JISAE
Online ISSN : 2597-8943
Print ISSN : 2442-4919
Journal of Indonesian Student Assessment and Evaluation

THE FUNCTIONALITY OF THE MIDDLE VALUE OF THE INDONESIAN VERSION OF EMOTIONAL LEARNING INSTRUMENT

Erwin Sulaeman
Universitas Negeri Jakarta
erwinsulaiman_pep17s2@mahasiswa.unj.ac.id

Wardani Rahayu
Universitas Negeri Jakarta
Wardani.rahayu@unj.ac.id

Erdawaty Kamaruddin
Universitas Negeri Jakarta
erda_kamaruddin@yahoo.com

Winona Amanda Tiara Widodo
Universitas Indonesia
winona.amanda@gmail.com

ABSTRACT

This article discusses the psychometric validity of the Indonesian version of emotional learning instruments with a scale of five and four response categories.The purpose of this study is to produce an Indonesian version of emotional learning instrument with an effective response category scale used by Indonesians. The instrument is a modification of the scale of the Learning Environment Research Questionnaire on Emotional Climate Classroom. This study is a survey of 1494 responses of 7th and 8th grade junior high school students.Samples were selected by random sampling and based on considerations of schools implementing the 2013 Curriculum.Modification instruments consisting of 43 items were tested in obtaining validity based on item difficulty estimations and psychometric criteria with Rasch modeling.The results of this study indicate that the Andrich threshold validity testing meets the monotonic characteristics and the Standardized Residual Correlation is higher, so the scale of the five response categories is more effective to measure the Indonesian version of emotional learning instruments than the scale of the four response categories.

Keywords: Functionality of middle value, ELVI, RaschModeling

INTRODUCTION

Emotions in the learning environment are formed from experiences and physical feelings. This condition must consider students' cognitive interests, aspirations and emotional lives to develop (Woodhouse, 2017). The importance of the learning environment influences student achievement and attitudes, (Ghosh, 2015; Koul, Fraser, Maynard, & Tade, 2018; Marchesi & Cook, 2012) reported that in the schools of Appalachian states in West Virginia, nearly 51000 students dropped out of high school due to less than 85 - 90% attendance, serious discipline violations, and stress in learning. Learning environment in classrooms embodies relationships between teachers, students, and student attitudes(López et al., 2018).

Subjective perceptions of teachers or students are felt with various important results regarding achievements (Jones et al., 2017), emotional and social aspects (Taylor, Oberle, Durlak, & Weissberg, 2017).The progress of practice in schools can be designed through emotional ability (Jones et al., 2017; Taylor et al., 2017; Yaeger, 2017), this becomes the basis for developing the Indonesian version of emotional learning instruments (ELVI).

Emotional learning in developed countries has been carried out, one of which is in Central Indiana and schools in the United States (Melnick, Cook-Harvey, &Darling-Hammond, 2017).In Indonesia, emotional learning is still theoretically introduced to character education(Suriyanti, 2015). Some research on emotions in relation to classroom environment is mostly concentrated on student anxiety.(Watt, Carmichael, & Callingham, 2017). The nature of emotional learning that influences how behavior is carried out leads to a learning environment or behavioral responses that appear on different time scales(Lowe, 2014). To get information about emotional learning, the right instrument is of course needed to be applied in Indonesia.The Learning Environment Research (LER) measurement scale was chosen in the modification of the ELVI instrument, based on recommendation of (Koul et al., 2018) about LER in Asia, that there is room for Asian researchers to modify the study environment study. To measure the level of latent nature related to the ability of emotional learning analysis using Rasch modeling. Its ability to predict missing data is based on a systematic response pattern, producing a standard measurement value of error and calibration in three ways, namely: the measurement scale, respondents, and items (Jae Jeong, 2016; Perera, Sumintono, & Jiang, 2018).

Instruments said to be valid must have a scaled concept.(Perera et al., 2018). The problem of the optimal number of response categories has not been resolved, as seen from the response patterns and information retrieval (Jae Jeong, 2016). A scale with more than two or three response categories can provide maximum information retrieval (Green, 2010). Odd and even category scale, with respect to functioning of mean. Odd numbers from the response category are generally preferred, because the functioning of the middle value is interpreted as a neutral point, thus providing an opportunity to represent respondents' emotions neutrally and discriminatively. Omission of the middle value forces respondents to be wiser, resulting in a more precise ranking(Andrich, 2016; Green, 2010). The ELVI instrument was designed with five and four-category response frequency type scales.

This has become a renewal in following up research (Adelson & McCoach, 2010) which previously compared the five-point scale and the four-point Likert type scale.The research has not yet investigated the effect of the number of response categories affecting the stability of student responses and helped answer whether the scale of the five response categories with functioning of middle values psychometrically outperformed the four response category scales. The

effectiveness of the scale used can be known through the validity of the Andrich threshold.The purpose of this study is to determine the differences in the validity of the Andrich threshold in an ELVI instrument with a scale of five and four response categories based on Rasch modeling.

Emotional learning is an inseparable component of cognitive process, testing how emotions during learning experience affect metacognitional progress that holds at the level of students' abilities(Chao, Dede, & Star, 2016). Cognitive processing is influenced by states of emotion(Lizzio, Wilson, & Simons, 2010). Emotional Learning is defined as an ability to help students recognize, express and regulate their own emotions, build relationships with peers and adults, empathize with other people's perspectives, maintain and focus attention (cognitive regulation), and understand the emotional perspectives of others.Recognizing how different situations are and deal with feelings in a prosocial way(Jones et al., 2017; Marchesi & Cook, 2012).

(Swartz, 2017) divided two emotional areas, namelypersonal competence and empathy.Personal competence includes self-awareness, self-management, and social awareness. Empathy is an awareness to give attention, needs or care to others and maintain social relationships. A rating scale that involves more than two response categories is a popular response format of measurement in education. Aresponse scales is closely related to building validity(Salzberger, 2014). (Revilla, Saris, & Krosnick, 2014) showed in their study that a few response categories tend to produce smaller validity. (Green, 2010; Neumann, Neumann, & Nehm, 2011) explained that odd numbers from the response category are generally preferred over even numbers because the middle category is interpreted as a neutral point so it tends to strengthen preferences for a scale of five categories.(Wakita, Ueshima, & Noguchi, 2012) explained that a scale without neutral intermediaries is preferred because respondents are forced to make definite choices.

(Sumintono, 2015) explained that the ranking scale validity analysis is conducted to verify whether the ranking of choice used confuse respondents or not.The Rasch model analysis provides a process of verifying the ranking assumptions given by looking at the Obsvd Avrge. Andrich Threshold tests whether the polytomic values used have been correct or not. (Distefano, Greer, Kamphaus, & Brown, 2015; DiStefano & Morgan, 2010) argue that the threshold as a moving point from one category to an adjacent category on the rating scale. The threshold number is equal to the number of scale categories (k-1).

