Development of Test Instruments to Test Students' Understanding of Macroscopic, Sub-Microscopic and Symbolic Levels in Acid-Base Titration Material Using the Rasch Model

Febi Edinda Putri*, Faizah Qurrata Aini

Chemistry Education Study Program FMIPA UNP, Padang, 25131, Indonesia

ARTICLE INFO

A B S T R A C T

This study created a test instrument that can test students' understanding of the macroscopic, sub-microscopic and symbolic levels in acid-base titration material that has been tested for validity, reliability, difficulty index and differential power of the questions . This development research uses the Rasch model. The subjects of this study were four lecturers, two high school chemistry teachers and 35 students of SMA N 2 Padang. The object is the quality of the test instrument that meets the criteria of validity, reliability, difficulty index and discriminatory power . Data analysis used the MiniFac and Ministep programs. This study has 10 stages, namely: (1) Defining the construct, (2) Identifying the construct, (3) Designing the items, (4) Testing the product, (5) Analyzing the data, (6) Reviewing the results of the analysis, (7) Reviewing the map wright, (8) Repeat steps 4-7, (9) Claim product quality, (10) Develop documentation. The results showed that the suitability of validation by experts using the MiniFac program was the exact agreement and expected agreement values were very close, namely 95.8% and 95.4%, this indicated that the results of this analysis fit the model and its estimation. Likewise testing of students using the Ministep program. Each item can be said to be valid because it meets the criteria of the MNSQ, ZSTD and PtMean Corr. The instrument is also claimed to be reliable because it has a value of 0.92 . The difficulty index and discriminating power of the items also varied from the easiest to the most difficult .

1. Introduction

In essence, natural phenomena which are related to the composition of matter, the shape of the structure, the properties of a substance, and also the energy changes

of a substance, are studied in chemistry. These natural phenomena can be understood if they involve students' skills and reasoning through solving problems that are tangible and can also be invisible or often called abstract, as stated by Johnstone, these phenomena cover three levels, namely macroscopic that can be observed or felt by the five human senses, sub-microscopic which is invisible or abstract and the last is symbolic (Johnstone, 2006). Based on the reality of three schools in Padang City, teachers teach at three levels, namely macroscopic, sub-microscopic and symbolic. Learning that applies these three levels is very helpful in increasing students' understanding of a material.

As stated by Gabel (1993), he said that if learning involves up to the particulate level (sub-microscopic level), it will help students connect understanding between the three levels of representation in chemistry learning and that will clearly increase students' understanding because this very related and interconnected and understanding at this level cannot be separated from the macroscopic or symbolic level. Even though learning has implemented all three levels of understanding, the test instrument used is inversely proportional to what has been taught. There are no questions that can test macroscopic and sub-microscopic understanding. The questions given are only at the symbolic level. Thus, it can be concluded that the application of test instruments that link the three levels of representation is still lacking and it is certain that the test instruments tested on students do not achieve KD in acid-base titration material, namely analyzing data from acid-base titration experiments.

The test instrument used only emphasizes calculation questions, so researchers develop test instruments that have been tested for validity and reliability, then have the right level of difficulty and discriminatory power. These four conditions will be created with the help of Rasch modeling analysis (Eliza and Yusmaita, 2021) . The Rasch model itself is a new measurement system, which aims to overcome the limitations of the classical measurement system or the Classical Test Theory (CTT) (Ashraf and Jaseem, 2020) . The choice of analysis using the Rasch model is because this model has four advantages, namely (1) It can overcome missing data. (2) Can identify error responses (3) The results of students' abilities are shown, not depending on the number of correct answers (4) Can identify careless and also predictive answers (Sumintono and Widhiarso, 2015) . It should be noted that if the data deviates greatly from the Rasch model, it is necessary that these items must be considered and or items that do not fit need to be deleted (Boone and Noltemeyer, 2017) . This study aims to produce a test instrument that can test students' understanding of acid-base titration material based on macroscopic, sub-microscopic, and symbolic levels using the Rasch model which is valid, reliable, has a good index of difficulty and item discrimination.

2.   Methodology

This development research uses the Rasch model adopted from the research of Wei et al. (2012) and modifications have been made according to the needs of this study. There are 10 stages of this research, which are as follows:

Defining Constructs
This initial stage will identify basic competencies and acid-base titration materials that will be tested on research subjects.

