Advances in Social Science, Education and Humanities Research, volume 640
Proceedings of the 5th International Conference on Current Issues in Education (ICCIE 2021)

An Application of the Rasch Model in Quality Testing of Polytomic Instruments

Yulinda Erma Suryani[1,2*], Amat Jaedun[1]

[1] *Universitas Negeri Yogyakarta, Indonesia*
[2]*Universitas Widya Dharma Klaten*
[*]*Corresponding author. Email: yulinda@unwidha.id*

ABSTRACT
This study aims to apply the Rasch Model in testing the quality of the polytomy instrument. The source of the data is the responses of 288 students of SMPN 2 Prambanan they gave to the Student Perception Scale of School Climate. The Student's Perception of School Climate Scale consists of 20 items. The results of the analysis of the quality of polytomy instruments using the Rasch Mmode can produce more complete instrument characteristics consisting of reliability, dimensionality testing, item difficulty level, item suitability level, item functionality, and rating scale analysis.

Keywords: *Instrument, polytomy, rasch.*

1. INTRODUCTION

The measurement process in quantitative research is essential since the numbers obtained are used to be processed and analyzed so as to get answers to the research questions. Without measurement, no data will be obtained and research will not continue. According to Rasch's modeling, a study relates to three things: instruments, items and respondents. In social sciences such as psychology and education, the object being measured is usually an invisible object or something hidden (latent). In Rasch modeling, the basic idea of the measurement process is called latent traits which are nothing but the main concept of item response theory (IRT). Although these characteristics cannot be observed empirically, the manifestation of these characteristics and their interaction with the environment will give rise to an empirical indicator that can be measured or observed. Thus, the measurement instrument is very important to measure these latent characteristics.

The measurement process in quantitative research is very important because numbers are only obtained from measurement results which are then processed and analyzed so research questions can get answers. In social sciences such as psychology and education, usually the object or attribute being measured is invisible or hidden (latent). Latent attribute means the attribute is hypothetical and cannot be observed directly.

The use of a quantitative approach requires high caution in the quantification process: the process of converting qualitative data into quantitative data. Thus, the instrument used for data retrieval must be scientifically justified. Data collection instruments are usually carried out using a scale. Psychological scales are widely used in education to measure psychological attributes for research purposes. For example research on student learning satisfaction [1], personal direction in learning or student self-esteem [2]. The measurement procedure using a psychological scale is done with several assumptions, for example the respondent knows himself well and understands the statement items accordingly, with those understood by the scalers [3]. Thus, to ensure these assumptions are met, before the psychological scale is given to respondents, several processes have been passed. First, a pilot study on several people to ensure that the items written on the scale are understood. Items difficult to understand are revised or excluded from the scale. Second, field studies (field tests) for the purposes of scale item analysis [4]. Items that do not support scale the whole is separated so that the items in the scale measure the attributes of the

measuring objective. These processes will produce a scale containing items understood, homogeneous, and can optimally distinguish psychological attributes between individuals.

Classical test theory (CTT) has developed widely and become the mainstream among psychologists and educators, as well as other fields of behavioral studies for two centuries [5]. CTT has a weakness because it is examinee sample dependent and item sample dependent ([5], [6], [7], [8). This weakness has triggered a new, more adequate theory, namely the modern test theory, also known as item response theory (TRA) or item response theory (IRT) and also known as latent trait theory (LTT).

In contrast to CTT, which focuses on information at the test level, TRA mainly focuses on information at the item level so that it is expected to cover the shortcomings contained in CTT. The application of the IRT model is based on several assumptions in the form of postulates: (1) the performance of a participant on an item can be predicted by a set of factors called traits, latent traits, or abilities; and (2) the relationship between participant performance on an item and a set of underlying latent abilities (ability) can be described by a monotonically attractive function called item characteristic function or item characteristic curve (ICC) ([6], [9], [10]). So ICC is a depiction in the form of a curve that describes the relationship between latent traits and the subject's performance on an item. The assumptions underlying TRA are unidimensionality, local independence, and parameter invariance. Meanwhile, the most basic assumptions are: (1) each item has a certain item characteristic curve (ICC); and (2) local independence.