(Lundgren-Nilsson, Dencker, Jakobsson, Taft, & Tennant, 2014) Threshold is a point between two categories that have the same possible response.When a threshold gets broken, items can be saved again by reducing the category.(Huang, 2016) states that the higher the estimated threshold parameters, the greater the defect measured.If the defect is not too severe, the item category with some or a little difficulty can dominate. (Gonza, Zabalegui-ya, Lo, & Siso, 2014) explained

that the number of responses in each category and the threshold for each item assessed the effectiveness of the rating scale.

METHOD

This research is a survey adopted from the post-positivism paradigm with a questionnaire method.Samples on a scale of five and four response categories on as much as 1494 student responses were taken at random in the province of Jakarta.Rasch Modeling (Kean, Bisson, Brodke, Biber, & Gross, 2018; Kutlay, Küçükdeveci, Gönül, & Tennant, 2018) describe the Rasch model, concerning the ability of nature, difficulty of items, and suitability of items used to examine psychometric properties of a collected instrument.(Andrich, 2016) explains the Rasch modeling put forward first by George Rasch from Denmark in the 1950s.According to (Kutlay et al., 2018) Rasch modeling relates to IRT as a modern measurement theory, while an existing measurement theory is stated as a classical measurement theory.According to (DiStefano & Morgan, 2010) that the Rasch model requires endurance of assumptions for accurate estimates, including (1) establishing unidimensionality, (2) monotonous scales, and (3) item fit. The ELVI instrument grids can be seen in Table 1 below:

Table 1 ELVI Instrument Grids

Dimension	Indicator	Item Number Before Modified	Total	Item Number After Modified	Total
Self-Awereness	Captivate	33,34,35,36, 37,38,39,40	8	33,34,35,36,37,38,39 ,40	8
Self-Management	Control	9,10,11,12,13,14,15 ,16	8	10,11,12,13,14,16,16 a	7
Social-Awereness	Care	1,2,3,4,5,6,7,8	8	1,2,3,4,4a,5,6,7,8	9
	Confer	41,42,43,44, 45,46,47,48	8	41,43,47,48	4
Relationship Skills	Challenge	25,26,27,28, 29,30,31,32	8	26,27,29	3
Decision Making Responsible	Clarify	17,18,19,20, 21,22,23,24	8	17,19,20,21,22,23,24	7
	Consolidate	49,50,51,52, 53	5	49,50,51,52,53	5
Total			53		43

RESULT

The basic requirement in construct validity is that instruments must be designed to measure one latent construct.**Unidimension** in Rasch modeling refers to invariant measurements(Kaliski et al., 2013). Unidimension becomes important as the essence of determining parameter estimation(Sinnema, Meyer, & Aitken, 2016). The importance of determining unidimension as proof of internal consistency(Huberty et al., 2013). The results of the unidimensional calculation of five and four response categories are shown in Table 2 below:

Table 2Unidimensions for scales of five and four response categories

Unidimension for five response category scale

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	64.2143	100.0%		100.0%
Raw variance explained by measures =	25.2143	39.3%		39.2%
Raw variance explained by persons =	8.3391	13.0%		13.0%
Raw Variance explained by items =	16.8751	26.3%		26.2%
Raw unexplained variance (total) =	39.0000	60.7%	100.0%	60.8%
Unexplned variance in 1st contrast =	2.5707	4.0%	6.6%	
Unexplned variance in 2nd contrast =	2.3281	3.6%	6.0%	
Unexplned variance in 3rd contrast =	2.2349	3.5%	5.7%	
Unexplned variance in 4th contrast =	1.9747	3.1%	5.1%	
Unexplned variance in 5th contrast =	1.7177	2.7%	4.4%	

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information unit

	Eigenvalue	Observed		Expected
Total raw variance in observations =	65.4540	100.0%		100.0%
Raw variance explained by measures =	28.4540	43.5%		43.1%
Raw variance explained by persons =	14.2432	21.8%		21.6%
Raw Variance explained by items =	14.2108	21.7%		21.5%
Raw unexplained variance (total) =	37.0000	56.5%	100.0%	56.9%
Unexplned variance in 1st contrast =	2.6960	4.1%	7.3%	
Unexplned variance in 2nd contrast =	2.3729	3.6%	6.4%	
Unexplned variance in 3rd contrast =	1.9289	2.9%	5.2%	
Unexplned variance in 4th contrast =	1.7761	2.7%	4.8%	
Unexplned variance in 5th contrast =	1.6282	2.5%	4.4%	

Unidimension for four response category scale

Unidimensional criteria are seen in "raw variance explained by measure.”The results in Table 2 are 39.3% for the scale of five response categories and 43.5% for the scale of four response categories.Both of them have a value greater than 20% so that the instruments meet the requirements for unidimension(Shih, Chen, Sheu, Lang, & Hsieh, 2013). Further dimensional analysis is proven through the *Eigenvalue units* column(Huberty et al., 2013; Kaliski et al., 2013), the value obtained is a scale of five response categories, namely: 2.6, 2.3, 2.2, 2.0, and 1.7.Variances that cannot be explained as follows: 4.0%, 3.6%, 3.5%, 3.1% and 2.7%.Eigenvalue units on a scale of four response categories: 2.7, 2.4, 1.9, 1.8, and 1.6, variance that cannot be explained: 4.1%, 3.6%, 2.9%, 2.7%, and 2.5%.An

unexplained variance of both scales is less than 15%(Sinnema et al., 2016). The value of variance is in the range of 3-5% in the very strong category(Seol, 2016). Thus empirically the ELVI instrument with a scale of five and four response categories of unidimension and building construct validity.

The monotonic nature of the modified ELVI instrument from the LER scale, Questionnaire on Classroom Emotional Climate. The use of frequency scales from five and four response categories can be seen in the following Table 3:

Table 3Rating expression in each scale

Scale	Response Category
5	Never
	Rarely
	Occasional
	Often
	Always
4	Never
	Rarely
	Often
	Always

In Table 3, the scale of the five response categories prioritizes the functioning of the middle value, placing a choice of three (3) with "occasional" indication (Naga, 2012).The scale of the four response categories negates the functioning of the middle value, so that students' responses are wiser and produce more precise rankings (Green, 2010).

