Psychology, Evaluation, and Technology in Educational Research
3 (1), 2020, 26-33

Available Online: http://petier.org/index.php/PETIER

Validity and reliability of cognitive tests study and development of elementary curriculum using Rasch model

Muhammad Erfan [a] *, Mohammad Archi Maulyda [b], Ida Ermiana [c], Vivi Rachmatul Hidayati [d], Arif Widodo [e]

Universitas Mataram. Jalan Majapahit No.62, Kota Mataram, Nusa Tenggara Barat. 83115, Indonesia
[a] muhammaderfan@unram.ac.id; [b] archimaulyda@unram.ac.id; [c] ida_ermiana@unram.ac.id;
[d] vivirachma@unram.ac.id; [e] arifwidodo@unram.ac.id
* Corresponding Author

Received: 3 June 2020; Revised: 18 September 2020; Accepted: 22 September 2020

Abstract: The instrument for measuring knowledge in the subject of Elementary School Curriculum Study and Development has been created to measure students' understanding of teacher candidates in applying concepts, implications and curriculum development at the elementary school level. In order to make reliable and feasible instrument for measuring students knowledge in the subject of Elementary School Curriculum Study and Development, this study aims to produce empirical evidence about the validity and reliability of test instruments using the Rasch model analysis. The study was conducted by testing 20 items on 142 elementary school teacher candidate at one of the State Universities in the City of Mataram, West Nusa Tenggara. The validity and reliability of the instrument were measured by the Rasch analysis model using the Winstep program. The one-dimensional testing of 20 items has a variance measured at 42.7% which exceeds the minimum points of 40.0% desired by the Rasch model. The reliability index of the respondents was 0.65 and the item reliability index was 0.98. All items show a positive value for Point Measure Correlation (PMC) in terms of item polarity which means there is no conflict between the item and the construct being measured. Outfit Mean Square value also shows that all items that almost all items have an MNSQ Outfit value smaller than 1.5 which means the measurement value can be said to be productive except for item 13 (3.77) and item 16 (3.77). Both of these items need to be re-examined because they have problems in measuring their validity. The results of this study have proven that the knowledge measurement test instrument in the Elementary School Curriculum Study and Development course has validity and reliability values that meet and are empirically feasible to be used in measuring Basic School Curriculum Study and Development knowledge for prospective teacher students.

Keywords: validity; reliability; curriculum; analysis; rasch model

How to Cite: Erfan, M., Maulyda, M., Ermiana, I., Hidayati, V., & Widodo, A. (2020). Validity and reliability of cognitive tests study and development of elementary curriculum using Rasch model. *Psychology, Evaluation, and Technology in Educational Research*, *3*(1), 26-33. doi:http://dx.doi.org/10.33292/petier.v3i1.51

INTRODUCTION

Teachers or educators as the spearhead in advancing the education system are required not only to be able to teach and educate but also must be able to evaluate the learning process because the success of the learning process can be seen from the results of evaluations conducted after the learning process activities are carried out (Widyaningsih & Yusuf, 2018). Learning evaluation is an organized and sequential process to determine the achievement of student learning objectives (Magdalena et al., 2020). In simple terms, evaluation is a means of measuring progress and the obstacles faced by students in meeting learning outcomes (Fitriani et al., 2017). Evaluation is very important for the teacher to plan before carrying out learning. This is because evaluation is one of three important components in improving learning (Hasibuan, 2016). With the presence of evaluation, teachers will be able to assess how far the learning programs in an educational unit have been implemented (Sulistyawati & Guntur, 2019). Furthermore, teachers must have superior competence, both in planning, implementing, and evaluating learning. The evaluation itself might also serve as a matter of consideration for defining the appropriate solutions for improvement (Hapsari et al., 2018). The ability to evaluate learning will

http://dx.doi.org/10.33292/petier.v3i1.51

determine how the next lesson plan will be made. Therefore, educators must be able to master a variety of evaluation techniques based on data collected during the learning process which will later be used to measure the extent to which the planned learning objectives have been previously achieved.

Teachers or educators in addition to being able to evaluate must also be able to develop measurement tools or instruments that will be used in evaluating the learning process following the expected learning outcomes. This is because the evaluation process is preceded by the measuring process. The design of a good measuring instrument is directly proportional to the quality of the evaluation to be carried out (Prijowuntato, 2020). The measurement process which is then continued with the assessment process, and ends with evaluation is a series of learning assessments that must be carried out by educators. The development of measuring instruments is an important step in carrying out learning evaluation. The measuring instrument or evaluation instrument developed in addition to being by the characteristics of the various components involved in the learning process must also be able to function as an indicator of the achievement of learning objectives as set out in the lesson plan (RPP).