Classical tests are still quite popular among researchers and practitioners because they are easy to apply in analyzing test quality. The problem is that behind the convenience there are a number of limitations that have the potential to bias information about test quality. Currently, Rasch modeling (Rasch Model) is available and can produce better and more accurate measurement instruments. So far, only the Rasch Model is an analytical tool that can test the validity and reliability of research instruments, and even test the suitability of persons and items simultaneously, which has not been matched by other analytical techniques. The Rasch Model also has several advantages because it fulfills the five principles of the measurement model, namely: first, it is able to provide a linear scale with equal intervals; second, it can make predictions on missing data; third, it can provide a more precise estimate; fourth, it can detect model inaccuracies; and fifth, it produces replicable measurements. These various advantages should be utilized by researchers to support higher quality research findings. Testing instruments validity is an inevitable and essential element before moving on to inferential statistics to get answers to the research questions posed.

In the concept of developing psychological measurement instruments, the constructs (attributes) related to humans have many characteristics, one of which is latent constructs. That is, these attributes are hypothetical and cannot be observed directly. Methodologically, the use of measurement instruments is a very important part of quantitative research. Reliable and valid instruments will provide reliable information. On the other hand, instruments that fail to meet the requirements will give biased or misleading results so that it can reduce the quality of the research.

Reliability is the accuracy of measurement regardless of what attribute is being measured [11]. Psychometrically, reliability has two meanings [12]: (1) self-consistency or internal consistency, and (2) stability. Consistency is the conformity between parts in a test. If one part of a test measures a certain variable, the other parts, if inconsistent with the first part, do not measure the same variable. Reliability that is based on the fit between the parts in a test of this kind is known as internal consistency reliability. According to [12], this concept of internal consistency reliability underlies the general principle in psychometry which states that reliability (high internal consistency) is one of the prerequisites for validity. Cattell (in [12]) states that the maximum validity will be obtained when the test items are not correlated with each other, but each item is positively correlated with the criteria; such a test will only have low internal consistency reliability. However, in practice the general proposition of psychometrics that valid tests usually have high (internal) consistency remains widely accepted.

The concept of reliability or unreliability or lack of reliability is included in the basic postulates of the classical test model (X + T + E). According to classical test theory, test scores reflect the influence of two factors [13], namely: (a) stable characteristics contained in the testee (pure characteristics), and (b) characteristics in the form of random or arbitrary events originating either from in the testee and from the test situation (random measurement error, abbreviated as RME). The impact of this random event is in the form of RME which results in unreliable measurement results, in the sense that the test score of a testee will fluctuate both positively (increasing score) and negatively (decreasing score) even though they are tested with the same test but on different occasions.

Reliability coefficients estimated using classical test theory have weaknesses due to dependence on the sample, non-linear raw scores, limitations in the range

of scores, and the price can be negative. In contrast to classical reliability which has a single price, reliability in Item Response Theory (IRT)/Rasch Model is between one level of ability with different abilities [14]. The same test will result in different measurement reliability when given to individuals with very high and very low abilities. A single reliability value reported by several IRT/Rasch software (eg Winstep) is a general summary of reliability per individual skill level being measured. A test cannot be estimated as reliable or not because reliability is not an attribute for the instrument but for the score or measurement.

The analysis technique based on the Rasch model is very useful for evaluating instruments or questionnaires used in research. Rasch modeling aims to develop objective measurements. Objective measurement is a measurement whose results depend on who is being measured (test dependent scoring). The percentage or number of correct answers on a test depends on the subject being measured (sample dependent) which is descriptive and applies to that subject. Objective measurement produces data that is free from the influence of the type of subject, the characteristics of the rater and the characteristics of the measuring instrument. The estimation and calibration techniques used in the Rasch modeling have eliminated the influence of these three factors, so by using Rasch modeling, the measurements made will have the same quality as the measurements made in the physical dimensions in the field of physics. Thus, this study aims to find out the results of the analysis of the quality of the polytomy instrument (Students' Perceptions of School Climate) based on Rasch modelling.