Table 4 Andrich thresholdin five and four scale response categories

Scale Category			
Obsvd Avrge (5)	Andrich Threshold (5)	Obsvd Avrge (4)	Andrich Threshold (4)
-0,83	NONE	-0,83	NONE
-0,15	-2,17	0,17	-2,48
0,33	-0,38	1,21	0,33
0,94	0,59	2,43	2,14
1,54	1,97		

(Andrich, 2011) explained that sequential threshold distances are not positively isolated and it is said that the response category can be interpreted as an ordinal scale.Table 4 shows that there was an increase in value on both scales, shown in the Observed Average column from negative to positive direction.Logit scores on a scale of five response categories start at -0.83 for choice of category 1 (never), 0.15 for category 2 (rare), 0.33 for category 3 (occasional), 0.94 for category 4 (often), and 1.54 for category 5 (always).Logit scores on a scale of four response categories start at -0.83 for category 1 (never), 0.17 for category 2 (rarely), 1.21 for category 3 (often), and 2.43 for category 4 (always). The Andrich threshold value on the

scale of five monotonous response categories rises from NONE towards negative logit direction (-2.17) and leads to positive logit (1.97).

The Andrich threshold value on the scale of the four monotonous response categories rises from NONE towards negative logit direction (-2.48) and leads to positive logit (2.14).Thus the increase in logit scores monotonically indicates that student responses can distinguish between the choices of response categories and verify the level of response of students who agree on the basis of both scales.This monotonic movement illustrates that items are in accordance with the students' choice of response categories for measurement.

Fit Item in Rasch modeling can see the quality of the item's conformity to the model, explaining whether the statement item is functioning normally in taking measurements or not. Examination of mismatch index is seen in the value of Outfit Mean Square (MNSQ), Estimated Outfit Z Standard (ZSTD), and Point Measure Correlation (DiStefano & Morgan, 2010; Perera et al., 2018).MNSQ through squared standardized residual assumptions aims to determine misfits in reporting actual data.Showing a match between items and student responses that are not standardized. Criteria for an item to be declard fit, MNSQ values has to be between 0.5 to 1.5 logit (Abd-el-fattah, 2015; Elisabet, Benito, & Miguel, 2012; Harachi, 2012; Seol, 2016). ZSTD with a value of -1.96 to +1.96 indicates that an estimation is accepted(Elisabet et al., 2012; Seol, 2016). ***Point Measure Correlation*** to measure the identification of internal consistency in items and student responses. Items with negative Point Measure Correlation (-) are misfit items.Estimation in the PT-MEASURE CORR column with acceptance criteria is $0.32 < x < 0.8$ (Abdullah et al., 2012; Boone & Noltemeyer, 2017).

This research tests 43 items on the ELVI instrument with a scale of five and four response categories, the results of the analysis can be seen in the following Table 5:

Table 5Fit and Unfit Items in Five and Four Response Category Scales

Category	MNSQ Out Fit Values (Fit Items)	PT-Measure Correlation Values (Item Fit)	Unfit Items	Fit Items
ELVI instruments with scale of five	0.68 to 1.4	0.24 to 0.60	B9, B10, B11, B27	B1, B2, B3, B4, B5, B6, B7, B8, B12, B13, B14, B15, B16, B17, B18, B19, B20, B21, B22, B23, B24, B25, B26, B28, B29, B30, B31, B32, B33, B34, B35, B36, B37, B38, B39, B40,B41, B42, B43
Total			4	39
ELVI instruments with scale of four	0.76 to 1.33	0.84 to 0.68	B9, B10, B11, B13, B15, B27	B1, B2, B3, B4, B5, B6, B7, B8, B12, B14, B16, B17, B18, B19, B20, B21, B22, B23, B24, B25, B26, B28, B29, B30, B31, B32, B33, B34, B35, B36, B37, B38, B39, B40,B41, B42, B43
Total			6	37

Based on Table 5, there are 39 fit items in the five response category scales. Four items are unfit, which are B9, B10, B11, and B27. MNSQ values from 0.68 logit to 1.4 logit and PT-Measure Correlation value from 0.24 logit to 0.60 logit.Fit items on a scale of four response categories contained 37 items.Six items are unfit, namely: B9, B10, B11, B13, B15, and B27. MNSQ values from 0.76 logit to 1.33 logit and PT-Measure Correlation value from 0.84 logit to 0.68 logit.Thus the ELVI instrument with a scale of five response categories has 39 fit items, while the scale of the four response categories has 37 fit items.No repairs for unfit items, so they are not used.Monotonic movements see an increase in average values of each item. The increase is described as in the Andrich threshold logit value from negative to positive logit direction. Take a look at the following Figure 1:

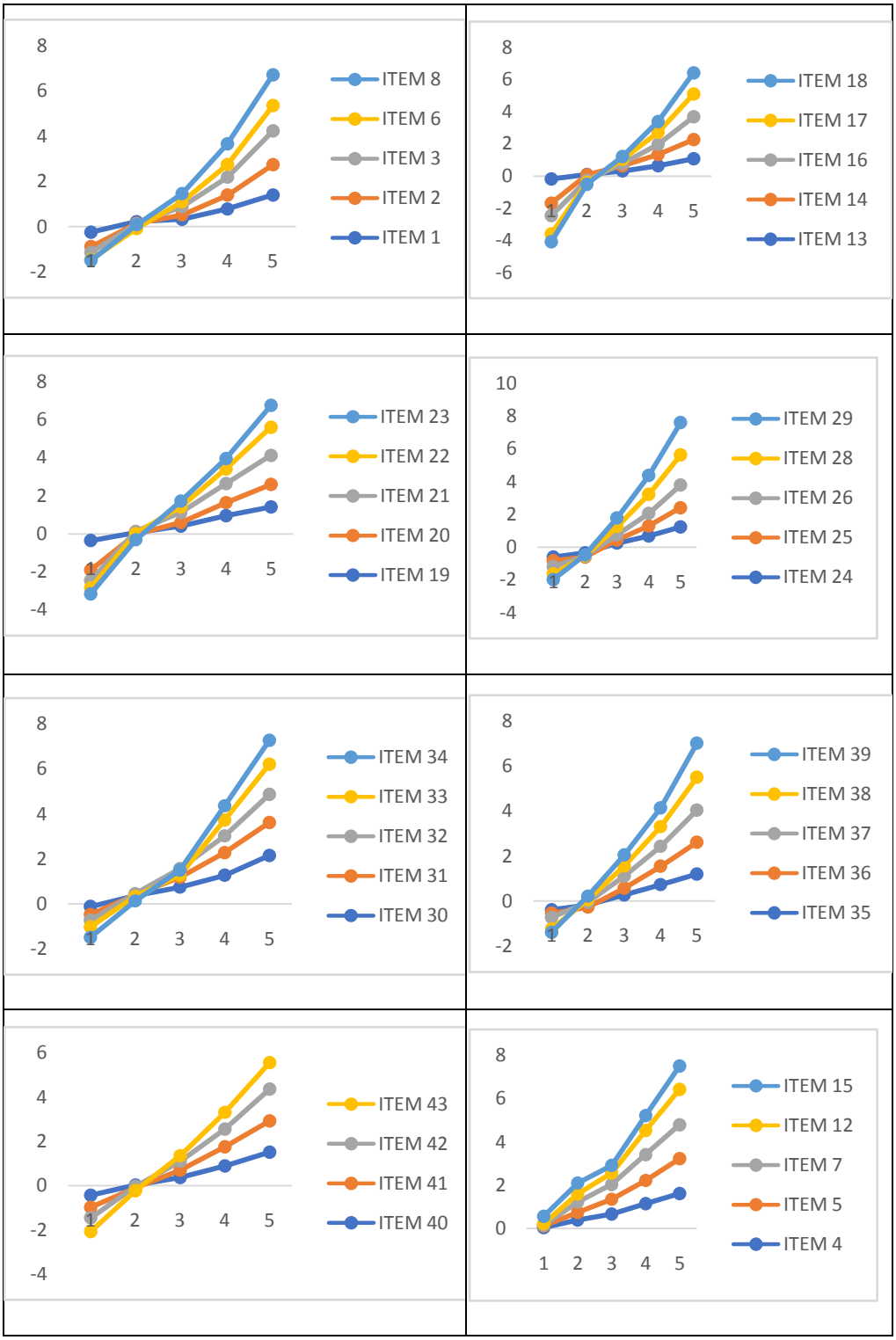
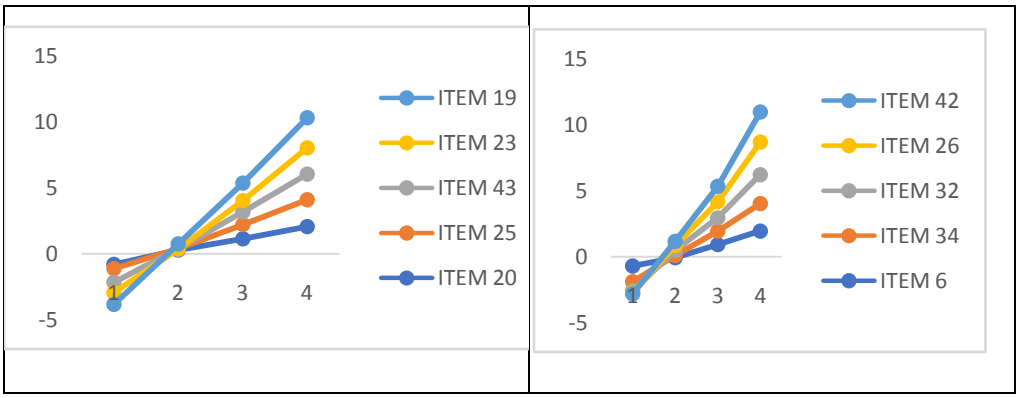