Identifying Constructs
At this stage, the preparation of competency achievement indicators based on the type and level will be compiled in *Learning Progression*.

Designing questions
At this stage the researcher designed three questions, in which each number had three other questions. These three questions are three different types of questions, namely (a) macroscopic, (b) sub-microscopic, (c) symbolic questions.

Test the product
Testing is carried out on predetermined subjects. The subjects chosen were six experts consisting of four Chemistry lecturers at UNP, two high school chemistry teachers from SMA N 2 Padang and SMA S Adabiah 2 Padang and testing the test instruments that had been validated by experts, namely 35 students from SMA N 2 Padang.

Analyze data
Rasch model used as a model for the development of test instruments by analyzing the responses of the items and the relationship between the level of ability of students and the level of difficulty of the items. The criteria to be analyzed are as follows:

1. **validity**
    In terms of validity, it was carried out by experts, data processing was carried out with the MiniFac program . The criteria seen in this program are *strata value* , *reliability* , *exact agreements* and *expected agreements* .
    In the Ministep program, data analysis uses *the output table* : Item Fit Order. In *the output table,* the values seen are MNSQ, ZSTD, and PtMean Corr.

2. **Reliability**
    On reliability, data analysis uses the Ministep program with *the output table* : Summary Statistics. In *the output table,* what is noticed is the item table in the reliability section of the real category.

3. **Difficulty Index**

   On the index of difficulty, data analysis uses the Ministep program with *the output table* : Item Measure. In *the output table,* the values that are seen are in the measure column and in the item column.

4. **Difference Power**

   At differential power , data analysis uses the same *output table as reliability, but the value seen here is in the separation section.*

Review analysis results
Review the results of the data analysis performed with the MiniFac and Ministep programs according to the criteria set by the Rasch modeling. Make revisions to the items that are less fit if needed.

View Wright's map
Review the results from the wright map. On this map you will see questions that are very difficult or very easy. Add or delete question items if needed.

Repeat steps 4-7
This step is carried out when needed. This means that if revisions are made to the items, then the research must be restarted from stages 4-7 until the desired results are obtained by the Rasch modeling.

Product quality claims
Determine the quality of the items, whether they are valid, reliable, have the right index of difficulty and discriminatory power.

Develop documentation
Develop documentation The intent is to provide information to assist users in applying the instrument appropriately. Important information included in this documentation is the purpose of using the test instrument, the definition of the construct, and guidelines for managing the test instrument along with the assessment rubric and the level of understanding of students.

3.   **Results and Discussion**

This study produced a test instrument that could test students' macroscopic, sub-microscopic and symbolic understanding of acid-base titration material. A quality test instrument must fulfill four conditions, namely valid, reliable, having differential power and the right index of difficulty. Specifically for validity conducted by experts, there will only be content/content validity and data analysis from validation by these experts using the MiniFac (Facets Rasch) program. For students, tests were carried out to analyze the four conditions mentioned above. Data analysis from testing students using Rasch modeling with the Ministep program. This research was conducted in 10 stages and at each stage the following results were obtained:

Defines a construct

In this early stage, the basic competencies of acid-base titration material will be identified which will be tested on research subjects. KD that researchers have determined namely KD 3.11 Analyzing data from acid-base titration experiments.

Identify constructs

At this stage, when the appropriate basic competencies have been determined. The next step is to reduce basic competence (KD) to become an indicator of competency achievement (GPA). Then after obtaining the GPA, determine the type of representation and cognitive level. GPA, cognitive level and its representation can be seen in Table 1.

Table 1. GPA, Cognitive Level and Representation

GPA	Cognitive Level	Representation
Analyzing strong acid-strong base titration curves (Key GPA)	C4	*Macroscopic* ; color change at the end point and pH *sub-microscopic* ; the number of solution molecules before, during and after the titration *symbolic* ; titration curve, reaction equation
Analyzing weak acid-strong base titration curves (Key GPA)	C4	*Macroscopic* ; color change at the end point and pH *sub-microscopic* ; the number of solution molecules before, during and after the titration *symbolic* ; titration curve, reaction equation
Analyzing weak base-strong acid titration curves (Key GPA)	C4	*Macroscopic* ; color change at the end point and pH *sub-microscopic* ; the number of solution molecules before, during and after the titration *symbolic* ; titration curve, reaction equation