2. METHOD

This research is a descriptive quantitative study that aims to evaluate the properties of the scale of students' perceptions of school climate based on Rasch modeling. The data collection method used in this research is the documentation method, namely by using data on the Student Perception Scale of School Climate which were filled out by 288 students of SMPN 2 Prambanan, Klaten. The Student's Perception of School Climate Scale consists of 20 items. In accordance with the research objective, which is to analyze the polytomy instrument using the Rasch model, there were several stages of analysis carried out, namely: reliability testing, dimensionality, item difficulty level, item suitability level, item functionality, and rating scale analysis. Based on Rasch modeling, there were several stages of instrument reliability testing, namely item reliability and person reliability. Based on Rasch modeling, there are several stages of instrument reliability testing, namely item reliability and person reliability. There were several stages in testing the validity of the instrument, namely testing the validity of the respondents, testing the validity of items, and testing the dimensionality of the instrument. Unidimensionality was done with the aim of seeing whether the instrument used for measurement only measured one dimension − a good instrument is if the instrument measures only one dimension. The criteria used were: If the value was 20%, the item was OK. If the value was 40%, it was good. If the value was 60%, it was very good. The purpose of the DIF analysis was to find out whether the items written were biased/favorable to one party, for example gender. The criteria used were: in the table see the PROB value, if prob < 0.05, the item contained bias. Rating Scale Analysis was conducted to find out whether respondents understand the difference in ratings. What was seen was the value of the Observe Average. The resulting Observ Average value had to increase. If the results were irregular then the rating had to be simplified, for example from five answer choices simplified to three answer choices.

3. RESULT AND DISCUSSION

Reliability coefficients estimated using the classical test theory have weaknesses due to dependence on the sample, non-linear raw scores, limitations in the range of scores, and the possibility of negative values. In contrast to classical reliability which has a single price, reliability in the Item Response Theory (IRT)/Rasch Model is between one level of ability with different abilities [14]. The same test will result in different measurement reliability when given to individuals with very high and very low abilities. The single reliability coefficient value reported by the Winstep software is a general summary of the reliability per individual level of ability being measured. A test cannot be estimated as reliable or not because reliability is not an attribute for the instrument but for the score or measurement.

Based on the results of the reliability analysis, the person reliability value is 0.75, the instrument reliability coefficient is 0.98, and the test reliability coefficient is 0.80. It can be concluded that the reliability of the instrument of student perception of the school climate is in the good category. The information function expresses the strength or contribution of the test in revealing the latent trait measured by the test. The results of the analysis of the information function of the School Climate instrument are presented in Figure 1.

Unidimensional is done with the aim of seeing whether the instrument used for measurement measures only one dimension. An instrument is good if it measures only one thing. The results of the dimensionality test show that the raw variance explained by measures was 33.9%. A good unidimensional value

is close to 40%. Based on the results of the analysis, it can be concluded that the dimensionality of the school climate instrument is in the good category.

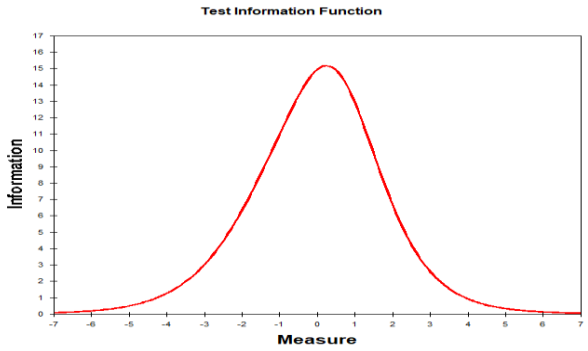

Figure 1 Test information function

Instrument validity is how far the measurement by the instrument can measure what attributes should be measured. There are various opinions regarding the validity of the instruments used in the measurement both in education and psychology. According to the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA, APA, and NCME) on standards for educational and psychological testing, validity refers to the degree to which facts and theory support the interpretation of test scores and is the most important consideration. Important in test development. Other experts argue that the validity of a measuring instrument is the extent to which the measuring instrument is able to measure what it is supposed to measure ([11], [15]-[17]). Validity refers to the adequacy and appropriateness of interpretations made from the assessment, with regard to specific uses [18]. Validity is an integrated evaluative policy about the extent to which empirical facts and theoretical reasons support the adequacy and suitability of inferences and actions based on test scores or scores of an instrument. Based on the various opinions above, it can be concluded that validity will show support for empirical facts and theoretical reasons for the interpretation of test scores or scores of an instrument, and is related to the accuracy of measurement.

