Implementation of Item Response Theory (IRT) Rasch Model in Psychometric Property Quality Analysis in Anger Management Scale (AMS-9) of Students

Dwi Yan Nugraha[1✉], **Dian Nugraha**[2], **Muh. Daud**[3], **Widyastuti Widyastuti**[4], **Ahmad Ridfah**[5]

[1,3,4,5]Department of Psychology, Universitas Negeri Makassar, Indonesia
[2]Department of Mathematics Education Postgraduate Program, Universitas Negeri Makassar, Indonesia

Article Info	Abstract
	Research on anger management has been widely carried out, especially in Indonesia, so it is necessary to re-evaluate the psychometric characteristics of the anger management scale made by researchers using the Rasch model method. This study aims to analyze the psychometric characteristics of the anger management scale which is used to measure students' ability to manage anger emotions using the Rasch model approach. Respondents in this study were 103 respondents consisting of 35 men and 68 women with an age range of 15-19 years. This study shows that the discriminatory power of items ranges from 0.57 to 0.72. This study also shows that there are three items that are invalid in terms of the validity of the content of the anger management scale, and are classified as valid in terms of construct validity and rating scales which indicate that the answer choices given are understandable and can be differentiated by research respondents. This study also shows that the overall anger management scale is relatively reliable (α = 0.89). Therefore, the anger management scale has good quality psychometric properties and is in accordance with the Rasch model.

✉Correspondence Address :
Department of Psychology, Universitas Negeri Makassar, Indonesia
90222
E-mail : dwiyannugraha20@gamil.com

INTRODUCTION

One of the psychological problems that is often found among high school students is juvenile delinquency (Ryan et al., 2013). Delinquency has become a universal problem that occurs in various countries, including Indonesia (Syariful, 2020). Juvenile delinquency occurs as a result of the individual himself (Alfianur et al., 2020; Gupta et al., 2022; Syariful, 2020; Anjaswarni et al., 2019; Gungea et al., 2017) who are unable to manage their anger, resulting in detrimental actions others and yourself. Goleman (2002) suggests that anger management is the ability to regulate feelings, calm down, escape from anxiety, depression, and offense, with the aim of balancing emotions.

Research on anger management has been carried out in Indonesia (Fitriansyah, 2008; Hudaya, 2015; Nugraha, 2017; Sihombing, 2018). However, there is no standardized anger management scale in Indonesia yet. Therefore, researchers consider it necessary to re-evaluate the psychometric characteristics of the anger management scale which was first created in 2017 using a classical theory approach (classical test theory, CTT). Considering that the classical measurement theory has weaknesses (Sumintono & Widhiarso, 2015), namely 1) The raw score is basically not the measurement result, but the number of responses made by the subject. 2) Raw scores are initial information in the form of summary data in the form of numbers, but do not provide data from a measurement. 3) The raw score has a weak quantitative meaning, because the raw score obtained varies depending on the number of items. Alagumalai et al. (2005) suggests that another weakness of the classical theory approach in analyzing items is in terms of the reliability coefficient which only depends on the number of samples, the measurement scale is not linear, the limited range of scores, and the balance of negative and positive correlations, and scoring. In addition, the classical test theory also has weaknesses in the form of measurement results that depend on the characteristics of the test used, the parameters of the test items depend on the abilities of the respondents, and measurement errors can only be described on a group scale, not individuals (Pratama, 2020).

Based on the weaknesses of the classical test theory approach, a modern theory emerged, namely item response theory to overcome these weaknesses. Sumintono and Widhiarso (2014) argue that Item Response Theory (IRT) is a general framework of mathematical functions that specifically carry out interactions between subjects and items. Unlike the classical theory which focuses on the result score obtained, IRT does not depend on a particular sample of items and subjects. Hambelton and Swaminathan (2013) suggest that IRT has several advantages, such as the score describes the ability of the respondent and does not depend on the difficulty of the item, can relate items to the ability of the respondent, and does not require parallel tests to determine the reliability value. In addition, the IRT also provides an overview related to the results of examining items and the relationship between the abilities measured and the items on the scale (Aprita & Haryati, 2021). Thus, individuals who have high abilities are more likely to give answers that strongly agree compared to individuals who have low abilities (Retnawati, 2014).