Figure 1 Andrich threshold monotonic graph on a scale of five response categories

In Figure 1 it shows 39 items on a scale of five response categories, item analysis is depicted in detail to show that the functioning of middle value on a scale of five response categories is more likely to be selected(Moors, 2008). Item description of students' responses in distinguishing between choices 'never', 'rarely', 'occasional', 'often', and 'always'.The findings based on Figure 1 show that the student response styles were clearly observed.Line graphs for each monotonous upward item indicate that the average is highest among the two response categories: 'often' and 'always'.

The difference in format on the scale of the five response categories is how the middle response category positions, an analysis of each item where the average for the response category 'occasional' rises monotonically.The highest average value falls on item B30 in the self-awareness dimension of the *captivate* indicator with the statement "I have an interesting homework to do." The functioning of the middle value is prominent in items with dimension of self-awareness. This shows that according to (Kupana, 2015; Lapoint & Butty, 2009) students'responses assess students' feelings, interests, values, and strengths accurately to maintain reasonable self-confidence.

Student responses that choose alternative middle values do not need to answer the question in the same way as other respondents.This shows that the response of students which are less intense is more influenced by presence or absence of the middle response category (Moors, 2008).Thus the instrument with a scale of five response categories that prioritizes the functioning of the middle value measures emotional learning based on self-awareness responses. The monotonous nature based on Andrick threshold requirements can be seen in Figure 2 below:



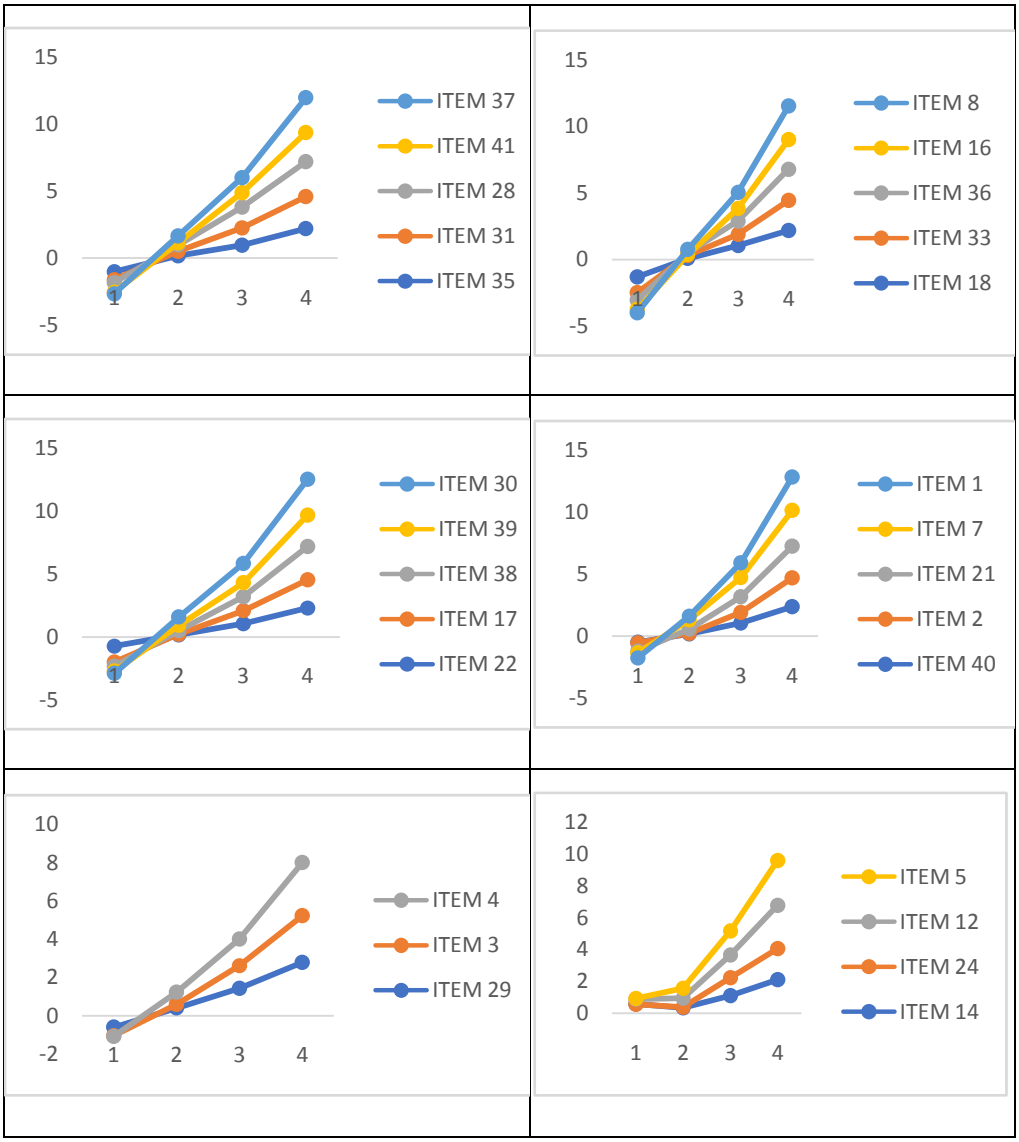


Figure 2 Andrich threshold monotonics on a scale of four response categories

Based on Figure 2, showing 37 items on a scale of four response categories, item analysis is depicted in detail to show that a scale of four response categories that negates the middle function is more likely to be chosen. Item description of students' responses in distinguishing between choices 'never','rarely,' 'often,' and 'always.'The findings based on Figure 2, the student response style is clearly observed.The line graph for each item is monotonically upwards, indicating that the average is highest between the two response categories: 'often' and 'always.'

The difference in format on the scale of the four response categories is to exclude the middle response function, an analysis of each item where the average for the monotonous response category rises. The highest average value fell on item B7 on the *care* indicator in the social-awareness dimension with the statement "My

teacher knows when something is bothering me."The scale of four response categories is prominent in items with social-awareness dimension. This shows that according to (Kupana, 2015; Lapoint & Butty, 2009) the choice of student responses is based on perspectives on individual and group differences.Student responses on a scale of four response categories are consistent in reflecting higher agreement (Moors, 2008). Thus the instrument with a scale of four response categories measures emotional learning based on social awareness responses.

Each item has an Andrich threshold value with a different monotonous increment distance. (Andrich, 2011) explained that sequential threshold distances from negative to positive were not isolated and it was said that the response category could be interpreted as an ordinal scale.The scale of five response categories in the figure shows that out of 39 items, there are five items that did not meet Andrich threshold requirements: items B4, B5, B7, B12, and B15. Item B4 with the statement "when I am sad, the teacher helps to feel better" is in the *care* dimension, with logit values (0.03, 0.38, 0.66, 1.14, to 1.61). Item B5 with "when I'm angry, the teacher helps to feel better" is in the *care* dimension, with logit values (0.08, 0.36, 0.68, 1.07, to 1.61). Item B7 with the statement "My teacher knows when something is bothering me" in the *care* dimension, with values (0.03, 0.48, 0.69, 1.20, to 1.57).Item B12 with the statement "friends in class behave according to my teacher's wishes" on the *control* dimension, with values (0.08, 0.38, 0.53, 1.11, to 1.63). Item B15 with the statement "My class is busy during the learning process" on the control dimension, with values (0.34, 0.50, 0.35, 0.69, to 1.08).

Figure on the scale of the four response categories shows that out of 37 items, four items did not meet Andrich threshold requirements: items B5, B12, B14, and B24. Item B5 with the statement "when I am angry, the teacher helps me to feel better" in the *care* dimension, with values (0.04, 0.63, 1.51, to 2.82).Item B12 with the statement "friends in class behave according to my teacher's wishes" in the *control* dimension, with values (0.31, 0.56, 1.43, to 2.73).Item B14 with the statement "every student knows what he must learn in class" in the *control* dimension, with values (0.57, 0.33, 1.10, to 2.10).Item B24 with the statement "My teacher does not let students give up when doing difficult tasks" in the *clarify* dimension, with values (0, 0.05, 1.13, to 1.96).

Thus in the scale of the five response categories, there are five items and in the scale of the four response categories, there are four items that have a positive upward movement, but not in accordance with Andrich threshold requirements. Monotonic movements of 34 items on a scale of five response categories and 33 items on a scale of four response categories show evidence of use of the two scales to be ordinal.The inappropriate items were isolated as ordinal scale response categories (Andrich, 2011).

Testing the information function scale of five response categories and the scale of four response categories is to find out which scale provides a lot of information (Gonza et al., 2014).A detailed description can be seen in Figure 3 below:

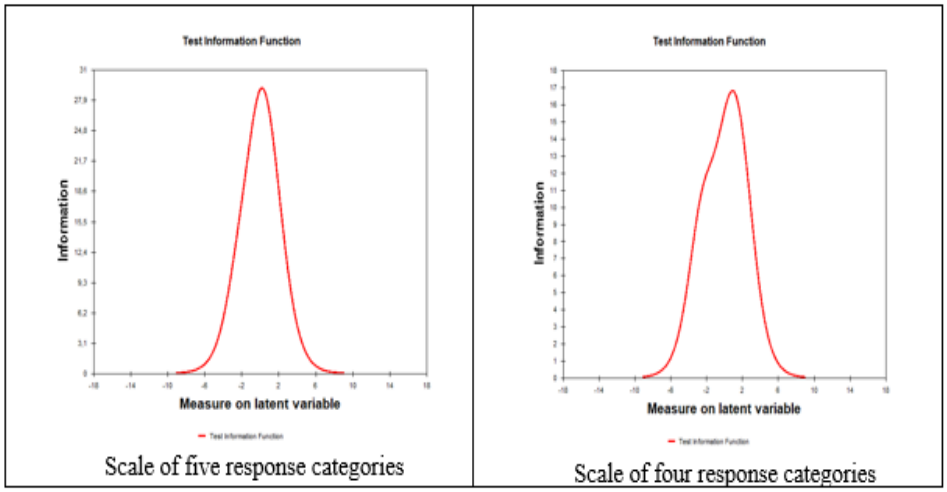


Figure 3 The informational function scale of five and four response categories

The functioning of the middle value compares the scale of the five and four response categories, to show that the information functions of the two scales must be calculated.The information function explains the strength of an item in uncovering the latent trait measured in the ELVI instrument, so that it is known which items are suitable for the model (Pretz et al., 2016).Different informational functions are seen from the plot of the two scales.The ability (emotional learning) level of student responses is shown on the X axis while the magnitude of the informational function is by the Y axis.Both plots show different pictures so that it can be interpreted that the informational function of the two plots is not optimal for each individual. The peak heights of both plots are different; the peaks of the scale of five response categories appear to be higher than the scale of the four response categories.This is in line with the opinion of (Vaughan, 2018) that the higher the peak of informational function, the higher the information can be given.Thus it can be concluded that the scale of the five response categories provides more information than the scale of the four response categories for various values of θ (emotional learning).

The validity of the Andrich threshold ELVI instrument with a scale of five and four response categories can prove that there is an effective response category scale used to measure emotional learning in Indonesia. The results of analysis of the scale of five and four response categories can be seen in the following Table 6.

