Item analysis of technologial pedagogical content knowledge (TPACK) in pre-service chemistry teachers using the Rasch Model application

Wahyuni Tri Lestari, Sulistyop Saputro, Mohammad Masykuri, Budi Hastuti, Maria Ulfa, Sri Mulyani, Sri Yamtinah

Master Program of Chemistry Education, Sebelas Maret University

Corresponding author: wahyuni.teel@gmail.com

Abstract. This research aims to analyze TPACK items using the Rasch model in terms of validity, reliability, item difficulty and bias items. The method used is descriptive quantitative with 34 pre-service chemistry teachers in semester VI consisting of 26 female students and 8 students from Sebelas Maret University (UNS). The instrument used in this study was the matter of Teachnological Pedagogical Content Knowledge (TPACK) which was adapted from the PPG Teachnological Pedagogical Content Knowledge (TPACK), which consisted of 30 multiple choice questions. The results showed that: 1) the validity of the suitability level of items was 27 questions that were fit and 3 questions that were not fit. 2) the results of raw variance data of 33.5% indicate that the requirements for Raw variance explained by measures of at least 20% can be met and the variance that cannot be explained by the instrument of 8.2% also meet the criteria ie, not to exceed 15%. 3) Reliability of pre-service chemistry teachers is 0.52 in the weak category, item reliability is 0.90 in the good category and reliability between pre-service chemistry teachers and items 0.64 in the sufficient category. 4) The level of difficulty of the questions is very difficult, difficult, easy, very easy and is dominated by questions that are categorized as difficult. 5) there are no questions that have a probability of less than 5% so there are no biased items. In conclusion, TPACK item analysis using the Rasch model is valid, reliable, item difficulty is very good and there are no biased items. So the questions that have been analyzed can be used to measure TPACK on pre-service chemistry teachers.

1. Introduction
One of the competencies of teachers as educators is that they must have the ability to carry out evaluation of learning. Evaluation or assessment is a process for making decisions using test and non-test instruments. UU no. 14/2005 states that "the main task of teachers as professional educators is to educate, teach, guide, train, assess and evaluate students in early childhood education, formal education, basic education and secondary education [1]. Government Regulation No. 19 of 2005 concerning National Education Standards, evaluation of learning outcomes is carried out by evaluating learning outcomes which aim to monitor the learning process, progress and improvement of learning outcomes and to assess and measure the level of competency achievement [2].

Evaluation is one of the important components and stages that must be taken to determine the effectiveness of learning [3]. The quality of the question instrument largely determines the outcome of learning in the form of mastery of the learning objectives formulated [4]. The item analysis process is
an activity carried out to examine and examine each item through the collection of information from students’ answers in order to obtain quality questions before the questions are used [5]. Item analysis can be done qualitatively, which is related to the content and form, and quantitatively, which is related to the statistical characteristics. This quantitative analysis includes the analysis of validity, reliability, level of difficulty, distinguishing features of the problem and the level of deception held by the question [6].

A valid item means that the instrument can be used to measure what it is intended to measure. While a reliable instrument is an instrument that if tested several times on the same object at different times will produce the same data [7]. Difficulty level of questions related to the ability to be able to answer or be able to solve the problem properly and correctly. The purpose of this study is to analyze the TPACK items using the Rasch model to test the validity, reliability, and person and items simultaneously. Rasch modeling can be used to measure validity and reliability directly based on the principle of probability [8].

In classical analysis testing raw scores in the form of rating (rating scale) are directly analyzed and treated as data that appear to have integer characters. However, in the Rasch Model, raw data cannot be directly analyzed, but must first be converted into an 'odds ratio' to then transform logarithms into logit units as a manifestation of the probability of respondents responding to an item [9-10] Through the Rasch model, a response which are ordinal in nature can be transformed into ratios that have a higher degree of accuracy by referring to the principle of probability. The difference between the Rasch model and the classical testing is that in the analysis of the data with the Rasch model, the data adjust the model, whereas in the classical testing the model is chosen based on the data [11]. Based on this, the use of the Rasch model in validating this instrument will produce more holistic information about the instrument and better meet the definition of measurement.