The most popular and widely used IRT model for analyzing the Likert scale is the Rasch model (Chan et al., 2021; Sumaryanto & Khumaedi, 2019). The rasch model is a data analysis method that is able to demonstrate the accuracy of the measurement scale concept, both in terms of subjects and items (Sumintono & Widhiarso, 2014). Measurement of the Rasch model in testing in terms of the subject can show patterns of consistent and inconsistent answers from the responses given by the subject, while in terms of items the Rasch model not only measures the reliability index (Cronbach's alpha) but item analysis is carried out to the level of each

item . Then, the Rasch model is very easy to implement and apply with accurate analytical results (Che Lah et al., 2021; Susdelina et al., 2018). This is because the Rasch model can provide data sizes with the same intervals, can overcome the problem of missing data, can provide precise estimates, can detect inaccuracies in measurement models, and provides a measurement scale of the parameters studied (Sumintono & Widhiarso, 2014; Abdullah et al., 2012).

Sumintono and Widhiarso (2014) argue that the concept of objective measurement in the social sciences has five criteria, namely 1) Provides a linear measure with equal intervals. 2) Overcoming missing data. 3) Finding the wrong item. 4) Carry out the proper estimation process. 5) Produce independent measurements of the parameters studied. These five conditions can only be explained by the Rasch model. This anger management measuring tool developed by researchers, besides having a major weakness in terms of its development method which uses classical measurement theory, also has a weakness in terms of the lack of a number of samples, so that there is no standardization of the measuring instrument properly. Sumintono and Widhiarso (2015) suggested that the criteria for sample size in the Rasch model, namely: a) the sample range of 16-36 has a 95% confidence level with calibration items ± 1 logit, b) the sample range of 27-61 has a 99% confidence level with ± 1 logit item calibration, c) the sample range of 64-144 has a 95% confidence level with ± 0.5 logit item calibration, d) the 108-243 sample range has a 99% confidence level with ± 0.5 logit item calibration. The results of the criterion for the number of samples form the basis that it is necessary to re-evaluate the anger management scale made by researchers in 2017 taking into account the number of samples in the Rasch model.

Sumintono and Widhiarso (2015) suggested that theoretically the stability of the item calibration corresponds to the standard error model. Coaley (2010) suggests that the greater the number of samples used, the error

in the SEmean will decrease and the observed mean will be closer to the true mean of the population, so that the results obtained will be more accurate. The sample size used in the Rasch model will determine the stability of the calibration which will determine the accuracy of the level being measured. This study will attempt to take samples that have a 95% confidence level with ± 0.5 logit calibration items.

In addition, the process of evaluating the psychometric characteristics of the anger management scale needs to be carried out properly in order to be able to measure the level of individual ability to manage emotions in real situations (Talakua et al., 2020). A good assessment requires a good scale (Agustin et al., 2018). A good scale is a measuring instrument that has a valid measurement level (Muluki et al., 2020). Meanwhile, a scale that does not have a valid measurement level certainly cannot provide any information regarding actual individual abilities (Solichin, 2017). In fact, the quality of the items on the scale of anger management is still not known, resulting in a pseudo-assessment that has an impact on the ability to disclose the true condition of individuals related to the level of ability to manage their angry emotions (Fauziana & Wulansari, 2021).

Based on the background of the problems in this study, the researcher wanted to analyze the quality of the psychometric characteristics of the anger management scale which is used to measure students' ability to manage anger using the Rasch model approach. This quality is measured based on several indicators, namely the power of the item according to the Rasch model, the validity and reliability of the item.

METHODS

Design

This study used a psychometric quantitative design. Borsboom and Molenaar (2015) suggest that psychometrics is a research design that aims to analyze the quality of the

psychometric property characteristics of a psychological scale (for example, the anger management scale) in order to produce accuracy and precision in the goals measured by that scale.

Respondents

The number of respondents who filled out the anger management scale was 12,533, consisting of 5,024 men and 7,509 women with an age range of 15-19 years. The data is then tabulated to be further analyzed by cleaning the data. Sumintono and Widhiarso (2014) suggested that the technique used in cleaning the data was in the Rasch Model approach, namely Person Fit Order analysis. Determination in conducting Person Fit Order analysis is based on the sum of the average value and standard deviation. These results are then compared with the logit values obtained from each individual. A logit value that is greater than this value indicates a misfit individual.

The results of cleaning the data using person fit order analysis showed that there were 12,430 subjects with indications of misfit, so data from these subjects could not be used for further analysis. Misfit subjects were then eliminated, resulting in data for 103 subjects to be used in conducting further analysis consisting of 35 males and 68 females with an age range of 15-19 years.