An item analysis was carried out with the aim of knowing whether the questions on the School Climate instrument could be understood well by students. With Rasch modeling, it can be seen that the quality of measurement information is good and informative. To determine the quality of the questions, the item analysis carried out is the level of difficulty of the questions, the level of suitability of the questions, and the detection of bias. The difficulty level of the item is a parameter that describes how difficult it is for a group of participants to agree on a statement that is in accordance with the item.

Based on the results of data analysis, item B1 has the highest measure value of 1.25, and thus it can be said that item B1 is the item with the highest difficulty so that it is the item that is the most difficult for respondents to agree with. Item A7 is the item most easily approved by respondents with a measuring value of -1.37. A high logit value indicates a high level of problem difficulty. This correlates with the total score column, which states how many statements the respondent agrees with.

In Rasch modeling, in addition to the item difficulty level, other valuable information is looking at the quality of item fit with the model or what is called item fit. Item fit explains whether the items function normally to measure or not. If a question is found to be unfit, it is an indication that there is a respondent's misconception of the item.

The value of outfit means square, outfit z-standard and point measure correlation are the criteria used to see the level of item fit ([19] and [20]). If the items in the three criteria are not met, it can be ascertained that the items are not good enough so they need to be repaired or replaced. This is done to ensure that the level of understanding of the respondents is indeed tested through appropriate and quality items. Item fit indicators for all items are Outfit Means Square ($0.5 < MNSQ < 1.5$); Outfit Z-standard ($-2.0 < ZSTD < +2.0$) and Point Measure Correlation ($0.4 < Pt\ Measure\ Corr < 0.85$). Based on these criteria, it can be seen that items A7 and B4 are misfit items because they do not meet the three criteria. Therefore, items A7 and B4 must be discarded. Items A4, B1, B2, C3, B6, B3, B5, B8, B9, A5, C1, A3 can still be revised or improved because only one criterion has not been met. It can be concluded that of the 20 items on the scale of student perceptions of the school climate, there are two items that fail, 12 items are revised, and six items are good. To be clearer, it can be seen from the item characteristic function or item characteristic curve (ICC).

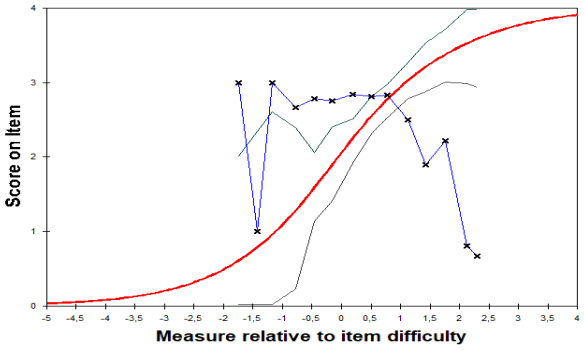

Figure 2 ICC items that must be discarded

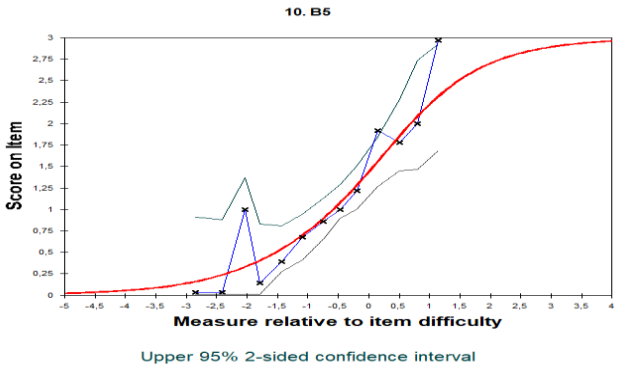

Figure 3 ICC items that must be revised

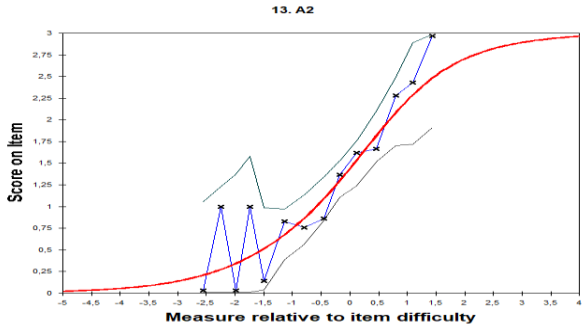

Figure 4 Good item ICC

A measurement is valid when the items used do not contain bias. An instrument or item is called biased if it is found that one individual with certain characteristics is more advantageous than individuals with other characteristics. The DIF analysis of the instrument in this study used gender and class demographic data. Both of these demographic data are used to detect bias. An item is said to contain bias if it is found that the probability value of the item is below 5% or 0.05.