The advantages of the Rasch model include: being able to predict missing data, which is based on a systematic response pattern; able to produce a standard error measurement value for the instrument used which can improve the accuracy of the calculation; and calibration which is done simultaneously in three ways, namely the measurement scale, respondents, and items [9-13]. The advantages of the Rasch model are very suitable for use in this study because they evaluate the ability of items. The use of the Rasch model is more effective than the classical analysis [14-16].

In this study the TPACK question instruments used in prospective pre-service chemistry teachers are the questions adapted from the TPACK PPG questions. Based on the explanation above, the problem that will be examined in this study is how the TPACK item analysis using the Rasch model in terms of validity, reliability, item difficulty level and item index refractive index.

2. Methodology
This research is a quantitative descriptive study to look for information and data that can be used to describe the quality of the matter of TPACK. The study was conducted on 34 pre-service chemistry teachers in semester VI consisting of 26 female students and 8 students from Sebelas Maret University (UNS). The instrument used in this study was the matter of TPACK which was adapted from the PPG TPACK, which consisted of 30 multiple choice questions. The data studied were analyzed using the Rasch model using the Winstep® version 3.73 application [17].

Validity can be seen from the construct validity and content validity. The results of construct validity analysis can be seen in the Winsteps program by using output tables 23. Item: Dimensionality. The construct validity can be determined by looking at Raw Variance and Unexplained variance. The minimum requirement for Raw variance explained by measures is 20% and the variance that cannot be explained by the instrument should ideally not exceed 15% [9-10].

The results of the analysis of content validity can be seen in the Winsteps program by using output tables 10, 13, and 30. Content Validity is determined through the item's Level of Conformity (Table 10. Item Fit Order) criteria used are:

a. Outfit mean square value (MNSQ) has a good condition that is 0.5 <MNSQ <1.5.

b. The Z-standard Output Value (ZSTD) received is -2 <ZSTD <+2.
c. The accepted Point Measure Correlation (Pt Mean Corr) value is 0.4 < Pt Mean Corr < 0.85.

The results of the reliability analysis can be seen in the Winsteps program by using the output table 1. Variable Map. Reliability can be determined through the separation value, reliability and the Cronbach α value. Separation is good if there are 4 groups of questions. To find separation (grouping indication) through calculation using the following equation:

\[ H = \frac{[(4 \times \text{SEPARATION}) + 1]}{3} \]  

The pre-service chemistry teacher reliability scores and items were determined using the following table:

| Table 1. Reliability Value |
|-----------------------------|
| Number | Range | Category |
| 1.     | <0,67 | Weak     |
| 2.     | 0,67-0,80 | Enough  |
| 3.     | 0,80-0,90 | good    |
| 4.     | 0,90-0,94 | Very good |
| 5.     | >0,94 | Special  |

To measure reliability, namely the interaction between pre-service chemistry teachers and the whole items used the Cronbach α value with the following conditions:

| Table 2. Cronbach's α values [9-10] |
|-------------------------------------|
| Number | Range | Category |
| 1.     | <0,5  | Bad      |
| 2.     | 0,5-0,6 | less good |
| 3.     | 0,6-0,7 | Enough  |
| 4.     | 0,7-0,8 | good    |
| 5.     | >0,8  | Very good |

The level of difficulty of the item (item measure) provides a detailed view of the logit value of each item. The output from this table provides information about items sorted from those that have the highest logit measure value to the lowest logit value, which indicates that the items are sorted from the difficult items to the easiest items. Classification of the level of difficulty of the questions can be determined based on the average value of the logit added by the value of the standard deviation (SD). This value is useful for identifying item groups (separation). Bias analysis of items (Table 30. Item DIF) states that an item can be considered biased if the probability value is <5% (<0.05) [9-10].

3. Result and Discussion

General description of the analyzed items can be seen in Figure 1. The picture illustrates the distribution of the abilities of 34 pre-service chemistry teachers and the distribution of the level of difficulty on the same scale. From the picture it can be seen that there is a pre-service chemistry teachers who has high ability, namely pre-service chemistry teachers with code 01L. The pre-service chemistry teachers is also outside the limits of the two Standard Deviations (SD) which are marked with T. This shows that this pre-service chemistry teachers has a high intelligence that is different (outlier) from the group tested. The pre-service chemistry teachers who has the lowest ability is code 07P with a logit value of almost -1 logit, but is still in the area between 1 elementary school and T.
Figure 1 also shows that question number 25 or S25 has the highest level of difficulty shown by the logit +4 logit value. Problems S23 and S5 have the lowest difficulty level which has a logit value below -3 logit and the possibility of all pre-service chemistry teachers working on this problem correctly.