Instrument

Data was collected by going through a scale that was created by researchers in 2017. The scale consists of 12 items and four aspects compiled based on Goleman's theory (2002), namely recognizing angry emotions, controlling angry emotions, defuse angry emotions, and expressing angry emotions assertively. This scale uses a 4-point Likert model (Strongly disagree = 1 to strongly agree = 4) with score intervals of 9 to 36. To determine the level of emotional management of respondents, the total score is calculated by adding up the answers from each item score or you can transform the total score into a score/logit value. A higher score indicates the better the respondent is in managing emotions.

Statistic analysis

The data in this study were analyzed using modern theories in psychometrics. The modern theory referred to in this study uses the Rasch model, which examines the quality of items in terms of both validity and reliability. This study used the help of the STATCAL and Winstep programs version 3.73 which were used to assist in analyzing research data.

RESULTS AND DISCUSSION

Characteristics of respondents

Based on the results of data cleaning using the person fit method using the Winstep version 3.73 program, it was found that 103 students collected research data. Then, the results of the descriptive analysis using the STATCAL program presented in Figure 1, show that the respondents consisted of 35 (33.961%) males and 68 (66.019%) females with an age range of 15-19 years, which were dominated by students aged 15 years (65.049%).
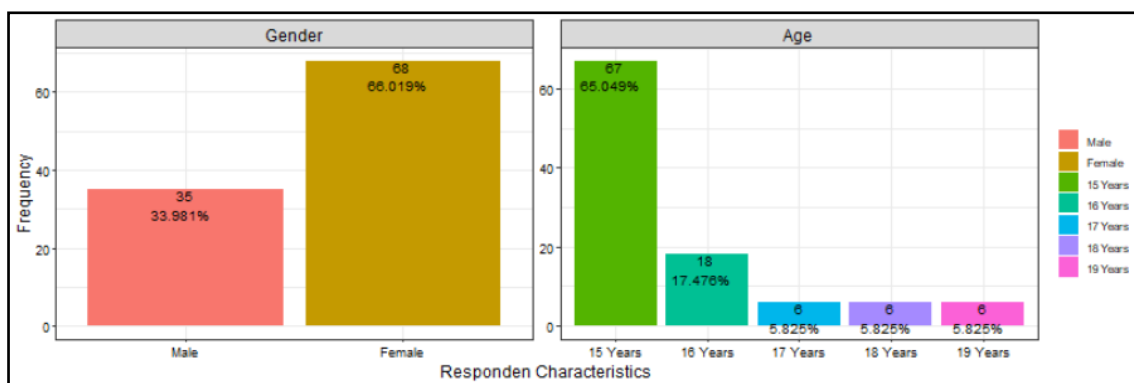
Figure 1. Characteristics of respondents

Item level quality

Based on the results of the analysis of discriminatory power of items using the WINSTEP program version 3.73 presented in table 1, it shows that the PT-Measure Correlation value moves from 0.57 to 0.72, so there are no items that need to be aborted. In addition, the results of the analysis also show that there are three items that do not fit, namely item 3, item 5, and item 9, so that the item is said to be invalid in content and must be discarded. Therefore, the final scale only leaves 9 fit items (see Appendix).

Table 1. Distribution of fit level items

No	Initial item serial number	Final item serial number	Discriminat PT-Measure Correlation	Kategori	Outfit MNSQ	ZSTD	Kategori
1	I1	I1	0.68	Fit	0.56	-0.4	Valid
2	I2	I2	0.67	Fit	0.54	-0.6	Valid
3	I3	-	0.67	Fit	0.02	-3.6	Tidak Valid
4	I4	I3	0.66	Fit	0.82	-0.1	Valid
5	I5	-	0.58	Fit	9.90	9.9	Tidak Valid
6	I6	I4	0.57	Fit	0.86	0.0	Valid
7	I7	I5	0.70	Fit	0.59	-0.5	Valid
8	I8	I6	0.62	Fit	0.69	-0.1	Valid
9	I9	-	0.60	Fit	0.08	-7.6	Tidak Valid
10	I10	I7	0.61	Fit	0.81	0.0	Valid
11	I11	I8	0.72	Fit	0.78	0.0	Valid
12	I12	I9	0.60	Fit	0.82	0.1	Valid

Validity and reliability

Based on the results of the construct validity analysis using the WINSTEP version 3.73 program presented in table 2, it shows that the Raw variance explained by measures empirically obtained 56.0% while the rasch model predicts 56.0%, so that the empirical construct validity is equal to the value obtained predicted by the Rasch model. This value has met the unidimensionality requirement of 100%. The variance values that cannot be explained by this scale are all below 15%, namely: 9.0%, 6.9%, 6.5%, 5.4% and 5.1%.