Based on the results of the DIF Gender analysis, it can be seen that there are six items whose probability values are below 5% or 0.05, namely items B1, B2, A1, C3, A6 and A7. So the six items contain a gender bias. A more complete explanation can be seen in the DIF Plot diagram below.

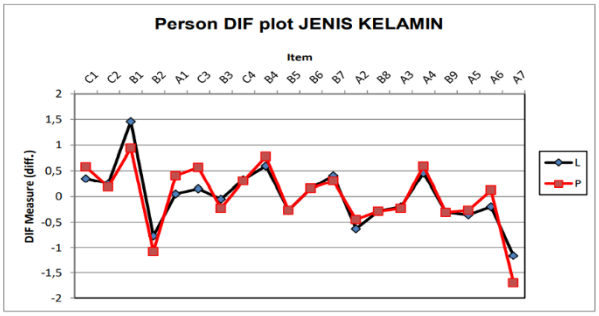

Figure 5 Person DIF plot gender

The DIF graph for gender demographic data in figure 5, it can be seen that item B1 is more difficult for male students to agree with than female students. Item B2 and item A7 are easier for female students to agree with. Meanwhile, items A1, C3, and A6 are more difficult to be approved by female students.

In addition to gender bias analysis, this study also analyzes bias based on class. Respondents in this study consist of classes VII, VIII and IX. There are 10 items that are class biased because the probability value of the items is below 5% or 0.05, namely: items C1, C2, B2, C3, C4, A2, B8, A3, A4 and A7. The results of the class DIF analysis can also be seen in the chart below.

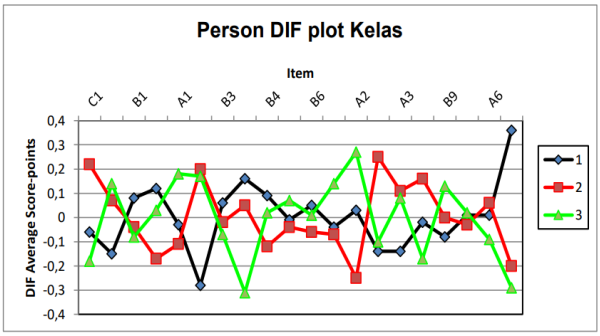

Figure 6 Person DIF Plot Class

Figure 6 shows that items C2, C3, A3 are more easily approved by grade VII students. Items B2 and A2 are more easily approved by grade VIII students, while items C4 are easier to be approved by grade IX students.

The rating scale analysis shows the validity of the response rating. The aim is to find out whether the respondents understand about the difference in ratings. The rating on the School Climate instrument is good so it can still be used by using four answer choices. The rating scale used in the instrument can be understood by the respondent's choice. The results of the rating scale analysis can be seen in Figure 7.

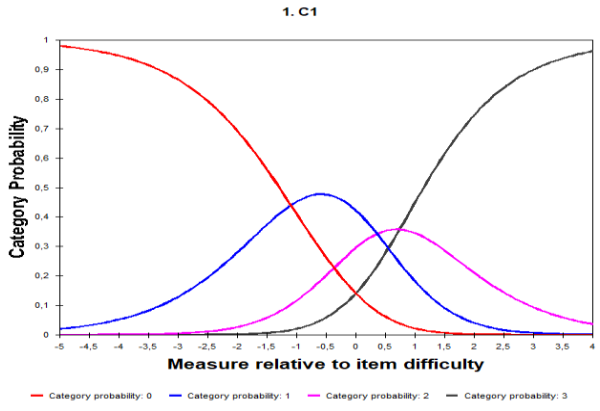

Figure 7 Rating Scale Analysis

Figure 7 shows that the separation between the answer choices can be seen clearly, which shows that respondents are not confused by the rating choices given.