![Diagram of Problem Item Analysis](image)

**Figure 1. Overview of Problem Items Analyzed**

From these results it can be seen that there are three outlier problems namely S5, S23, and S25 which are above 2SD or T. There are questions that have the same difficulty level, one of which is S4, S9, S12, S19, and S20 which has a logit value below +1 logit. Figure 1 also shows that the average logit ability of pre-service chemistry teachers is above 0.0 logit. This shows that the average achievement of sixth semester chemistry education pre-service chemistry teachers is above the average level of difficulty of standard questions.
3.1. Validity
Analysis of the validity generated by the Winstep® program is an analysis of construct validity and content validity.

3.1.1. Construct validity
The results of the analysis conducted with the Rasch model obtained the results in the output items dimension dimensionality, the results of the analysis can be seen in Figure 2.

![Figure 2. Shows the results of construct validity on the value of Raw variance explained by](image)

Measures empirically obtained 33.5% and the Rasch model predicts 33.9% in this case the construct validation empirically is almost the same as the value predicted by the Rasch model. The minimum requirement for Raw variance explained by measures is 20% and the variance that cannot be explained by the instrument should ideally not exceed 15% [9-10].

Based on these explanations, the results of the raw variance data of 33.5% indicate that the requirements for Raw variance explained by measures of at least 20% can be met and the variance that cannot be explained by the instrument of 8.2% also meet the criteria ie, not exceeding 15%. So it can be concluded that the TPACK question is construct valid.

3.1.2. Content Validity
Content validity analysis includes the level of appropriateness of the questions that serves to see the quality of the level of conformity of the items with the model. The information provided in the form of information about the suitability of the items with the criteria is to see the value of the mean square outfit, z-standard outfit, and point measure correlation. The item is said to be valid or accepted if it has met at least 2 criteria and corrected if it meets one of the three criteria, and is discarded if no one meets the criteria. Grain suitability value is strongly influenced by the amount of data, the larger the sample used, the better the level of suitability.

The results of the analysis of the content validity that can be seen the level of appropriateness of the items are shown in the output items item fit as in Table 3.

| Number | Question | Outfit MNSQ | Outfit ZSTD | Explanation |
|--------|----------|-------------|-------------|-------------|
| 1      | S25      | 1.65        | 0.9         |              |
| 2      | S5       | 1.38        | 3.1         | Item Question Unfit |
| 3      | S23      | 1.25        | 2.2         |              |
| 4      | S22      | 1.17        | 0.5         | Item Question Fit |
| 5      | S6       | 1.14        | 0.7         |              |
| 6      | S29      | 1.11        | 0.7         |              |
| 7      | S2       | 1.06        | 0.3         |              |
| 8      | S16      | 1.05        | 0.3         |              |
Table 3 shows the results of the validity analysis of the suitability level of the items, from the 30 TPACK items obtained as many as 27 items that were declared fit and 3 items that were not fit because they did not meet the MNSQ Outfit and ZSTD Outfit criteria. In questions number 25, 5, and 23 the outfit value of ZSTD does not fit the accepted condition, which is between $-2 < \text{ZSTD} < +2$. Whereas in question number 25 the MNSQ outfit value cannot be accepted because it is worth more than 0.5. So that the three questions that are not fit need to be corrected. For the other 27 questions, they can be declared fit because they meet the MSQ outfit and ZSTD outfit criteria. A question that has been declared fit means fulfilling the criteria and can guarantee that the level of understanding of the prospective chemistry teacher is indeed tested through appropriate and quality items. Rasch modeling is very effective because it can be obtained the results of a validity analysis that can be trusted [19]. The Rasch model is also easier to use because it has been accompanied by direct analysis using a computer application [20].