Table 2. Construct validity results

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)					
			-- Empirical --		Modeled
Total raw variance in observations	=	20.5	100.0		100.0
Raw variance explained by measures	=	11.5	56.0		56.0
Raw variance explained by persons	=	8.4	40.9		40.8
Raw Variance explained by items	=	3.1	15.1		15.1
Raw unexplained variance (total)	=	9.0	44.0	100.0	44.0
Unexplned variance in 1st contrast	=	1.8	9.0	20.4	
Unexplned variance in 2nd contrast	=	1.4	6.9	15.7	
Unexplned variance in 3rd contrast	=	1.3	6.5	14.7	
Unexplned variance in 4th contrast	=	1.1	5.4	12.4	
Unexplned variance in 5th contrast	=	1.0	5.1	11.6	

Then, the results of the validity analysis of the rating scale using the WINSTEP version 3.73 program are presented in table 3, showing that the Andrich Threshold value moves from None to negative and continues to positive sequentially, so that the options given are valid for respondents.

Table 3. Results of the rating scale validity

Category Label	Observed Score	Count	%	Obsvd Avrge	Sample Expect	Infit Mnsq	Outfit Mnsq	Andrich Threshold	Category Measure
SD	1	32	3	-11.71	-11.7	0.96	0.82	NONE	(-12.38)
D	2	222	24	-3.31	-3.31	1.00	0.78	-11.28	-6.89
A	3	664	72	1.44	1.44	1.00	0.97	-2.50	5.64
SA	4	9	1	15.43	15.43	0.90	0.06	13.77	(14.87)

Note. SD = Strongly Disagree; D = Disagree; A = Agree; SA = Strongly Agree.

In addition, the results of the reliability analysis showed that the reliability value of the respondent's level was 0.00 and the item level reliability value was 0.00, which means that the consistency of the respondents' answers is weak, so the quality of the items on this scale is also weak because it is smaller than 0.67. However, the total reliability value between the respondent and item levels is classified as reliable based on the Cronbach Alpha (α) value of 0.90 greater than 0.70.

Scale normalization

Based on the results of data analysis using the WINSTEP program version 3.73 which is presented in table 4, the norm for the anger management scale is obtained. The results of this scale normalization were formed into five categories, namely 2 (1.942%) students who were in the very high category, 27 (26.214%) students who were in the high category, 44 (42.718%) students who were in the medium category, 24 (23.301%) students who are in the low category, and 6 (5.825%) students who are in the very medium category.

Table 4. Normalization of the anger management scale

Formula	Norm	Category	Frequency	Percentage (%)
$1.5 < \mu$	8 >	Very High	2	1.942
$0.5\sigma < \mu \leq 1.5\sigma$	3-7	High	27	26.214
$-0.5\sigma < \mu \leq 0.5\sigma$	-1-(2)	Medium	44	42.718
$-1.5\sigma < \mu \leq -0.5\sigma$	-6-(-2)	Low	24	23.301

$\mu \leq -1.5\sigma$	< -7	Very Low	6	5.825

Note. μ = average logit score of students; σ = standar deviasi.

Categorization based on the characteristics of the respondents

Based on the results of data analysis using the STATCAL program presented in Figure 2, the distribution of respondents' categorization based on sex and age characteristics was obtained. The results obtained show that the majority of male students (27.18%) and female students (15.53%) are in the medium category. In addition, Figure 2 also reports that the majority of students aged 15 years (27.18%), 17 years (2.91%), and 19 years (3.88%) are in the medium category. However, students who are 18 years old are in the same two categories, namely medium and high with a percentage of 2.91% each. Meanwhile, the majority of students aged 16 years (8.74%) were in the low category for managing emotions.
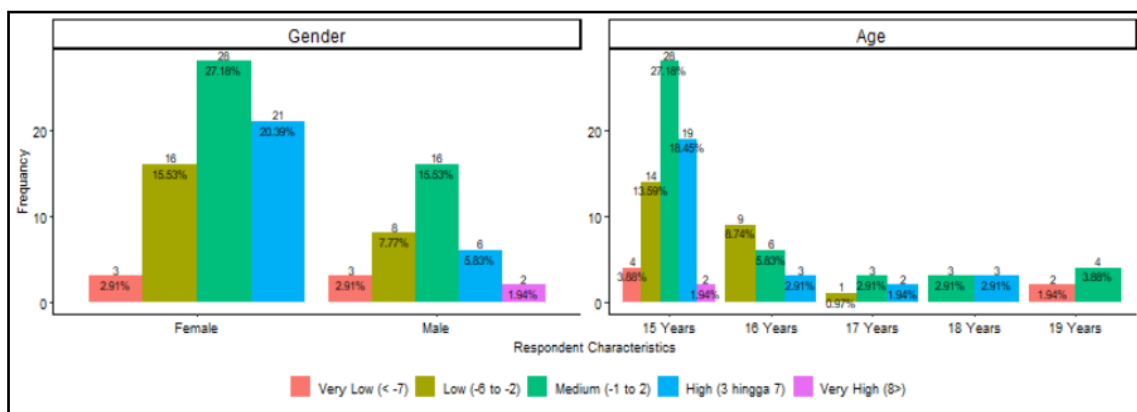