Measurement of the achievement of learning objectives in evaluation activities is generally carried out with test and non-test techniques. The test is a tool that aims to collect information about the achievement of educational goals or learning objectives (Wahyudi, 2010), besides the test is also a certain way that can be used or procedures that need to be taken in the context of measurement and assessment in the field of education (Kadir, 2015). In simple terms, the test technique presents questions that have wrong or right answers. This allows the teacher to find out the students' understanding of something. This test technique has various models, including written and oral tests. The types of questions on the written test also vary, namely multiple choice, short answers, and descriptions. Another test technique is the oral test, which requires students to answer questions orally. In addition to the test technique there are also non-test techniques which are techniques for assessing student learning outcomes that are carried out without "testing" students, but by systematically observing what is commonly known as observation, interviews, questionnaire distribution, scale document analysis (both attitude scale and rating scale), case studies, and sociometry (Mania, 2008). Apart from observing and distributing questionnaires, another concrete form of non-test technique is inventory. Inventory is an instrument that contains reports on student progress during the learning process (Prijowuntato, 2020). Not only inventory, other types of non-test instruments include project assignments. This type is commonly used by teachers when they want to measure student skills. Project assignments may also be intended to assess processes and outcomes only. In measuring activities, educators are given the freedom to choose measurement techniques and are free to develop assessment instruments in the process of evaluating learning outcomes.

Instruments or measuring instruments that are good in an assessment process must have main characteristics namely, valid, reliable, and have a high level of usefulness (Gronlund et al., 2009). Besides, there are economic and practical aspects that a good instrument must have (Azwar, 2015). Setyosari (2013) argues that the two most important things an instrument must have are valid and reliable. In this case also applies to assessment instruments. An instrument is used to reveal a phenomenon or fact which will later be summarized into data. This is why an instrument must have validity and reality (Arifin, 2017). Good learning outcomes assessment instruments must meet several criteria including assessment instruments have good item validity, items must be steady in the sense of having good item reliability values. The validity and reliability of these items are important because they involve the level of confidence in the outcomes measured in the process of evaluating learning outcomes. The higher the value of the validity and reliability of an instrument, the more accurate the data obtained from a study or measurement (Hayati & Lailatussaadah, 2016). Besides, validity and reliability are also important factors in determining whether a measurement or test that has been carried out has good criteria or not (Wahyuningsih, 2015).

The item or item is said to be valid if the question measures something that must be measured, meaning that if the desired learning outcome is a change in the aspects of knowledge, skills, and attitudes, then the items developed must also include all three of these things (aspects of knowledge, skills, and attitudes) (Sumintono & Widhiarso, 2015). Another example of the validity of a measurement instrument is that when we want to measure length, the measuring instrument we use is a ruler or meter. A ruler or meter can't be used to measure the mass of an object because of course the ruler or meter is invalid and cannot be trusted to measure the mass of an object.

The validity of the instrument can be fulfilled from three aspects or parts. The three aspects are content validity (content), constructs, and criteria (Yusup, 2018). The validity of the content or content provides evidence on each item on the measuring instrument whether it reflects all the achievements to be measured. The validity of the content was assessed by an expert. There are several things that are usually considered in validating content. These include the representation of whether it is in accordance with the expected achievement indicators; the number of questions is appropriate; the format of the answer is clear; scoring is correct and clear; instructions for filling the instrument are visible; time for processing questions; as well as the layout and grammar used.

In contrast to content validity, construct validity focuses on the extent to which an instrument shows measurement results in accordance with the definition of the variable. Construct validity is very important to do. A simple example is, when we measure the length and width of a book, it is enough to use a ruler in cm. When measuring body weight, you can use the scale in kg. These concrete items are easy to measure and the units to be used are clear. It's different when we try to measure creativity, thought processes, mathematical communication skills, or students' critical thinking abilities. These abstract things must be reconstructed into something more concrete and can be measured quantitatively. That is the importance of construct validity in an instrument or measuring instrument.

The validity of the criteria focuses on comparing the instrument in question with other instruments that are considered comparable (Yusup, 2018). The validity of the criteria, as the name implies, focuses on whether the instrument is in accordance with the desired criteria (Arifin, 2017). The validity of this criterion needs to be done with the consideration that each instrument has criteria that are not always ideal. The instruments developed may have high content and construct validity values but are impractical and expensive. This requires the validity of the criteria to compare it with instruments with ideal criteria so that further instrument development can meet the expected criteria. The validity of these criteria is divided into transient and predictive validity.

The reliability of an essential test relates to the test of the constancy of test questions in which it is a set of items if repeatedly given to the same object (Nuswowati et al., 2010). The constancy or consistency of an instrument is if the items in the same instrument are tested several times to the same or almost the same subject or respondent (Rosseni et al., 2009). For example, an exam given today to a student by a teacher should give a value that is not much different if given the next day (because there are no learning activities or forget at the same time in one day) (Sumintono & Widhiarso, 2015). There are two aspects measured in terms of reliability, namely internal consistency and stability. Stable means when measuring two of the same objects, the measuring instrument has a tendency to show the same results. Internal stability is required, as is external stability.

