of females in the sample was twice the percentage of males due to the nature of their percentages in the study population (A'Sharqiyah University). The study recommends the use of other criteria such as convergent and divergent validity as these are associated with critical thinking skills. The properties of the scale in lower age groups such as school students can also be considered.

References

Andrich, D. (2005). The rasch model explained. In S. Alagumalai, D. D. Durtis, & N. Hungi (Eds.), *Applied rasch measurement: A book of exemplars* (pp. 308–328). Springer-Kluwer.

Andrich, D. (2010). Understanding the response structure and process in the polytomous rasch model. In M. L. Nering, & R. Ostini (Eds.), *Handbook of Polytomous Item Response Theory Models* (pp.123-152). Routledge.

Ariffin, S. R. (2008). *Inovasi dalam pengukuran danpenilaian: Inovation in measurement and evaluation.* UKM Press.

Baker, F. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation. USA.

Bond, T. G., & Fox, C. M. (2007). *Applying the rasch model: Fundamental measurement in the human sciences* (2nd ed.). Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how?. *Life Sciences Education, 15*(4), 14-21. https://doi.org/10.1187/cbe.16-04-0148

Bowell, T., & Kemp, G. (2005). *Critical thinking g*. Tracey Bowell: USA and Canada.

Christensen, K., 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

Conrad, K. M., Conrad, K. J., Passetti, L. L., Funk, R. R., & Dennis, M. L. (2015). Validation of the full and short-form self-help involvement scale against the rasch measurement model. *Evaluation Review, 39*(4), 395-427. https://doi.org/10.1177/0193841X15599645

Demircioğlu, E., & Kilmen, S. (2015). Examination of the factor structure of critical thinking disposition scale according to different variables. *American Journal of Theoretical and Applied Statistics (Special Issue: Computational Statistics), 4* (2), 1-8. http://dx.doi.org/10.11648/j.ajtas.s.2015040101.11

Edelen, M., & Reeve, B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16,* 5–18. https://doi.org/10.1007/s11136-007-9198-0

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8,* 341–349. https://psycnet.apa.org/doi/10.1037/1040-3590.8.4.341

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Ennis, R. H. (2003). Critical thinking assessment. In D. Fasko (Ed.), *Critical thinking and* reasoning (pp. 293–310). Cresskill, NJ: Hampton Press.

Fisher, J. W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*(1), 1095. https://www.rasch.org/rmt/rmt211m.htm

Fox, C. M., & Jones, J. A. (1998). Uses of rasch modeling in counseling psychology research. *Journal of Counseling Psychology, 45*(1), 30-45. https://psycnet.apa.org/doi/10.1037/0022-0167.45.1.30

Goroshit, M. (2018). Academic procrastination and academic performance: An initial basis for intervention. *Journal of Prevention & Intervention in the Community, 46*(2), 131–142. https://doi.org/10.1080/10852352.2016.1198157

Gustafson, J. E. (1980). Testing and obtaining fit of data to the rasch model. *British Journal of Mathematical and Statistical Psychology, 33*(2), 205-233. https://doi.org/10.1111/j.2044-8317.1980.tb00609.x

Halpern, D. F. (2003). *Thought & knowledge: An introduction to critical thinking* (4nd ed.). New Jersey: Lawrence Erlbaum Associates.

Hambleton, R. K., & Swaminathan, H. (1985*). Item response theory: Principles and application.* Kluwer Nijhoff Publishing, Boston, U.S.A.

Harlow, L. (2005). *The essence of multivariate thinking basic themes and methods.* Lawrence Erlbaum Associates, Publishers Mahwah, New Jersey.

Hashemiannejad, F., Oloomi, S., & Oloomi, S. (2016). Examine the relationship between critical thinking and happiness and social adjustment. *International Academic Journal of Social Sciences, 3*(6), 42–47. https://doi.org/10.9756/IAJSS/V6I1/1910003

Heidari, M., & Ebrahimi, P. (2016). Examining the relationship between critical-thinking skills and decision-making ability of emergency medicine students. *Indian Journal of Critical Care Medicine, 20* (10), 581–586. https://doi.org/10.4103/0972-5229.192045

Hurst, P. (1999). *Philosophy of Education: The main themes in the tradition of analytical* (B. Shabani Varaki, & M. R. Shoja Razavi, Trans). Mashhad: Ferdowsi University of Mashhad Press

Irani, T., Rudd, R., Gallo, M., Ricketts, J., Friedel, C., & Rhoades, E. (2007). *Critical thinking instrumentation manual.* http://step.ufl.edu/resources/critical_thinking/ctmanual.pdf

Karami M., & Shakurnia A. (2020). Critical thinking disposition in the pharmacy faculty members of ahvaz jundishapur university of medical sciences, Iran. *Educational Research in Medical Sciences, 9*(2). http://dx.doi.org/10.5812/erms.109691

Kim, S., & Kyllonen, P.C. (2006). *Rasch rating scale modeling of data from the standardized letter of recommendation.* ETS Research Report Series. https://doi.org/10.1002/j.2333-8504.2006.tb02038.x

Kwon, N., Onwuegbuzie, A.J., & Alexander, L. (2007). Critical thinking disposition and library anxiety: Affective domains on the space of information seeking and use in academic libraries. *College & Research Libraries, 68*(3), 268-278. https://doi.org/10.5860/crl.68.3.268

Lai, E. R. (2011). *Critical thinking: A literature review research report.* London: Parsons Publishing.