Figure 2. Distribution of respondent's characteristic categorization

Data processing in this study produces psychometric characteristics of the anger management scale obtained from the Rasch model. This study obtained a value of the discriminatory power of items that moved from 0.57 to 0.72. Sumintono and Widhiarso (2014) suggested that the discriminatory power of items in the Rasch model method can be seen from the PT Measure Correlation value of each item. This can be known through the criteria set by Sumintono and Widhiarso (2014), namely 0.40 < PT-Measure Correlation < 0.85. The PT Measure Correlation value generated on this scale indicates that each item meets the criteria, so there are no items that need to be aborted. Sumintono and Widhiarso (2014) suggested that content validity in the Rasch model method can be seen from the Outfit MNSQ-ZSTD value of each item. This can be known through the criteria set by Sumintono and Widhiarso (2014), namely 0.5 < MNSQ < 1.5 and -2.0 < ZSTD < +2.0. The MNSQ-ZSTD Outfit value generated on this scale indicates that there are three items that do not meet the criteria, namely item 3, item 5, and item 9, so these items need to be aborted.

Items that do not meet the content validity criteria are considered to have poor internal language structure (e.g., item 3: *"If my task is difficult, I want to tear it all apart"*), cannot describe the measuring construct of the measurement scale (e.g., item 5: "I hold grudges against the actions of people who make me angry"), and cannot reveal the actual condition of the respondent (e.g., item 9: "I rarely do activities to relieve the angry emotions I experience"). This is in line with previous research which reported that there are five parameters of content validity, namely 1) the extent to which the contents of the items are relevant and represent the constructs targeted in the assessment objectives, 2) respondent responses, 3) internal structure, 4)

relationships with other variables, and 5) consequences (Yusoff, 2019).

Sumintono and Widhiarso (2014) argued that the construct validity in the rasch model method can be seen from the value of the Raw variance explained by measures. This can be seen through the criteria set by Sumintono and Widhiarso (2014), that is, if the value is at least 20%, it means that it can be fulfilled, if the value is more than 40%, it means better, if the value is more than 60%, it means special. The construct validity on this scale as seen from the Raw variance explained by measures empirically obtained 56.0% while the rasch model predicts 56.0%, so that this value meets the unidimensionality requirement of 100%. The variance values that cannot be explained by the scale are all below 15%, namely: 9.0%, 6.9%, 6.5%, 5.4% and 5.1%. The unidimensionality value indicates that this scale measures one dimension, namely anger management.

These results are consistent with the study of Stith and Hamby (2002) who obtained a good construct validity value using the confirmatory factor analysis (CFA) method on the anger management scale (AMS) conducted on a sample of college students in Southeast United States. In addition, the results of Boudreaux's research (2014) also reported adequate construct validity on the Attitudes Toward Anger Management Scale (ATAMS), in samples of undergraduate students in Asian and American Indian/Alasan Native. Another study also found good construct validity on the anger management skills test conducted on Shiraz students, Iran (Farschi-Tabrizi, 2020). Then, the results of research in Indonesia also found that the anger management scale had good quality construct validity in student samples. high school (Nugraha, 2017).

Sumintono and Widhiarso (2014) suggested that the validity of the rating scale in the Rasch model method can be seen from the Andrich Threshold value. This can be known through the criteria set by Sumintono and Widhiarso (2014), that is, if you move from None then negative and continue to lead

to positive sequentially, so that the options given are valid for respondents. The value shown from the Andrich Threshold value meets the criteria, so the options given on this scale are valid.

Linacre (2014) suggests that the rasch model approach can determine the appropriateness of the assessment on the reliability index of each answer option given by each statement item, including the frequency of categories and average size, average squared infit and outfit, and threshold calibration. . The threshold value must move from none to negative and continue to positive sequentially representing each category of answer options, for example strongly disagree to strongly agree (Zile-Tamsen, 2017; Bond & Fox, 2012).