Reliability or validity is a measure of the credibility of a measuring instrument. Reliable measuring instruments are not necessarily valid. Measuring instruments that have been declared valid must be reliable, so they can be used in the evaluation of learning. This is why the first test that must be done in determining the credibility of a measuring instrument is the validity test. The items that were declared valid were then tested for reliability. The aspect of validity that relates to reliability is the validity of the content (content), while the aspect of reliability that relates to validity is the internal reliability between items, item-total and split half.

Analysis of test instruments in the evaluation process in the field of education can be done through two approaches namely the classical test theory (CTT) approach and the modern approach with Rasch modelling. One of the weaknesses of classical test theory is that sometimes there are inconsistencies in the characteristics of items that depend on the ability of the respondent or test takers at the time of working on the questions. This inconsistency can be overcome in measurements by Rasch modelling.

Rasch modeling is a measurement model that measures continuously estimates the validity and reliability of each respondent candidate who answers items/questions and the difficulty of items/ questions for each question/item (Searing, 2008). Analysis by Rasch modelling produces fit statistics analysis which provides information to the researcher whether the data obtained is ideally illustrated that people who have the high ability provide patterns of answers to items according to the level of difficulty (Misbach & Sumintono, 2014). In Rasch modelling, the validity and reliability of a test instrument can be determined by looking at analyses such as item polarity, unidimensional, item-individual/respondent mapping, item-individual reliability, and several other forms of analysis (Bond & Fox, 2007). Therefore, this research was conducted to obtain empirical evidence related to the validity and reliability of items developed to measure the knowledge of prospective teacher students and as an effort to improve the quality of evaluation tools, develop evaluation tools that can measure the ability of prospective elementary school teachers in Elementary studies and curriculum development courses.

METHOD

This research is a quantitative study in the form of a survey conducted on 142 prospective elementary school teacher students as research samples. Students who were respondents were students who received the elementary school curriculum study and development course. The instrument of the questions in this study consisted of 20 items/questions in the form of multiple-choice and aimed at measuring student learning outcomes of prospective teachers from the aspect of knowledge. The objective questions consist of four answer choices (A, B, C, and D) and there is only one correct answer choice from the four options and the whole question is created and given in the form of a quiz on Google Form. The research data was obtained from the answers answered by respondents on the research instrument questions. Item evaluation is based on the correct or wrong answers from each respondent where the correct answer is given a value (1) and if it is wrong then the value is zero (0). The research data were then analyzed using the Winstep program to obtain the results of the analysis of the validity and reliability of items by the Rasch Model.

RESULT AND DISCUSSION

Construct validity

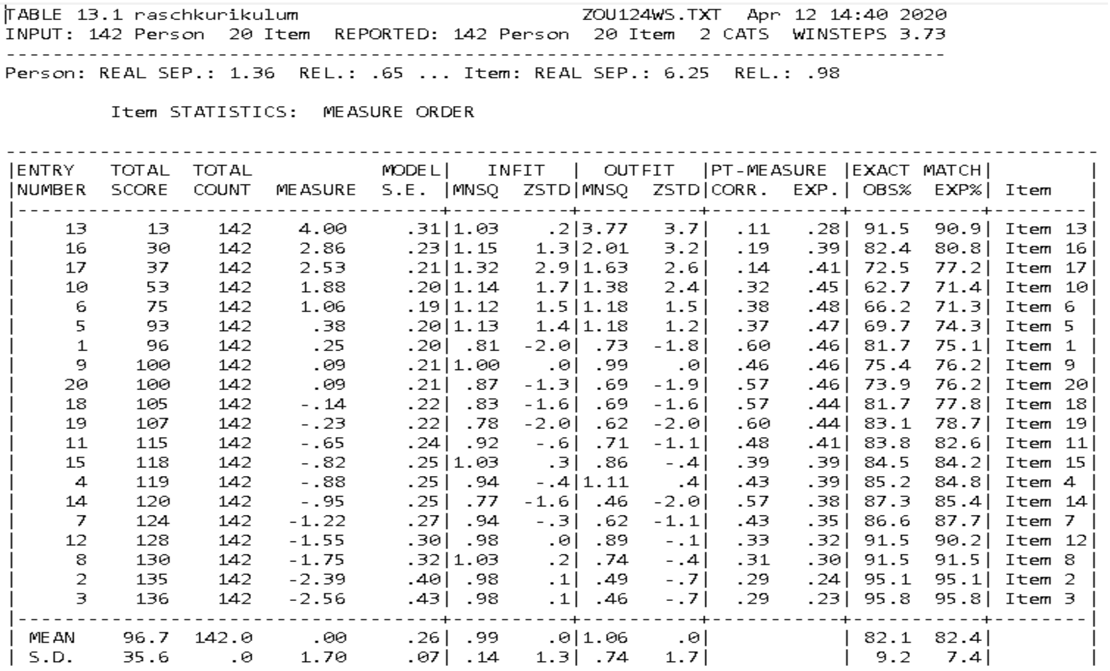
According to Nurfaizin (2019) construct validity is validity that concerns about how far the test items are able to measure what they really want to measure according to a specific concept or a predetermined conceptual definition. Construct validity is commonly used for instruments intended to measure conceptual variables, both typical of performance, such as instruments for measuring attitudes, interests, self-concept, locus control, leadership style, achievement motivation, etc., as well as those with performance characteristics. maximum such as instruments to measure aptitude (aptitude test), intelligence (intellectual intelligence), emotional intelligence and others (Kintner & Sikorskii, 2008).