Lai, J. S., & Eton, D. T. (2002). Clinically meaningful gaps. *Rasch Measurement Transactions, 15*(4), 850. https://www.rasch.org/rmt/rmt154e.htm

Lee, S. (2009). *Examining the relationships between metacognition, self-regulation and critical thinking in online Socratic seminars for high school social studies students* [Unpublished doctoral dissertation]. The University of Texas, Austin.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean?. *Rasch Measurement Transactions, 16*(2), 878. https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2005). *Winsteps rasch measurement computer program.* MESA Press.

Linacre, J. M. (2010). When to stop removing items and persons in Rasch misfit analysis?. *Rasch Measurement Transactions, 23*(4), 1241. https://www.rasch.org/rmt/rmt234g.htm

Linacre, J. M. (2012). *A user's guide to WINSTEPS: Rasch model computer programs.* MESA Press.

Lismayani, I., Parno, P., & Mahanal, S. (2017). The correlation of critical thinking skill and science problem-solving ability of junior high school students. *Journal Pendidikan Sains, 5* (3), 96–101. http://dx.doi.org/10.17977/jps.v5i3.10338

Lyutykh, E., (2009). Practicing critical thinking in an educational psychology classroom. *Journal of educational studies, 45*, 377-391. https://doi.org/10.1080/00131940903066263

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11.

Mahal, R., Chawla, A., & Kanwar, V. (2015). Critical thinking as a correlate of stress management among rural adolescent girls. *Advance Research Journal of Social Science, 6*(1), 32–35. http://dx.doi.org/10.15740/HAS/ARJSS/6.1/32-35

Norris, S. P. (2003). The meaning of critical thinking test performance: The effects of abilities and dispositions on scores. In D. Fasko (Ed.), *Critical thinking and reasoning: Current research, theory and practice* (pp. 315-329). Cresskill, NJ: Hampton Press.

Page, A. (2007). Promoting critical thinking skills by using negotiation exercises. *Journal of education for business, 82*(5), 251-257. http://dx.doi.org/10.3200/JOEB.82.5.251-257

Paul, R., & Elder, L. (2014). *The miniature guide to critical thinking: Concept and tools* (7th Ed.). Dillon Beach, CA: Foundation for Critical Thinking Press.

Paul, R., Elder, L., & Bartell, T. (1997). *California teacher preparation for instruction in critical thinking: Research findings and policy recommendations.* Sacramento, CA: Commission on Teacher Credentialing.

Profetto, M.J. (2003). The relationship of critical thinking skills and critical thinking dispositions of baccalaureate nursing students. *Journal Advance Nurse, 43*(6), 569-577. https://doi.org/10.1046/j.1365-2648.2003.02755.x

Ricketts, J. C., & Rudd, R. D. (2005). Critical thinking of selected youth leaders: The efficacy of critical thinking dispositions, leadership, and academic performance. *Journal of Agricultural Education, 46*(1), 33-44. http://dx.doi.org/10.5032/jae.2005.01032

Rincker, L. (2014). *Critical thinking dispositions of students receiving livestock evaluation training* [Unpublished Master Thesis]. Agricultural Education, California State University, Chico.

Sick, J. R. (2009). Rasch measurement in language education Part 3: The family of rasch models. *SHIKEN, 13*(1), 4-10. Retrieved from http://jalt.org/test/sic_3.htm.

Sk, S., & Halder, S. (2020). Critical thinking disposition of undergraduate students in relation to emotional intelligence: Gender as a moderator. *Heliyon, 6*(11), e05477. https://doi.org/10.1016/j.heliyon.2020.e05477

Stedman, N. L. P., & Andenoro, A. C. (2006). *Linking emotional intelligence to critical thinking: Balancing our curriculum within leadership education.* Proceedings of the Association of Leadership Educators, Big Sky, MT.

Stedman, N. L. P., & Andenoro, A. C. (2007). Identification of relationship between emotional intelligence skill & critical thinking disposition in undergraduate leadership students. *Journal of Leadership Education, 6*(1), 190-208. http://dx.doi.org/10.12806/V6/I1/RF10

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial [rasch model application for social disciplines]*. Trim Komunikata Publishing House.

Taghva, F., Rezaei, N., Ghaderi, J., & Taghva, R. (2014). Studying the relationship between critical thinking skills and students' educational achievement (Eghlid Universities as Case Study). *International Letters of Social and Humanistic Sciences, 25,* 18–25. https://doi.org/10.18052/www.scipress.com/ILSHS.25.18

Van den Wollenberg, A. (1982). A simple and effective method to test the dimensionality axiom of the rasch model. *Applied psychological measurement, 6*(1), 83-91. https://doi.org/10.1177/014662168200600109

Von Davier, M. (2016). Rasch Model. In Wim J. van der Linden (ed.), *Handbook of Item Response Theory* (Boca Raton: CRC Press), Routledge Handbooks.

Wilson M. (2005). *Constructing measures: An item response modelling approach*. Lawrence Erlbaum Associates, Inc.

Wu, M., & Adams, R. (2007). *Applying the rasch model to psychosocial measurement: A practical approach.* Educational Measurement Solutions.

Yüksel, G., & Alci, B. (2012). Self-efficacy and critical thinking dispositions as predictors of success in school practicum. *International Online Journal of Educational Sciences, 4*(1), 81-90.