Sumintono and Widhiarso (2014) suggested that the criteria for the value of Person Reliability and Item Reliability are, if the reliability value obtained is less than 0.67 it means weak, if the reliability value obtained is 0.67-0.80 it means sufficient, if the reliability value is obtained 0.81-0.90 means good, if the reliability value obtained is less than 0.91-0.94 it means very good, and if the reliability value obtained is greater than 0.94 it means special. The criterion for the Cronbach Alpha value is, if the reliability value obtained is less than 0.50 it means bad, if the reliability value obtained is 0.50-0.60 it means bad, if the reliability value obtained is 0.61-0.70 it means enough, if the reliability value obtained is 0.71-0.80 it means good, and if the reliability value obtained is greater than 0.80 it means very good. Reliability on this scale using the Rasch model method shows that the reliability value of the subject obtained is 0.00 and the item reliability value is 0.00 which means that the consistency of the answers from the respondents is classified as weak, so the quality of the items on this scale is also relatively weak. The overall reliability value of the anger management scale obtained from the interaction between subjects and items is obtained from the Cronbach Alpha value of 0.90 which means reliable and belongs to the very good category.

Some of the results of previous studies also explained that the anger management scale has an internal consistency value that is included in the reliable category, such as the study of Stith and Hamby (2002) on students in Southeast United States (AMS 36 item α = 0.87; AMS 20 item α = 0 , 70). However, for the AMS 12 item (α = 0.51) it has a low internal consistency value. Likewise with the results of the study (Farschi-Tabrizi, 2020) which also reported that there was a low internal consistency value on the anger management skills test scale (α = 0.62). In contrast to the Boudreaux study (2014) which obtained high internal consistency values for each factor in ATAMS (Belief in Treatment, α = 0.93; Receptiveness, α = 0.73), and is also in line with other studies which obtained values good internal consistency (Nugraha, 2017), among high school students in Indonesia, particularly in the South Sulawesi region.

The norm used in this study is the normalization with five categories that move from very low to very high. Azwar (2017) suggests that categorization is based on the assumption that individual scores in the group are estimates of the individual scores in the population which are normally distributed, so that theoretical score category boundaries are formed which are distributed according to the standard norm model with the formula: a) $1.5 < \mu$ with very high category. b) $0.5\sigma < \mu \leq 1.5\sigma$ with high category. c) $-0.5\sigma < \mu \leq 0.5\sigma$ in the moderate category. d) $-1.5\sigma < \mu \leq -0.5\sigma$ with low category. e) $\mu \leq -1.5\sigma$ with very low category. The norms obtained from this scale can be used appropriately if there is compatibility with the characteristics of the subjects in this study, so it is necessary to pay attention to the characteristics of the subjects in this study. The characteristics of the subjects in this study are illustrated by the subject's demographic data, namely students who are dominated by women (66.019%), and are 15 years old (65.049%).

In general, this study provides good quality evidence of psychometric characteristics in a sample of high school students in Indonesia. However, this research has limitations, including that the respondents involved were only a limited sample of students in the South Sulawesi region. Therefore it is very interesting to expand the research location in several cities in Indonesia. This research can also be extended to the undergraduate level and to the general adolescent population.

CONCLUSION

Based on the results of the analysis, it was obtained that the overall anger management scale had good quality psychometric properties and was in accordance with the Rasch model. This is indicated by the value of the discriminating power of the items in this study moving from 0.57 to 0.72, so that there are no items that need to be aborted. However, the results of the content validity analysis showed that there were three items that did not meet the criteria, namely items 3, 5 and 9, so that the three items were declared invalid and had to be aborted. The results of the construct validity analysis in this study indicate that the unidimensionality value of 100% has been fulfilled, so that this scale measures one construct, namely angry management. Then, analysis of the validity of the rating scale also shows that the answer choices contained in this scale are valid for respondents. In addition, the results of the reliability analysis show that overall this scale is in the reliable category based on the Cronbach's alpha reliability value (α = 0.90). However, when viewed based on the reliability of the subject and item, unsatisfactory results are obtained, meaning that the consistency of the respondents' answers is classified as weak. Therefore, future research must re-evaluate the reliability value of this scale using various methods such as test-retest reliability. Then, the norms used from this scale must of course pay attention to the characteristics of the research respondents if they are to be applied for research and identification purposes.

REFERENCES

Abdullah, H., Arsad, N., Hashim, F. H., Aziz, N. A., Amin, N., & Ali, S. H. (2012). Evaluation of Students' Achievement in the Final Exam Questions for Microelectronic (KKKL3054) using the Rasch Model. *Procedia - Social and Behavioral Sciences*, *60*(c), 119–123.