In order to determine the construct validity of an instrument, a theoretical review process must be carried out of a concept of the variable to be measured, starting from the formulation of the construct, determining dimensions and indicators, to the elaboration and writing of the items of the instrument. The formulation of the construct must be done based on the synthesis of the theories regarding the concept of the variable to be measured through a logical and careful analysis and comparison process. Listening to the theoretical review process as has been stated, the construct validation process of an instrument must be carried out through expert review or justification or through the assessment of a group of panels consisting of people who master the substance or content of the variables to be measured (Ariffin et al., 2010).

The first analysis carried out on the items was an analysis of construct validity which was done by looking at polarity items. As shown by Figure 1 it is known that all item items have a positive Point Measure Correlation (PT-MEASURE CORR) or PMC value. This shows that there is no conflict between item problems with the measured question construction.

Furthermore, if seen from the OUTFIT value in the Mean Square column, it is known that almost all items about the value are smaller than 1.5, there are only three items with Outfit-MNSQ values above 1.5, namely item 13 (3.77), item 16 (2.01), and item 17 (1.63). Item 13 has ZSTD 3.7, item 16 ZSTD 3.2, and item 17 has ZSTD of 2.6. therefore, the researcher decided to review the problems for the three items before all three were aborted from the research instrument.

Subsequent analysis to see the unidimensional construct of knowledge and development of elementary school curriculum, researchers look at the Principal Component Analysis of Rasch Residuals as shown in Figure 2. Unidimential test is one of the tests that need to be seen for the validity of the instrument (Andrich, 1988). From Figure 2 it is known that the total items involved are 20 items and have a measured variance at 42.7%, and all items exceed the minimum points of 40% desired by the Rasch Model.