Tabel 6 Comparison of Andrich Threshold's psychometric validity

Fit persons	Fit Items	Response Category	Andrich Threshold	Fit Itemsbased on Andrick Threshold validity	Standardized Residual Correlation (SRC)
907	39	Five	-2.17, -0.38, 0.59, 1.97	34	-0.177 and 0.638
976	37	Four	-2.48, 0.33, 2.14	33	-0.195 and 0.565

Testing on 43 items of the ELVI instrument using the scale of five response categories resulted in 39 fit items with 907 fit persons, while the use of scale of four response categories produced 37 fit items with 976 fit persons.The functioning of response categories in the scale used can be seen through the Andrich threshold parameter (Gonza et al., 2014; Meiser, 2017).Comparison of threshold results responding to the existence of one of the response categories that is effectively used in measuring the ELVI instrument (Andrich, 2011).The results in Table 6 show the Andrich threshold value of the scale of five response categories moving from negative (-) towards positive direction, namely: (-2.17, -0.38, 0.59, to 1.97), while on a scale of four response categories moves from the values (-2.48, 0.33, to 2.14).The estimated threshold value on the scale of the four response categories runs higher than in the scale of the five response categories. This is consistent with the opinions of (McDonald, Vidacovich, Ascione, Williams, & Green, 2015) that the higher the estimated threshold parameters, the greater the disability measured.It was concluded that the scale of the five response categories with estimated threshold parameter values, for each item assessing the effectiveness of the rating scale in measuring ELVI instruments (Gonza et al., 2014).

Following up on the Andrich threshold value for the two scales, an Andrich threshold value was analyzed for each item. This can be ensured by looking at the increase in the average value of each item as required by the Andrick threshold, which moves from negative to positive direction.Thus as described in Figure 2 before, fit items are obtained from 39 to 34 items for emotional learning instruments with a scale of five response categories. Fit items for emotional learning instruments with scale of four response categories from 37 items to 33 items. The acquisition is reviewed based on the suitability of each item that meets the Andrick threshold requirements. Further analysis of the Andrich threshold differences can be seen in Figure 4 below:

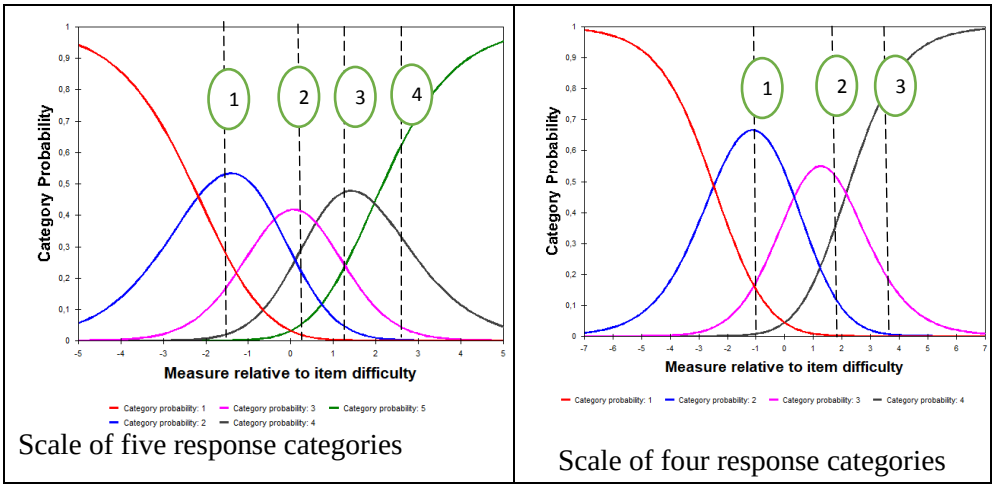


Figure 4 Andrich probability threshold of scale of five and four response categories

Figure 4 is about the threshold probability of scale of five and four response categories, both of which have different probability information. Opinion of (DiStefano & Morgan, 2010; Meiser, 2017) that the number of Andrich threshold is equal to the number of scale categories (k-1).This means that the Andrich threshold on the scale of the five response categories in the figure has four lines, while the scale of the four response categories with three lines. The meaning of the cut lines can provide decisions about which scale is more effectively used in measuring ELVI instruments.The distance between the cut lines should be noted, this is in line with the explanation (Meiser, 2017) the distance between adjacent thresholds is not significant with alpha = 0.05. The probability of the scale of four response categories indicates close distances compared to the scale of five response categories. Thus, it can be stated that the polytomous type scale of five responses is more effectively used to measure ELVI instruments that have a functioning mean value which is "occasional."

In this table, the Standardized Residual Correlation (SRC) value is presented to measure the mismatch of the scale of five and four response categories in the ELVI instrument (Maydeu-olivares et al., 2017). SRC values are based on item correlation values from negative to positive ranges which are then compared with significance values ($p < \alpha = 0.05$) (Gonza et al., 2014).Rasch modeling with the Winsteps program version 4.0.1 obtained SRC value of items -0.177 to 0.638 on scale of five response categories and -0.195 to 0.565 on scale of four response categories. Testing the correlation of the two scales can be seen in Table 7 below:

Table 7 Correlation of scale of five and four response categories

Reliability (x)	Reliability (y)	Correlation	Disattenuated Correlation
0,92	0,95	0,095	0,098

Table 7 shows the reliability value for the scale of five response categories of 0.92 and reliability of 0.95 for the scale of four response categories, both of which have an ideal reliability value.The correlation of ELVI instruments with both scales is of 0.095 or 0.95% of the variance distributed, but this correlation is weakened by measurement error.Erasing the measurement error through disattenuated correlation with the resulting value of 0.098 or 0.98% with an increase of 0.3% from the previously observed correlation.Thus it shows that the correlation scale of the five response categories based on items is higher than scale of the four response categories.The observed correlation based on both scales is 0.098 after correction of attenuation.This was interpreted as not statistically significant ($p < \alpha = 0.05$) in the expected direction.

Following up on this, referring to research conducted by (Wang et al., 2014) to support a significant SRC value at $\alpha = 0.05$, Bonferroni correction was used. Bonferroni correction determines the cut-off is significant at α / n with a value of $\alpha = 0.05$ and n is the number of independent variables tested (Maydeu-olivares et al., 2017), due to the increased risk of type I errors, namely: concluding that errors made in research rejecting the null hypothesis (H_0), even though the null hypothesis is true (Armstrong, 2014; Maydeu-olivares et al., 2017). Calculations with Bonferroni produce significant values, namely: $0.05 / 2 = 0.025$. This can be interpreted that the statistical correlation value with Bonferroni ($0.025 < 0.05$) then H_0 is rejected, thus the SRC validity of Andrich threshold based on Rasch modeling of the response of the ELVI instrument with scale of five response categories is more effective than scale of the four response categories.