Agustin, N., Susilogati, S., & Addiani, K. (2018). Desain Instrumen Tes Bermuatan Etnosains Untuk Mengukur Kemampuan Berpikir Kritis Siswa Sma. *Jurnal Inovasi Pendidikan Kimia*, *12*(2), 2159-2169.

Alagumalai, S., Curtis, D. D. & Hungi, N. (2005). *Applied rasch measurment: A book of examplars*. Papers in Honour of John P. Keeves. Dordrecht Netherlands: Springer.

Alfianur, A., Ezalina, E., & Fitriami, E. (2020). Kekerasan emosional menyebabkan kenakalan pada remaja. *Holistik Jurnal Kesehatan, 14*(1), 52-58.

Andrich, D. (1978). Rating formulation for ordered response category. *Psychometrika, 43*(4), 561–573.

Anjaswarni, T., Nursalam, N., Widati, S., & Yusuf, A. (2019). Analysis of the Risk Factors Related to the Occurrence of Juvenile Delinquency Behavior. *Jurnal Ners, 14*(2), 129-136.

Aprita, Y. M., & Haryati, T. (2021). Analisis Tingkat Kesesuaian Butir Soal (Item Fit) Menggunakan Model Rasch. *Jurnal AKRAB JUARA*, *6*(1), 180–193.

Azwar, S. (2017). *Penyusunan skala psikologi edisi 2 cetakan XI*. Yogyakarta: Pustaka Pelajar.

Bond, T. G., & Fox, C. M. (2012). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd Ed.). New York: Routledge.

Borsboom, D., & Molenaar, D. (2015). Psychometrics. In J. O. Ramsay (Eds.). *International Encyclopedia of the Social & Behavioral Sciences* (2nd Ed.). (vol. 18, pp. 12416–12422). Elsevier Ltd.

Boudreaux, D. J., Dahlen, E. R., Madson, M. B., & Bullock-Yowell, E. (2014). Attitudes toward anger management scale: Development and initial validation. *Measurement and Evaluation in Counseling and Development*, *47*(1), 14-26.

Chan, S. W., Looi, C. K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. *Journal of Computers in Education*, *8*(2), 213–236.

Che Lah, N. H., Tasir, Z., & Jumaat, N. F. (2021). Applying alternative method to evaluate online problem-solving skill inventory (OPSI) using Rasch model analysis. *Educational Studies*, *00*(00), 1–23.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. London: SAGE Publications.

Farschi-Tabrizi, Y., Borazjani, S., & Rahimi, C. (2020). Validation and Psychometric Properties of the Anger Management Skills Test. *Social Determinants of Health*, *6*(1), 1-8.

Fauziana, A., & Wulansari, A. D. (2021). Analisis Kualitas Butir Soal Ulangan Harian di Sekolah Dasar dengan Model Rasch. *Jurnal Ibriez: Jurnal Kependidikan Dasar Islam Berbasis Sains*, *6*(1), 10-19.

Fitriansyah, N. (2008). Deskripsi kemampuan mengenal emosi dan mengelola emosi pada siswa kelas X SMA Pangudi Luhur Sedayu Yogyakarta tahun ajaran 2007/2008 dan implikasinya terhadap usulan topik bimbingan klasikal. *Skripsi*. Yogyakarta: Program Studi Bimbingan dan Konseling Jurusan Ilmu Pendidikan Fakultas Keguruan dan Ilmu Pendidikan Universitas Sanata Dharma.

Goleman, D. (2002). *Kecerdasan emosional*. Terjemahan oleh T. Hermaya. Jakarta: PT Gramedia Pustaka Utama.

Gungea, M., Jaunky, V. C., & Ramesh, V. (2017). Personality traits and juvenile delinquency. *International Journal of Conceptions on Management and Social Sciences, 5*(1), 42-46.

Gupta, M. K., Mohapatra, S., & Mahanta, P. K. (2022). Juvenile's Delinquent Behavior, Risk Factors, and Quantitative Assessment Approach: A Systematic Review. *Indian Journal of Community Medicine*, *47*(4), 483-490.

Hambelton, R. K., & Swaminathan, H. (2013). *Item Response Theory: Principles and Applications*. Springer Science & Business Media.

Hudaya, N. V. (2015). Peningkatan kemampuan mengelola emosi marah melalui teknik *anger management* pada siswa kelas X TKJ SMK Muhammadiyah 1 Moyudan. *Jurnal Bimbingan dan Konseling Edisi 7 Tahun Ke-4*, 1-9.

Linacre, J. M. (2014). *A user's guide to Winsteps Ministeps Rasch-Model computer programs*. Chicago, IL: Author.