3.2. Reliability

The reliability value in the Rasch modeling is indicated by the value of individual separation (person separation) and item separation. Individual separations indicate how well a set of items in the test spreads along the logit range or continuum. The item separation shows how large the sample is subject to such measurements along a linear interval scale [9-10]. The results of the reliability analysis of the items shown in the summary statistics can be seen in Table 4.

| Number | Question | $Outfit$ MNSQ | $Outfit$ ZSTD | Explanation |
|--------|----------|---------------|---------------|-------------|
| 9      | S4       | 1.04          | 0.4           |             |
| 10     | S24      | 1.04          | 0.3           |             |
| 11     | S10      | 1.00          | 0.1           |             |
| 12     | S3       | 0.99          | 0.0           |             |
| 13     | S15      | 0.91          | 0.0           |             |
| 14     | S8       | 0.97          | 0.0           |             |
| 15     | S17      | 0.99          | -0.1          |             |
| 16     | S11      | 0.92          | -0.4          |             |
| 17     | S9       | 0.98          | -0.1          |             |
| 18     | S30      | 0.64          | -0.3          |             |
| 19     | S13      | 0.88          | -0.4          |             |
| 20     | S26      | 0.76          | -0.2          |             |
| 21     | S18      | 0.93          | 0.0           |             |
| 22     | S7       | 0.94          | -0.5          |             |
| 23     | S14      | 0.91          | -0.7          |             |
| 24     | S1       | 0.90          | -0.5          |             |
| 25     | S28      | 0.61          | -0.5          |             |
| 26     | S21      | 0.48          | -0.6          |             |
| 27     | S19      | 0.54          | -0.6          |             |
| 28     | 12       | 0.68          | -0.8          |             |
| 29     | S20      | 0.85          | -1.4          |             |
| 30     | S27      | 0.77          | -2.0          |             |

| Variable                        | Logit Average (SD) | Separation | Reliability | $\alpha$ Crombach |
|---------------------------------|--------------------|------------|-------------|-------------------|
| pre-service chemistry teacher   | 0.38               | 0.7        | 0.52        | 0.64              |
| question items                  | 1.56               | 3.05       | 0.90        |                   |
Table 4 shows the results of the pre-service chemistry teachers reliability value of 0.52 which means that the pre-service chemistry teachers reliability is weak because the value is <0.67. This shows that the subjects giving answers have not been consistent, but it can also be caused by the number of respondents used in this study is too small, thus affecting the value of reliability of the person [9-10]. While the reliability value of the questions obtained a value of 0.90 which means that the reliability of the questions is also good which is in the range of 0.80-0.90. This means that the item questions are consistent, which means they can provide the same information if tested in different times.

Cronbach's alpha value is 0.64 which means that the reliability value between pre-service chemistry teachers and items is categorized as sufficient. The assessment is based on Cronbach’s alpha value which can measure the interaction between the person and the whole item. This shows that the interaction between person and question items as a whole is good enough. The value of separating the grouping of pre-service chemistry teachers and items can be determined as in equation (1).

\[ H = \left( \frac{4\times SEPARATION + 1}{4} \right) ^ \frac{3}{3} \]

\[ H \text{ pre-service chemistry teacher} = \left( \frac{4\times 0.77 + 1}{4} \right) ^ \frac{3}{3} = 1.36 \]

\[ H \text{ question items} = \left( \frac{4\times 3.05 + 1}{4} \right) ^ \frac{3}{3} = 4.4 \]

The pre-service chemistry teacher H scores 1.36 rounded 1, which means that the pre-service chemistry teacher separation is not good because the cognitive abilities of pre-service chemistry teachers are only dominated by one ability. This can be caused by the number of respondents used in this study is too small. Separation of pre-service chemistry teachers is said to be good if the cognitive abilities of pre-service chemistry teachers are divided into high, medium, and low abilities. H items are as many as 4.4 rounded up 4, which means the separation of questions is of good value because dividing the questions into 4 groups is very easy, easy, difficult, very difficult.

### 3.3. Difficulty level items

Based on the results of data analysis, the results of the level of difficulty of the questions on the output items measure. The level of difficulty of items can be grouped by combining logit mean values and Standard Deviation (SD) values. This value is useful for identifying item groups (separation). From these results obtained an average logit measure value of 0.00 and an SD value of 1.96.