CONCLUSION

This study shows the differences in the validity of the Andrick threshold of the ELVI instrument on a scale of five and four response categories based on the Rasch modeling. This model provides an overview of the characteristics of items and respondents on a rating scale (logit scale). Rasch modeling combined from various empirical opinions provides effective information on psychometric concepts.Modification of the ELVI instrument from LER can be used as empirical support to state that this emotional learning measurement instrument has a good psychometric assurance.

This can be shown in the Cronbach alpha reliability values possessed by the two scales in the very ideal category. The validity of Andrick threshold plays an important role in following up the effectiveness of using one of the scales in measuring ELVI instruments. The results show that a scale of five response categories is effectively used to measure the ELVI instrument.Developed phenomenon of testing the ELVI instrument showed significant results in its use. In addition, the modified statement items meet the psychometric criteria, thus the ELVI instrument can be used to measure emotional learning.

REFERENCES

Abd-el-fattah, S. M. (2015). Rasch Rating Scale Analysis of the Arabic Version of the Physical Activity Self-Efficacy Scale for Adolescents : A Social Cognitive Perspective. (December), 2161–2180.

Abdullah, H., Arsad, N., Hanim, F., Abdul, N., Amin, N., & Hamid, S. (2012). Evaluation of Students ' Achievement in the Final Exam Questions for Microelectronic (KKKL3054) using the Rasch Model. 60(c), 119–123. https://doi.org/10.1016/j.sbspro.2012.09.356

Adelson, J. L., & McCoach, D. B. (2010). Measuring the mathematical attitudes of elementary students: The effects of a 4-point or 5-point likert-type scale. Educational and Psychological Measurement, 70(5), 796–807. https://doi.org/10.1177/0013164410366694

Andrich, D. (2011). Rating scales and Rasch measurement. 11(5), 571–585.

Andrich, D. (2016). Georg Rasch and Benjamin Wright's Struggle With the Unidimensional Polytomous Model With Sufficient Statistics. Educational and Psychological Measurement, 76(5), 713–723. https://doi.org/10.1177/0013164416634790

Armstrong, R. A. (2014). When to use the Bonferroni correction. https://doi.org/10.1111/opo.12131

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis : A primer for school psychology researchers and practitioners. Cogent Education, 25, 1–13. https://doi.org/10.1080/2331186X.2017.1416898

Chao, T., Dede, C., & Star, J. R. (2016). Using Digital Resources for Motivation and Engagement in Learning Mathematics : Reflections from Teachers and Students. Digital Experiences in Mathematics Education, 253–277. https://doi.org/10.1007/s40751-016-0024-6

DiStefano, C., & Morgan, G. B. (2010). Evaluation of the BESS TRS-CA Using the Rasch Rating Scale Model. School Psychology Quarterly, 25(4), 202–212. https://doi.org/10.1037/a0021509

Distefano, C., Greer, F. W., Kamphaus, R. W., & Brown, W. H. (2015). Using Rasch Rating Scale Methodology to Examine a Behavioral Screener for Preschoolers At Risk. https://doi.org/10.1177/1053815115573078

Elisabet, L., Benito, G., & Miguel, A. (2012). An Outcomes-Based Assessment of Quality of Life in Social Services. 81–93. https://doi.org/10.1007/s11205-011-9794-9

Ghosh, P. (2015). Historical Perspectives of Classroom Learning Environment (1920-Present). (July), 436–437.

Gonza, L., Zabalegui-ya, A., Lo, J. A., & Siso, M. D. N. A. (2014). Ethical behaviour in clinical practice : a multidimensional Rasch analysis from a survey of primary health care professionals of Barcelona (Catalonia ,

Spain). https://doi.org/10.1007/s11136-014-0720-x
Green, P. E. (2010). Information and to Categories. Response, 34(3), 33–39.
Harachi, T. W. (2012). NIH Public Access. 7(1), 16–38.
Huang, H. (2016). Mixture Random-Effect IRT Models for Controlling Extreme Response Style on Rating Scales. 7(November), 1–15. https://doi.org/10.3389/fpsyg.2016.01706
Huberty, J., Vener, J., Gao, Y., Matthews, J. L., Ransdell, L., & Elavsky, S. (2013). Developing an instrument to measure physical activity related self-worth in women : Rasch analysis of the Women ' s Physical Activity Self-Worth Inventory. Psychology of Sport & Exercise, 14(1), 111–121. https://doi.org/10.1016/j.psychsport.2012.07.009
Jae Jeong, H. (2016). Item Response Theory-Based Evaluation of Psychometric Properties of the Safety Attitudes Questionnaire—Korean Version (SAQ-K). Biometrics & Biostatistics International Journal, 3(5). https://doi.org/10.15406/bbij.2016.03.00079
Jones, S. M., Doolittle, E. J., Greenberg, M. T., Domitrovich, C. E., Weissberg, R. P., Durlak, J. A., … Yeager, D. S. (2017). Social and Emotional Learning 3 Social and Emotional Learning: Introducing the Issue 33 SEL Interventions in Early Childhood 73 Social and Emotional Learning Programs for Adolescents. 27(1).
Kaliski, P. K., Wind, S. A., Engelhard, G., Morgan, D. L., Plake, B. S., & Reshetar, R. A. (2013). Educational and Psychological Measurement. https://doi.org/10.1177/0013164412468448
Kean, J., Bisson, E. F., Brodke, D. S., Biber, J., & Gross, P. H. (2018). An Introduction to Item Response Theory and Rasch Analysis: Application Using the Eating Assessment Tool (EAT-10). Brain Impairment, 19(1), 91–102. https://doi.org/10.1017/BrImp.2017.31
Koul, R. B., Fraser, B. J., Maynard, N., & Tade, M. (2018). Evaluation of engineering and technology activities in primary schools in terms of learning environment, attitudes and understanding. Learning Environments Research, 21(2), 285–300. https://doi.org/10.1007/s10984-017-9255-8
Kupana, N. (2015). Social Emotional Learning and Music Education. SED Journal of Art Education, 3(1). https://doi.org/10.7816/sed-03-01-05
Kutlay, S., Küçükdeveci, A. A., Gönül, D., & Tennant, A. (2018). Adaptation and validation of the Turkish version of the Rheumatoid Arthritis Quality of Life Scale. Rheumatology International, 23(1), 21–26. https://doi.org/10.1007/s00296-002-0247-2
Lapoint, V., & Butty, J. M. (2009). Encyclopedia of Cross-Cultural School Psychology. In Encyclopedia of Cross-Cultural School Psychology. https://doi.org/10.1007/978-0-387-71799-9
Lizzio, A., Wilson, K., & Simons, R. (2010). Studies in Higher Education University Students ' Perceptions of the Learning Environment and