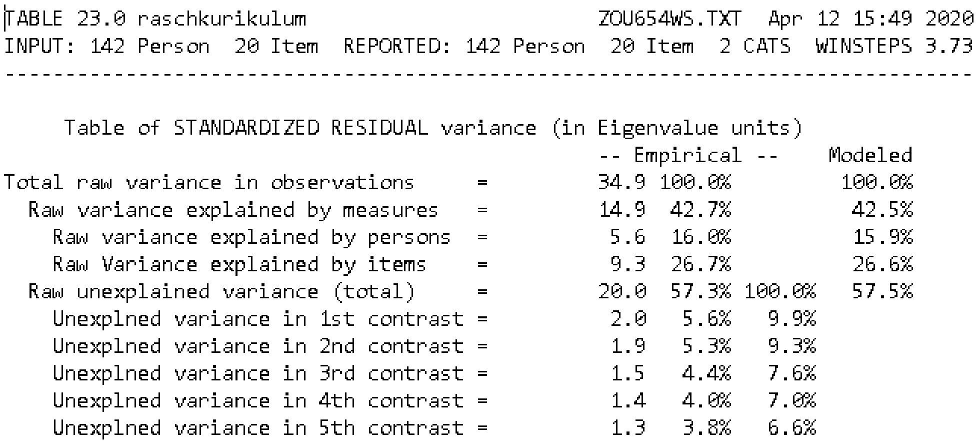


```
TABLE 13.1 raschkurikulum                        ZOU124WS.TXT  Apr 12 14:40 2020
INPUT: 142 Person  20 Item  REPORTED: 142 Person  20 Item  2 CATS  WINSTEPS 3.73
--------------------------------------------------------------------------------
Person: REAL SEP.: 1.36  REL.: .65 ... Item: REAL SEP.: 6.25  REL.: .98

         Item STATISTICS:  MEASURE ORDER

---------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT  MATCH|       |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item  |
|------------------------------------+----------+----------+-----------+-----------+-------|
|    13     13    142    4.00     .31|1.03    .2|3.77   3.7|  .11   .28| 91.5  90.9| Item 13|
|    16     30    142    2.86     .23|1.15   1.3|2.01   3.2|  .19   .39| 82.4  80.8| Item 16|
|    17     37    142    2.53     .21|1.32   2.9|1.63   2.6|  .14   .41| 72.5  77.2| Item 17|
|    10     53    142    1.88     .20|1.14   1.7|1.38   2.4|  .32   .45| 62.7  71.4| Item 10|
|     6     75    142    1.06     .19|1.12   1.5|1.18   1.5|  .38   .48| 66.2  71.3| Item 6 |
|     5     93    142     .38     .20|1.13   1.4|1.18   1.2|  .37   .47| 69.7  74.3| Item 5 |
|     1     96    142     .25     .20| .81  -2.0| .73  -1.8|  .60   .46| 81.7  75.1| Item 1 |
|     9    100    142     .09     .21|1.00    .0| .99    .0|  .46   .46| 75.4  76.2| Item 9 |
|    20    100    142     .09     .21| .87  -1.3| .69  -1.9|  .57   .46| 73.9  76.2| Item 20|
|    18    105    142    -.14     .22| .83  -1.6| .69  -1.6|  .57   .44| 81.7  77.8| Item 18|
|    19    107    142    -.23     .22| .78  -2.0| .62  -2.0|  .60   .44| 83.1  78.7| Item 19|
|    11    115    142    -.65     .24| .92   -.6| .71  -1.1|  .48   .41| 83.8  82.6| Item 11|
|    15    118    142    -.82     .25|1.03    .3| .86   -.4|  .39   .39| 84.5  84.2| Item 15|
|     4    119    142    -.88     .25| .94   -.4|1.11    .4|  .43   .39| 85.2  84.8| Item 4 |
|    14    120    142    -.95     .25| .77  -1.6| .46  -2.0|  .57   .38| 87.3  85.4| Item 14|
|     7    124    142   -1.22     .27| .94   -.3| .62  -1.1|  .43   .35| 86.6  87.7| Item 7 |
|    12    128    142   -1.55     .30| .98    .0| .89   -.1|  .33   .32| 91.5  90.2| Item 12|
|     8    130    142   -1.75     .32|1.03    .2| .74   -.4|  .31   .30| 91.5  91.5| Item 8 |
|     2    135    142   -2.39     .40| .98    .1| .49   -.7|  .29   .24| 95.1  95.1| Item 2 |
|     3    136    142   -2.56     .43| .98    .1| .46   -.7|  .29   .23| 95.8  95.8| Item 3 |
|------------------------------------+----------+----------+-----------+-----------+-------|
| MEAN     96.7  142.0     .00     .26| .99    .0|1.06    .0|           | 82.1  82.4|       |
| S.D.     35.6     .0    1.70     .07| .14   1.3| .74   1.7|           |  9.2   7.4|       |
```

Figure 1. Item polarity



```
TABLE 23.0 raschkurikulum                        ZOU654WS.TXT  Apr 12 15:49 2020
INPUT: 142 Person  20 Item  REPORTED: 142 Person  20 Item  2 CATS  WINSTEPS 3.73
--------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                              -- Empirical --    Modeled
Total raw variance in observations     =      34.9 100.0%         100.0%
  Raw variance explained by measures   =      14.9  42.7%          42.5%
    Raw variance explained by persons  =       5.6  16.0%          15.9%
    Raw Variance explained by items    =       9.3  26.7%          26.6%
  Raw unexplained variance (total)     =      20.0  57.3% 100.0%   57.5%
    Unexplned variance in 1st contrast =       2.0   5.6%   9.9%
    Unexplned variance in 2nd contrast =       1.9   5.3%   9.3%
    Unexplned variance in 3rd contrast =       1.5   4.4%   7.6%
    Unexplned variance in 4th contrast =       1.4   4.0%   7.0%
    Unexplned variance in 5th contrast =       1.3   3.8%   6.6%
```