**Table 5. The results of the item measure analysis of item difficulty analysis**

| Item Statistics | Measure |
|-----------------|---------|
| Mean            | 0.00    |
| S.D.            | 1.96    |

**Table 6. Description of the results of the analysis of the difficulty level of the questions**

| Category       | Item Number Question | amount | Percentage (%) |
|----------------|----------------------|--------|----------------|
| Outlier        | 25                   | 1      | 3.33           |
| very difficult | 22,18                | 2      | 6.66           |
| difficult      | 2,16,3,29,14,7,4,9,12,19,20,17,27 | 13     | 43.33          |
| easy           | 10,1,11,6,13,8,15    | 7      | 23.33          |
| very easy      | 24,26,28,21,30       | 5      | 16.66          |
| Outlier        | 5,23                 | 2      | 6.66           |

Based on the results of the analysis, the questions already have difficulty levels that are very difficult, difficult, easy, and very easy. The level of difficulty of the questions is dominated by questions that are categorized as difficult. While the outlier category means that the item cannot be
used or thrown away. Rasch modeling can provide detailed information about the difficulty level of a given instrument [21]. Rasch modeling can be obtained by a good question instrument to measure the level of understanding of respondents in research so that it is highly recommended its use especially in the field of education [22].

3.4. Analysis of Bias Item Problem
A problem can be said to be biased if it is obtained that one pre-service chemistry teachers with certain characteristics is more advantageous than pre-service chemistry teachers with other characteristics. TPACK questions are given to pre-service chemistry teachers by looking at the sexes, namely male (L) and female (P). This gender variable can be used to detect item bias.

An item can be considered biased if the probability value of the item is below 5% (0.05) [9-10]. Table 7 shows that there are no problems that have a probability of less than 5%.

Table 7. Bias Item Questions

| Item Number | Name |
|-------------|------|
|             |      |

If seen from the sex, the problem situation can be seen in Figure 3. Figure 3 shows that S25 has a high level of difficulty and S4 and S22 have a low level of difficulty for female and male pre-service chemistry teachers.
An educator must master the technique of drafting the questions so that the questions produced are in the form of quality questions. This can be accomplished by following the training or by reading the book preparation of questions so that it can help the teacher's task in analyzing the item [23]. If the educator wants to use the questions in the books, then an analysis should be done first to ensure the quality of the questions. Item analysis aims to determine the role of each item related to the whole test [24].

4. Conclusion
Based on the results of the analysis of items about the quality of items can be concluded that: the validity of the suitability level of items as many as 27 questions that are fit and 3 questions that are not fit. From the results of raw variance data of 33.5% shows that the requirements for Raw variance explained by measures of at least 20% can be met and the variance that cannot be explained by the instrument of 8.2% also meet the criteria ie, not to exceed 15%. So it can be concluded that the questions are construct valid. The reliability of pre-service chemistry teachers 0.52 means weak, the reliability of item 0.90 means good and the reliability between pre-service chemistry teachers with item 0.64 means sufficient. Questions have had 4 levels of difficulty, namely very difficult questions on questions S22 and S18. Questions with difficult categories are questions S2, S16, S29, S14, S7, S4, S9, S12, S19, S20, S17, and S27. Questions that have easy categories are in questions S10, S1, S11, S6, S13, S8, and S15, while questions with categories are very easy in questions S24, S26, S28, S21, S30, S30. This problem is dominated by questions with difficult categories. From these questions there are no questions that have a probability of less than 5% so there are no biased items. In conclusion, TPACK item analysis using the Rasch model is valid, reliable, item difficulty is very good and there
are no biased items. So the questions that have been analyzed can be used to measure TPACK on pre-service chemistry teachers.

Acknowledgments
The author would like to thank the Institute for Research and Community Service (LPPM) Sebelas Maret University Surakarta for providing grant funds in this research activity.