Academic Outcomes : Implications for theory and practice. (September 2013), 37–41. https://doi.org/10.1080/03075070120099359

López, V., Torres-Vallejos, J., Ascorra, P., Villalobos-Parada, B., Bilbao, M., & Valdés, R. (2018). Construction and validation of a classroom climate scale: a mixed methods approach. Learning Environments Research, 21(3), 407–422. https://doi.org/10.1007/s10984-018-9258-0

Lowe, R. (2014). Embodiment in emotional learning, decision making and behaviour: The "what" and the "how" of action. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8515 LNCS(PART 3), 672–679. https://doi.org/10.1007/978-3-319-07446-7_64

Lundgren-Nilsson, Å., Dencker, A., Jakobsson, S., Taft, C., & Tennant, A. (2014). Construct validity of the Swedish version of the revised piper fatigue scale in an oncology sample - A rasch analysis. Value in Health, 17(4), 360–363. https://doi.org/10.1016/j.jval.2014.02.010

Marchesi, B. A. G., & Cook, K. (2012). Social and emotional learning as a catalyst for academic excellence. White Paper: ICF International, 1–9.

Maydeu-olivares, A., Shi, D., Rosseel, Y., Maydeu-olivares, A., Shi, D., & Rosseel, Y. (2017). Assessing Fit in Structural Equation Models : A Monte-Carlo Evaluation of RMSEA Versus SRMR Confidence Intervals and Tests of Close Fit Assessing Fit in Structural Equation Models : A Monte-Carlo Evaluation of RMSEA Versus SRMR Con fi dence Intervals and Tests of Close Fit. Structural Equation Modeling: A Multidisciplinary Journal, 00(00), 1–14. https://doi.org/10.1080/10705511.2017.1389611

McDonald, S. E., Vidacovich, C., Ascione, F. R., Williams, J. H., & Green, K. E. (2015). The children's treatment of animals questionnaire: A rasch analysis. Anthrozoos, 28(1), 131–144. https://doi.org/10.2752/089279315X14129350722172

Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models : A review and tutorial. 159–181. https://doi.org/10.1111/bmsp.12086

Melnick, H., Cook-Harvey, C., & Darling-Hammond, L. (2017). Encouraging Social and Emotional Learning In the Context of New Accountability. The Learning Insitute, (April), product/encouraging-social-emotional-learning-new. Retrieved from https://learningpolicyinstitute.org/product/encouraging-social-emotional-learning-new-accountability-brief

Moors, G. (2008). Exploring the effect of a middle response category. 779–794. https://doi.org/10.1007/s11135-006-9067-x

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. International Journal of Science Education, 33(10), 1373–1405.

https://doi.org/10.1080/09500693.2010.511297

Perera, C. J., Sumintono, B., & Jiang, N. (2018). The Psychometric Validation Of The Principal Practices Questionnaire Based On Item Response Theory. International Online Journal of Educational Leadership, 2(1), 21–38. https://doi.org/10.22452/iojel.vol2no1.3

Pretz, C. R., Kean, J., Heinemann, A. W., Kozlowski, A. J., Bode, R. K., & Gebhardt, E. (2016). A Multidimensional Rasch Analysis of the Functional Independence Measure Based on the National Institute on Disability, Independent Living, and Rehabilitation Research Traumatic Brain Injury Model Systems National Database. Journal of Neurotrauma, 33(14), 1358–1362. https://doi.org/10.1089/neu.2015.4138

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the Number of Categories in Agree-Disagree Scales. Sociological Methods and Research, 43(1), 73–97. https://doi.org/10.1177/0049124113509605

Salzberger, T. (2014). The Rasch model (Rasch, 1960) is a model for the measurement of quantitative latent variables (. 43(1978).

Semiar, P., Pendidikan, N., Sebelas, U., & Surakarta, M. (2015). " EMOTIONAL LEARNING " SEBAGAI PENGEMBANGAN PENDIDIKAN KARAKTER Yulia Suriyanti STKIP Persada Khatulistiwa Sintang. (November).

Seol, H. (2016). Using the Bootstrap Method to Evaluate the Critical Range of Misfit for Polytomous Rasch Fit Statistics. https://doi.org/10.1177/0033294116649434

Shih, C., Chen, C., Sheu, C., Lang, H., & Hsieh, C. (2013). Validating and Improving the Reliability of the EORTC QLQ-C30 Using a Multidimensional Rasch Model. Value in Health, 16(5), 848–854. https://doi.org/10.1016/j.jval.2013.05.004

Sinnema, C., Meyer, F., & Aitken, G. (2016). Capturing the Complex , Situated , and Active Nature of Teaching Through Inquiry-Oriented Standards for Teaching. https://doi.org/10.1177/0022487116668017

Sumintono, B. (2015). Pemodelan Rasch pada Asesmen Pendidikan : suatu pengantar. (December).

Swartz, M. K. (2017). Social and Emotional Learning. Journal of Pediatric Health Care, 31(5), 521–522. https://doi.org/10.1016/j.pedhc.2017.06.001

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting Positive Youth Development Through School-Based Social and Emotional Learning Interventions: A Meta-Analysis of Follow-Up Effects. Child Development, 88(4), 1156–1171. https://doi.org/10.1111/cdev.12864

Vaughan, B. (2018). A Rasch analysis of the Revised Study Process Questionnaire in an Australian osteopathy student cohort. Studies in Educational Evaluation, 56(July 2017), 144–153. https://doi.org/10.1016/j.stueduc.2017.12.003

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological Distance Between Categories in the Likert Scale. Educational and Psychological Measurement, 72(4), 533–546. https://doi.org/10.1177/0013164411431162

Wang, L., Ertmer, P. A., Newby, T. J., Wang, L., Ertmer, P. A., Newby, T. J., … Newby, T. J. (2014). Increasing Preservice Teachers ' Self-Efficacy Beliefs for Technology Integration Increasing Preservice Teachers ' Self-Efficacy Beliefs for Technology Integration. 1523. https://doi.org/10.1080/15391523.2004.10782414

Watt, H. M. G., Carmichael, C., & Callingham, R. (2017). Students ' engagement profiles in mathematics according to learning environment dimensions : Developing an evidence base for best practice in mathematics education. https://doi.org/10.1177/0143034316688373

Woodhouse, H. (2017). Contrasting Views of Emotion in Learning: Alfred North Whitehead and Jerome Bruner. Interchange, 48(3), 217–230. https://doi.org/10.1007/s10780-016-9299-1

Yaeger, D. (2017). Social and Emotional Learning Programs for Adolescents. Future of Children, 27(1), 73–94. Retrieved from http://web.b.ebscohost.com.proxygw.wrlc.org/ehost/detail/detail?vid=0&sid=2b78f5d2-09a2-45cc-a273-ae3d107fc5cb%40sessionmgr103&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3D%3D#AN=123568102&db=sih