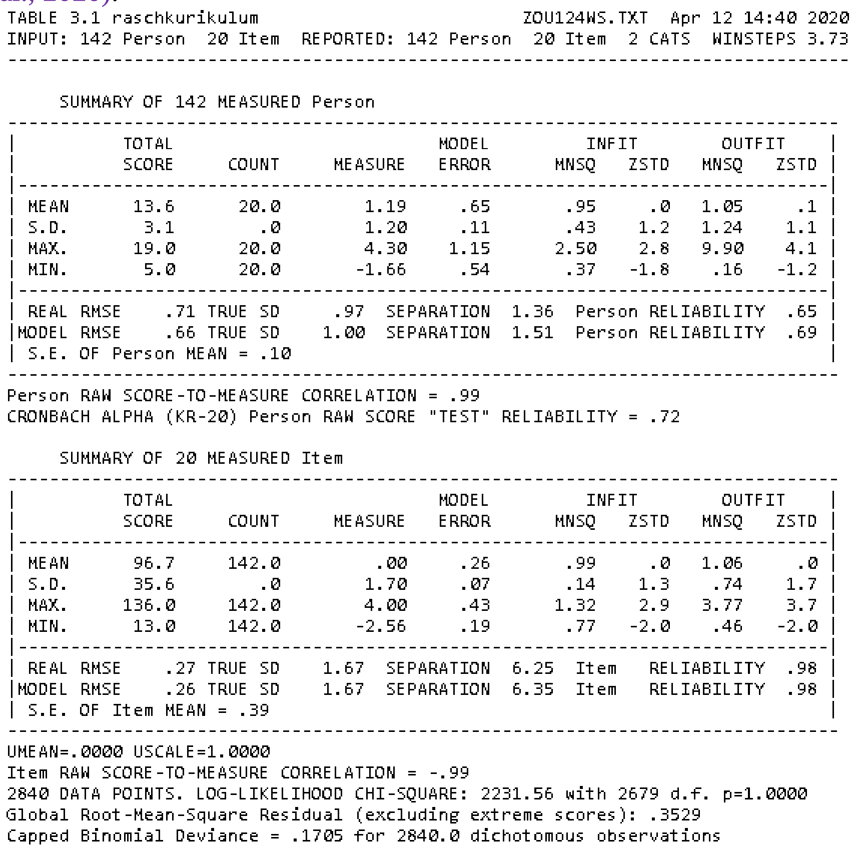
Figure 2. Principal Component Analysis of Rasch Residual

Constructive Reliability

Reliability is the consistency of measurement (Anjos et al., 2016). Fahruna and Fahmi (2017) state that reliability refers to an understanding that the instruments used in research to obtain the information used can be trusted as data collection tools and are able to reveal real information in the field. Azwar (2015) states that reliability is a tool for measuring a questionnaire which is an indicator of variables or constructs. A questionnaire is said to be reliable or reliable if a person's answer to a statement is consistent or stable over time. The reliability of a test refers to the degree of stability, consistency, predictive power, and accuracy. Measurements that have high reliability are measurements that can produce reliable data.

Reliability is an index that shows the extent to which a measuring instrument can be trusted or reliable (Khaeruman & Saefullah, 2017). If a measuring device is used twice - to measure the same symptoms and the measurement results obtained are relatively consistent, then the measuring device is reliable. In other words, reality shows the consistency of a measuring device in the same symptom meter. According to Nuswowati et al. (2010) reliability shows the extent to which measurement results with these tools can be trusted. The measurement results must be reliable in the sense that they must have a level of consistency and stability.

Summary statistics as shown in Figure 3 shows the results of the analysis of items/questions and individual respondents. Item reliability can be seen in the item reliability of 0.98 where the reliability number of this magnitude is included in the category of very good or special (Sumintono & Widhiarso, 2015). Besides, a separation index value of 6.25 was also obtained, in which the separation index could differentiate test items into 8.67 rounded up to 9 (nine) difficulty items. The greater the value of item separation, the quality of the instrument in terms of the overall respondents and the items the better, because it can identify the group of respondents and item groups (Sumintono & Widhiarso, 2015; Erfan et al., 2020).



```
TABLE 3.1 raschkurikulum                          ZOU124WS.TXT  Apr 12 14:40 2020
INPUT: 142 Person  20 Item  REPORTED: 142 Person  20 Item  2 CATS  WINSTEPS 3.73
--------------------------------------------------------------------------------

     SUMMARY OF 142 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE  ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      13.6       20.0        1.19    .65       .95     .0   1.05     .1 |
| S.D.       3.1         .0        1.20    .11       .43    1.2   1.24    1.1 |
| MAX.      19.0       20.0        4.30   1.15      2.50    2.8   9.90    4.1 |
| MIN.       5.0       20.0       -1.66    .54       .37   -1.8    .16   -1.2 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .71 TRUE SD     .97  SEPARATION  1.36  Person RELIABILITY  .65 |
|MODEL RMSE    .66 TRUE SD    1.00  SEPARATION  1.51  Person RELIABILITY  .69 |
| S.E. OF Person MEAN = .10                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .72

     SUMMARY OF 20 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE  ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      96.7      142.0         .00    .26       .99     .0   1.06     .0 |
| S.D.      35.6         .0        1.70    .07       .14    1.3    .74    1.7 |
| MAX.     136.0      142.0        4.00    .43      1.32    2.9   3.77    3.7 |
| MIN.      13.0      142.0       -2.56    .19       .77   -2.0    .46   -2.0 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .27 TRUE SD    1.67  SEPARATION  6.25  Item   RELIABILITY  .98 |
|MODEL RMSE    .26 TRUE SD    1.67  SEPARATION  6.35  Item   RELIABILITY  .98 |
| S.E. OF Item MEAN = .39                                                     |
-------------------------------------------------------------------------------
UMEAN=.0000 USCALE=1.0000
Item RAW SCORE-TO-MEASURE CORRELATION = -.99
2840 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 2231.56 with 2679 d.f. p=1.0000
Global Root-Mean-Square Residual (excluding extreme scores): .3529
Capped Binomial Deviance = .1705 for 2840.0 dichotomous observations
```

Figure 3. Individual Reliability and Item Reliability

The results of the item analysis shown in Figure 3 on the Cronbach Alpha value (KR-20) show a value of 0.72 more than a minimum value of 0.70 (Pallant, 2010). This figure shows that if the items were analyzed using classical test theory, the results of a good or steady reliability analysis were obtained (Erfan et al., 2020).