References
[1] Arikunto, S 2012 Dasar-dasar Evaluasi Pendidikan (Jakarta: Bumi Aksara)
[2] BNSP 2009 Peraturan Pemerintah Republik Indonesia Nomor 19 Tahun 2005 Tentang Standar Nasional Pendidikan (Jakarta: BNSP)
[3] Zainal Arifin 2012 Evaluasi Pembelajaran (Bandung: PT Remaja Rosdakarya)
[4] Yusuf, I., & Widyaningsih, S. W 2018 Profil Kemampuan Mahasiswa dalam Menyelesaikan Soal HOTS di Jurusan Pendidikan Fisika Universitas Papua, Jurnal Komunikasi Pendidikan, 2(1), 42-49
[5] Ratumanan, T. G., & Laurens, T 2011 Penilaian Hasil Belajar pada Tingkat Satuan Pendidikan (Ambon: Unesa University Press)
[6] Rahayu, Dwi T., Purnomo H. B., & Sukidin 2014 Analisis Tingkat Kesukaran dan Daya Beda pada Soal Ujian Tengah Semester Ganjil Bentuk Pilihan Ganda Mata Pelajaran Ekonomi Kelas X di SMA Negeri 5 Jember Tahun Ajaran 2012-2013 Jurnal Edukasi UNEJ, 1(1)
[7] Sugiyono 2014 Metode Penelitian Kombinasi (Bandung: AlfaBeta)
[8] Tennant, A., McKenna, S. P., & Hagell, P 2004 Application of Rasch Analysis in the Development and Application of Quality of Life Instruments. Value in Health, 7, S22-S26
[9] Sumintono, B., and Widhiarso, W 2014 Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial (Cimahi: Trim Komunikata)
[10] Sumintono, B., & Widhiarso, W 2015 Aplikasi Pemodelan Rasch pada Assessment Pendidikan (Bandung: Trim Komunikata)
[11] Bond, T.G., & Fox, C 2014 Applying the Rasch Model. Fundamental measurement in the Human Sciences. Lawrence Erlbaum Associates, Publishers Mahwah New Jersey
[12] Boone, W. J., Straver, J. R., and Yale M. S 2014 Rasch Analysis in the Human Sciences (Dordrecht: Springer)
[13] Bond, T. G., and Fox C. M 2015 Applying the Rasch Model, Fundamentals Measurement in the Human Sciences (3rd edition) (New York: Routledge)
[14] Fisher 1993 A. G. The Assessment of IADL Motor Skills: An Application of Manyfaceted Rasch Analysis American Journal of Occupational Therapy, 47(4), 319-329
[15] Fisher 2007 W. P. Jr “Rating Scale Instrument Quality Criteria,” Rasch Meas. Trans., vol. 21, no. 1, p. 1095
[16] Dong, N. T., and Seong, D. N. F 2014 “Applying the Rasch Model to Investigate Singapore Principals ’ Instructional Leadership Practices,” vol. 20, no. 2, pp. 1–26
[17] Linacre, J. M 2013 FACETS Version 3.71.3 [Computer Software and Manual] (Chicago: Chicago: Winsteps.com)
[18] Andersen, E. B 1973 A Goodness of Fit Test for the Rasch Model. Psychometrika, 38(1), 123-140
[19] Sari, D. R, Sekarwana, N., Hinduan, Z. R., & Sumintono, B 2016 Analisis Tingkat Kepuasan Masyarakat terhadap Dimensi Kualitas Pelayanan Tenaga Pelaksana Eliminasi Menggunakan Pemodelan Rasch. JSK, 2(1), 47-55
[20] Mair, P., & Hatzinger, R 2007 Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. Journal of Statistical Software, 20(9), 1-20
[21] Curtis, D. D 2007 X-ray Your Data with Rasch. International Education Journal, 8(2), 249-259
[22] Ismail, I., Idrus, R. M., & Johari, S. S. M 2010 Acceptance on Mobile Learning Via SMS: a Rasch Model Analysis International Journal of Interactive Mobile Technologies (iJIM),
4(2), 10-16

[23] Purwanti, M 2014 Analisis Butir Soal Ujian Akhir Mata Pelajaran Akuntansi Keuangan Menggunakan Microsoft Office Excel 2010 *Jurnal Pendidikan Akutansi Indonesia*, 7 (1), 81-94

[24] Damopolii, I 2016 Analisis Butir Soal Ujian Tengah Semester Mata Kuliah Biologi Umum Di Universitas Papua. *Prosiding Seminar Nasional*, 2 (1), 145-153