Designing questions

At this stage designing the items by reducing the GPA that has been determined previously to become the question indicators, after that from the item indicators a question item is designed. From the decrease in GPA to be an indicator of this question, three indicator questions were obtained, namely for question number 1 which is a matter for a strong acid-strong base titration. Problem number 2 is a matter for a weak acid-strong base titration. Problem number 3 is a matter for a weak base-strong acid titration. Each question has three sub-tasks, namely (a)macroscopic, (b)sub-microscopic and (c)symbolic questions. One of the questions developed can be seen in Figure 1.
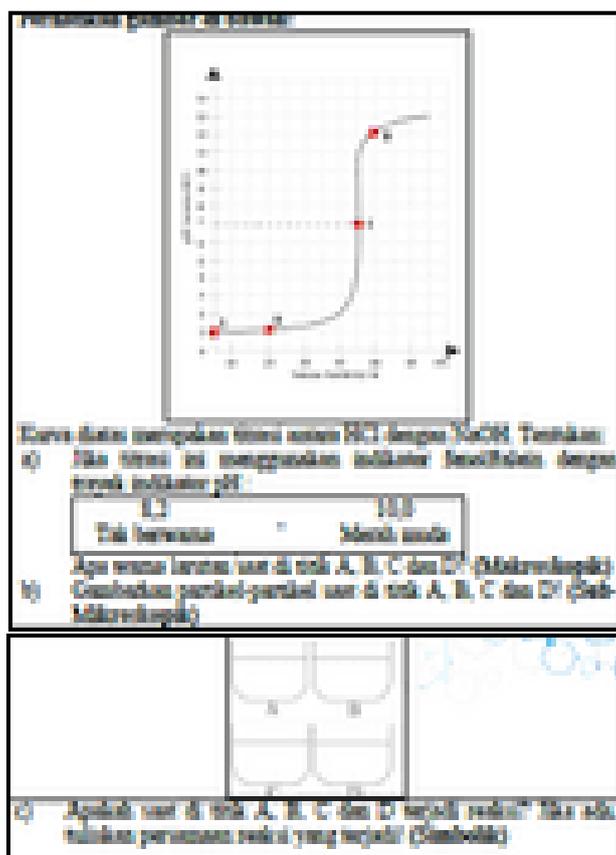
Figure 1. Question Number 1

Each question has an interconnected level of chemical representation in the sub-items of the question. This problem design tests understanding at the macroscopic level for sub-item (a), sub-microscopic for sub-item (b) and symbolic for sub-item (c). The GPA achieved in item number 1 is to analyze the strong acid-strong base titration curve. Sub-point no. 1(a) is a question to test understanding of concepts at the macroscopic level. Students are required to be able to know the color of the solution in each titration process that takes place. Question no 1(b) requires students to be able to explain/describe how the particles are in each titration process. Question no 1(c) requires students to write down the equation for the reaction that occurs during each titration process, this is knowledge at the symbolic level.

Test the product
At this stage, tests were carried out on predetermined subjects, namely with six experts consisting of four chemistry lecturers from FMIPA UNP, two high school chemistry teachers from SMA N 2 Padang and SMA S Adabiah 2 Padang and 35 students from SMA N 2 Padang. The selection of experts also goes through consideration, namely in expert expertise such as material experts and media experts. Likewise in the selection of 35 students, the students who were the subject were recommendations from teachers at schools from the high , medium and low ability levels of students .

The first test was carried out to experts , namely by testing the validity of the content/content by giving a questionnaire containing four aspects, namely material, construction, language and additional rules . After validation is complete, then proceed to try out the product on SMA N 2 Padang students. Testing is carried out by deploying test instruments that have been validated by experts. This test was carried out at the beginning of the even semester of the 2022/2023 academic year in class XII MIPA. The subjects were taken by class XII MIPA because it was the class XII students who had studied acid-base titration material, while class XI had not yet entered this material when the researchers conducted the research.

Analyze with the Rasch model and review the results of the analysis
At this stage, the data obtained from testing the subject analyzed in a different way. Because the experts carry out content validity tests while with students, they carry out product trials that have been validated by experts. To test content validity, data analysis used the Rasch model with the MiniFac (Facets Rasch) program. As for product trials for students, data analysis was carried out using the Rasch model with the Ministep program.