The high value of item reliability is not accompanied by the high value of Person Reliability. Based on Figure 3 also obtained that the value of Person Reliability is 0.65 which is included in the sufficient category (Sumintono & Widhiarso, 2015). Also, the separation index value of the Measured Person is only 1.36, which if included in the strata equation, a value (H) of 2.14 is rounded to 2. This number 2 indicates that in general all respondents could only be divided into two groups. Conditions, where items are not able to separate individuals or respondents into more than two strata, may be caused by the quality of items/items that are low for good individual separation (Jailani, 2011). However, if seen from the reliability value of the items included in the special category shows that this instrument is sufficient and can be used in measuring the domain of study knowledge and elementary curriculum development for elementary school teacher candidates.

CONCLUSION

The conclusions that can be drawn from this study are based on the results of the construct validity test of 20 items obtained 17 items/test items that there is no conflict between item items and item construction measured by Outfit-MNSQ values that are at less than or equal to 1 .5 In addition to the item reliability test it was found that the item/question reliability index was 0.98 with a special category and the individual/respondent reliability index was 0.65 with a sufficient category. The instrument can be trusted and used to measure the domain of knowledge and elementary curriculum development for elementary school teacher candidates. Based on the results of this study, 17 items were valid and reliable measurement items were obtained to measure students' knowledge about curriculum development in elementary schools and based on the quantity of the items, the lecturer must develop more valid and reliable measurement items using much more specific criteria such as the Rasch Model.

REFERENCES

Andrich, D. (1988). *Rasch models for measurement*. Sage Publication Inc.

Anjos, D. B. M. dos, Rodrigues, R. C. M., Padilha, K. M., Pedrosa, R. B. dos S., & Gallani, M. C. B. J. (2016). Reliability and construct validity of the instrument to measure the impact of valve heart disease on the patient's daily life. *Revista Latino-Americana de Enfermagem*, *24*(6), 12–25. https://doi.org/10.1590/1518-8345.0624.2730

Ariffin, S. R., Omar, B., Isa, A., & Sharif, S. (2010). Validity and reliability multiple intelligent item using rasch measurement model. *Procedia - Social and Behavioral Sciences*, *9*, 729–733. https://doi.org/10.1016/j.sbspro.2010.12.225

Arifin, Z. (2017). Kriteria instrumen dalam suatu penelitian. *Jurnal Theorems (the Original Research of Mathematics)*, *2*(1), 28–36. https://doi.org/10.31949/th.v2i1.571

Azwar, S. (2015). *Reliabilitas dan validitas*. Pustaka Pelajar.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis kualitas soal kemampuan membedakan rangkaian seri dan paralel melalui teori tes klasik dan model rasch. *Indonesian Journal Of Educational Research and Review*, *3*(1), 11. https://doi.org/10.23887/ijerr.v3i1.24080

Fahruna, Y., & Fahmi, M. (2017). Validitas dan reliabilitas konstruk pengukuran perpustakaan ideal berbasis pemakai dengan pendekatan LIBQUAL. *Jurnal Ekonomi Bisnis Dan Kewirausahaan*, *6*(2), 161. https://doi.org/10.26418/jebik.v6i2.22989

Fitriani, C., AR, M., & Usman, N. (2017). Kompetensi profesional guru dalam pengelolaan pembelajaran di MTs Muhammadiyah Banda Aceh. *Jurnal Administrasi Pendidikan : Program Pascasarjana Unsyiah*, *5*(2), 88–95. http://e-repository.unsyiah.ac.id/JAP/article/view/8246

Gronlund, N. E., Linn, R. L., & Miller, M. D. (2009). *Measurement and evaluation in teaching* (10th ed.). Macmillan Publishing Co., Inc.

Hapsari, S. I., Sugiyarto, K. H., & Kosaka, N. (2018). An evaluation of application of information technology and communication of learning science with the theme of solar system. *Psychology, Evaluation, and Technology in Educational Research*, *1*(1), 41. https://doi.org/10.33292/petier.v1i1.18

Hasibuan, H. (2016). Studi kompetensi guru Pendidikan Agama Islam dalam pelaksanaan evaluasi pembelajaran. *Forum Paedagogik*, *08*(02), 14–38. https://doi.org/10.24952/paedagogik.v8i2.571

Hayati, S., & Lailatussaadah, L. (2016). Validitas dan reliabilitas instrumen pengetahuan pembelajaran aktif, kreatif dan menyenangkan (pakem) menggunakan model rasch. *Jurnal Ilmiah Didaktika*, *16*(2), 169. https://doi.org/10.22373/jid.v16i2.593

Jailani, M. K. M. (2011). *Manual pengenalan pengukuran rasch & winsteps*. Fakulti Pendidikan Universiti Kebangsaan Malaysia.