Muluki, A., Bundu, P., & Sukmawati, L. (2020). Analisis Kualitas Butir Tes Semester Ganjil Mata Pelajaran IPA Kelas IV Mi Radhiatul Adawiyah. *Jurnal Ilmiah Sekolah Dasar*, *4*(1), 86–96.

Nugraha, D. Y. (2017). *Laporan konstruksi alat ukur psikologi (Alat ukur anger management)*. Tugas Akhir Mata Kuliah Konstruksi Alat Ukur Psikologi tidak diterbitkan. Makassar: Program Studi Psikologi Fakultas Psikologi Universitas Negeri Makassar.

Pratama, D. (2020). Analisis Kualitas Tes Buatan Guru Melalui Pendekatan Item Response Theory (IRT) Model Rasch. *Tarbawy: Jurnal Pendidikan Islam*, *7*(1), 61–70.

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Nuha Medika.

Ryan, J. P., Williams, A. B., & Courtney, M. E. (2013). Adolescent neglect, juvenile delinquency and the risk of recidivism. *Journal of youth and adolescence*, *42*(3), 454-465.

Sihombing, D, N. (2018). Kemampuan mengelola emosi (Studi deskriptif pada mahasiswa yang sedang menyelesaikan skripsi dan yang baru saja lulus dan program kemampuan mengelola emosi program studi bimbingan dan konseling universitas sanata dharma angkatan tahun 2013). *Skripsi*. Yogyakarta: Program Studi Bimbingan dan Konseling Jurusan Ilmu Pendidikan Fakultas Keguruan dan Ilmu Pendidikan Universitas Sanata Dharma.

Stith, S. M., & Hamby, S. L. (2002). The anger management scale: Development and preliminary psychometric properties. *Violence and victims*, *17*(4), 383-402.

Solichin, M. (2017). Analisis Daya Beda Soal Taraf Kesukaran,Butir Tes, Validitas Butir Tes, Interpretasi Hasil Tes Valliditas Ramalan dalam Evaluasi Pendidikan. *Jurnal Manajemen Dan Pendidikan Islam 2*, *2*(2), 192–213.

Sumaryanto, T., & Khumaedi, M. (2019). the Analysis of Instrument of the Ability To Acting and Thinking Creatively Based Rasch Model. *Journal of Research and Educational Research Evaluation*, *8*(1), 48–56.

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model rasch untuk penelitian ilmu-ilmu sosial edisi revisi*. Cimahi: Trim Komunikata Publishing House.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan cetakan I*. Cimahi: Trim Komunikata.

Susdelina, Perdana, S. A., & Febrian. (2018). Analisis Kualitas Instrumen Pengukuran Pemahaman Konsep Persamaan Kuadrat Melalui Teori Tes Klasik Dan Rasch Model. *Jurnal Kiprah*, *6*(1), 41–48.

Syariful., Khilmiyah, A., & Masruri, S. (2020). Juvenile Delinquency Measurement in Indonesian High Schools. *Systematic Reviews in Pharmacy*, *11*(9), 522-527.

Talakua, M., Ratumanan, T. G., & Tamalene, H. (2020). Komparasi Hasil Belajar Siswa Yang Menggunakan Media Swishmax Dan Tanpa Swishmax Pada Materi Operasi Hitung Pecahan Di Kelas Vii Smp Negeri 3 Ambon. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *14*(1), 031–038.

Van Zile-Tamsen, C. (2017). Using Rasch analysis to inform rating scale development. *Research in Higher Education*, *58*(8), 922-933.

Yusoff, M. S. B. (2019). ABC of content validation and content validity index calculation. *Education in Medicine Journal*, *11*(2), 49–54.

Appendix

Anger Management Scale (AMS-9)

No	Item	SD	D	A	SA
1	When my physical condition began to change, I tried to control myself so that I wasn't angry				
2	I do other activities to reduce my angry emotions				
3	If I'm angry I will be quiet and think positive				
4	Although angry, I have never vilifying people who have made me angry				
5	When I'm angry, I choose to read story books or watch television				
6	If I'm angry, I prefer to sleep and listen to music				
7	I will express the anger I felt in an appropriate way without hurting the feelings of others				
8	If I feel that I disagree with the people's opinion I will them in good words so as not to offend				

9	I don't like to blame others for no apparent reason

Note. SD = Strongly Disagree; D = Disagree; A = Agree; SA = Strongly Agree. Aspect recognizing angry emotions = I1 dan I2; controlling angry emotions = I3 dan I4; defuse angry emotions = I5 dan I6; expressing angry emotions assertively = I7, I8, dan I9.