4. CONCLUSION

The reliability coefficient value of the school climate instrument of 0.98 is included in the good category. The dimensionality test results are close to 40%, including in the good category. Item B1 is the item with the highest difficulty with a measure value of 1.25. While item A7 is the easiest item to be approved by respondents with a measure value of -1.37. Items A7 and B4 are misfit items that must be discarded, while items A4, B1, B2, C3, B6, B3, B5, B8, B9, A5, C1, A3 need to be revised. Items B1, B2, A1, C3, A6 and A7 contain gender bias. Items C1, C2, B2, C3, C4, A2, B8, A3, A4 and A7 contain class bias. The analysis rating scale on the School Climate instrument is good, so it can still be used by using four answer choices.

The results of the analysis of the quality of polytomy instruments using the Rasch Model can produce more complete instrument characteristics. Researchers in the social and educational fields can use the Rasch Model in testing the quality of the research instruments to be used. Researchers who will use the School Climate instrument can use it by paying attention to the items that must be discarded and those that must be revised.

AUTHORS' CONTRIBUTIONS

YES performs data collection, data analysis and prepares the manuscript; AJ conducted research direction, experimental design, and completion of manuscript.

ACKNOWLEDGMENTS

Our gratitude goes to the students of SMPN 2 Prambanan Klaten, who have taken the time to become respondents in this study.

REFERENCES

[1] Hostetter, Busch, Journal of Scholarship of Teaching and Learning, Vol. 6, No. 2, pp. 1 – 12, 2006.

[2] P. Kususanto, M. Chua, Students' self-esteem at school the risk, the challenge, and the cure, Journal of Education and Learning, 6, pp.1-14, 2012.

[3] S. Hadi, Analisis butir untuk instrumen angket, tes dan skala nilai dengan basica, Yogyakarta: Andi Offset, 1991.

[4] T. M. Haladyna, Developing and validating multiple-choice test items. British Journal of Educational Technology (3rd ed., Vol. 31), New Jersey: Lawrence Erlbaum Associates Publishers, 2004, https://doi.org/10.4324/9780203825945

[5] R.K. Hambleton, H. Swaminathan, Item response theory, Boston, MA: Kluwer Inc.1985.

[6] R.K. Hambleton, H. Swaminathan, H.J. Rogers, Fundamentals of item response theory, Sage Publications, Inc.1991.

[7] R.K. Hambleton, F. Robin, D. Xing, Item response models for the analysis of educational and psychological test data, In H. E. Tinsley, & S. D. Brown, Handbook of applied multivariate statistics and mathematical modeling (hal. 553-581). San Diego, CA: Academic Press, 2000.

[8] F.M. Lord, Application of item response theory to practical testing problems, Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers, 1980.

[9] R.J. Harvey, A.L. Hammer, Item response theory, The Counseling Psychologist, 27 (3), pp.353-383, 1999

[10] S. Suryabrata, Pengembangan alat ukur psikologis, Yogyakarta: Andi Offset, 1999.

[11] J.C. Nunnally, Introduction to psychological measurement, McGraw-Hill, New York, 1970.

[12] M.W. Klein, Labeling theory and delinquency policy: An experimental test, Criminal Justice and Behavior, 13, pp. 47-79, 1986.

[13] L. Friedenberg, Psychological testing: Design, analysis, and use. USA: Allyn & Bacon, 1995.

[14] B. Sumintono, W. Widhiarso, Aplikasi pemodelan rasch pada assesment pendidikan, 2015

[15] M.J. Allen, W.M. Yen, Introduction to measurement theory, Monterey: Brooks/Cole, 1979.

[16] F.N. Kerlinger, Foundations of behavioral research, 3rd edition, Holt, Rinehart and Winston, New York, 1986.

[17] S. Azwar, Reabilitas dan validitas, Yogyakarta: Pustaka Belajar, 2000

[18] Gronlund, Linn, Meassurement and assesment in teaching, New. Jersey : Prentice Hall, 1995

[19] W.J. Boone, M. Staver, Yale. Rasch analysis in the human sciences, Springer Publishers, 2014.

[20] T.G. Bond, C.M. Fox, Applying the rasch model fundamental measurement in the human sciences (3rd ed.). Mahwah, NJ L. Erlbaum, 2015.