Kadir, A. (2015). Menyusun dan menganalisis tes hasil belajar. *AL-TA'DIB : Jurnal Kajian Ilmu Kependidikan*, *8*(2), 70–81. https://doi.org/10.31332/atdb.v8i2.411

Khaeruman, K., & Saefullah, E. (2017). Analisis lokasi usaha terhadap penjualan pedagang buah-buahan di sepanjang jalan ciptayasa serang. *Sains Manajemen*, *3*(2), 15–37. https://doi.org/10.30656/sm.v3i2.255

Kintner, E. K., & Sikorskii, A. (2008). Reliability and construct validity of the Participation in Life Activities Scale for children and adolescents with asthma: an instrument evaluation study. *Health and Quality of Life Outcomes*, *6*(1), 43. https://doi.org/10.1186/1477-7525-6-43

Magdalena, I., Crismaningrum, O. D., Chairunnisa, N., & Jannah, N. (2020). Evaluasi pembelajaran dalam keterampilan berbicarapada mata pelajaran bahasa Indonesia siswa kelas IV SDN Balaraja i. *Jurnal Halaqah*, *2*(3), 349–356. https://doi.org/10.5281/zenodo.3940559

Misbach, I. H., & Sumintono, B. (2014). Pengembangan dan validasi instrumen "persepsi siswa tehadap karakter moral guru" di Indonesia dengan model rasch. *PROCEEDING Seminar Nasional Psikometri*, 148–162.

Nurfaizin, N. (2019). Uji validitas konstruk self control terhadap prokrastinasi akademik dengan metode confirmatory factor analysis (CFA). *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia (JP3I)*, *7*(1), 41–48. https://doi.org/10.15408/jp3i.v7i1.12107

Nuswowati, M., Binadja, A., Soeprodjo, S., & Ifada, K. E. N. (2010). Pengaruh validitas dan reliabilitas butir soal ulangan akhir semester bidang studi kimia terhadap pencapaian kompetensi. *Jurnal Inovasi Pendidikan Kimia*, *4*(1), 566–573. https://journal.unnes.ac.id/nju/index.php/JIPK/article/view/1314

Pallant, J. (2010). *SPSS survival manual a step by step guide to data analysis using the SPSS Program*. McGraw-Hill Education.

Prijowuntato, S. W. (2020). *Evaluasi pembelajaran*. Sanata Dharma University Press.

Rosseni, D., Ahmad, M., M.Faisal, K., Sidek, N. M., Karim, A. A., Johar, N. A., Jusoff, K., Zakarian, M. S., Mastor, K. A., & Ariffin, S. R. (2009). Kesahan dan kebolehpercayaan soal selidik gaya e-pembelajaran (eLSE) versi 8.1 menggunakan model pengukuran rasch. *Journal of Quality Measurement and Analysis*, *5*(1), 15–27. http://journalarticle.ukm.my/1930/

Searing, L. M. (2008). *Family functioning scale validation: A Rasch analysis*. University of Illinois.

Setyosari, H. P. (2013). *Metode penelitian pendidikan & pengembangan*. Prenadamedia Group.

Sulistyawati, S., & Guntur, G. (2019). Sports education learning program evaluation in senior high school. *Psychology, Evaluation, and Technology in Educational Research*, *2*(1), 22. https://doi.org/10.33292/petier.v2i1.31

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Trim Komunikata.

Wahyudi, W. (2010). Assesment Pembelajaran Berbasis Portofolio di Sekolah. *Jurnal Visi Ilmu Pendidikan*, *2*(1), 288–296. https://doi.org/10.26418/jvip.v2i1.370

Wahyuningsih, E. T. (2015). *Analisis butir soal tes objektif buatan guru ulangan semester ganjil mata pelajaran ekonomi kelas X di SMA Negeri 1 Mlati tahun ajaran 2013/2014*. Universitas Negeri Yogyakarta.

Widyaningsih, S. W., & Yusuf, I. (2018). Analisis soal modul laboratorium fisika sekolah I menggunakan racsh model. *Gravity : Jurnal Ilmiah Penelitian Dan Pembelajaran Fisika*, *4*(1). https://doi.org/10.30870/gravity.v4i1.3116

Yusup, F. (2018). Uji validitas dan reliabilitas instrumen penelitian kuantitatif. *Jurnal Tarbiyah : Jurnal Ilmiah Kependidikan*, *7*(1). https://doi.org/10.18592/tarbiyah.v7i1.2100