1. Validity
Validity can be interpreted as the extent to which the accuracy of the value of an instrument to be able to carry out a measurement function (Azwar, 2012) . In this validity there are differences in testing between validation with experts and students, which can be seen in the results below:

a. Expert Expert
Validation with experts was analyzed using the Rasch model with the MiniFac program (Facets Rasch). This program is another development of the Rasch model for *Multi-Rater data analysis* . This program can display which validators provide consistent assessments compared to other validators (Sumintono & Widhiarso, 2015) .

Figure 2. Wright Map Question Items

The results of this analysis can be seen in the figure. 1 , this is a Wright Map map which shows the distribution of item items, criteria and expert experts who provide assessments. This Wright map has four columns. The first column (far left) is called the measure column which has a logit scale of -2 to +2. Furthermore, in the second column, is the item column. This column shows the composition/distribution of the quality of the items that have been assessed by the validator. The higher the location of the item, the better the quality of the item. Furthermore, the third column is the assessment criteria column. The reading of this column can be seen from the criteria that are getting to the top, according to the validator, which are the most difficult criteria to achieve, and vice versa, the lower the location of the criteria, it indicates that according to the validator, these criteria are easy to achieve or easy to fulfill. Finally, in the fourth column, is the part that describes the validator's assessment scheme. The way of reading is the same as the third column. The higher the name of the validator, it can be said that the validator is difficult to give an assessment and the validator's name below is the easy validator to give an assessment.

On average, each item item met the criteria, but for items number 1A, 2A, and 3A it still did not achieve the criteria 'Images and text are presented clearly' according to the validator's assessment and it can also be said that this criterion is a criteria that is difficult to achieve.

Figure 3. Measurement Report Expert Items

Furthermore, this analysis can be summarized in a table. To see the results of the testers' measurements using the Rasch model in the form of strate value , reliability, *exact agreements* , and *expected agreements* can be seen in Figure 2 and can be summarized as Table 3 .

Table 2. Summary of Validator Analysis Results (Items)

Level Value	reliability	exact agreements	Expected agreements
2.61	0.74	95.8%	95.4%

Stratum value shows a value of 2.61 which indicates that the value is included in the sufficient category. Likewise with the reliability with a value of 0.74 which is also categorized as sufficient. The values of *the exact agreements* and *expected agreements* are also not much different. This indicates that the analysis shows a fit (*fit*) between the model and its estimation (Desnita et al., 2021) . From the results of the validation analysis of the contents of these items, it can be concluded that

there are no items that need to be revised because they are in accordance with the existing criteria.

b. Learners

In testing this validity, what is considered in Ministep is the *output table* : Item Fit Order. There are three *outfit criteria that must be met to get fit* or valid item items , namely *the Mean Square outfit* (MNSQ) *, the Z-Standard outfit* (ZSTD), and *outfit Point Measure Correlation* (Pt Measure Corr .). In looking at the validity of these item items, there is relief, namely the item 'may' be claimed to be valid if only one criterion has been met. That is, it doesn't matter if of the three criteria, there are two criteria that are not met (Sumintono & Widhiarso, 2015) .

The results of the analysis of the validity of these students can be seen in table 3. From this table it can be seen that the item with number 3A has a tendency to be unfit compared to the other items. Judging from the MNSQ outfit value 1 .61 ; ZSTD 1.87; and PtMeasure Corr. 0.56; only MNSQ did not meet the criteria. Therefore, item number 3A may still be maintained. For item 1A it also tends to be unfit with an MNSQ value of 0.45 ; ZSTD -0.3; Pt Measure Corr. 0.23. However, because the ZSTD value is still within the permissible limits, item 1A is still maintained without revision. For item 2B, it is an item that tends to fit because the MNSQ value is 0.61 ; ZSTD -1.46; Pt Measure Corr. Enter the fit category.

Figure 4. Fit Order Items

It can be seen that all item items have *outfit* values that meet the criteria and are in accordance with existing rules. So it can be concluded that all of these items can be claimed to be valid.

2. Reliability

Is a measuring tool that shows the results of measurements with an instrument whether it can be trusted or not (Friatma et al., 2017) . The results of data analysis using *the output table* : Summary Statistics can be seen in Figure 4. In the table it can be seen that the reliability value is 0.92 which indicates that it is in the very good category. So it can be concluded that this test instrument can be claimed to be reliable.

Figure 5. Output Table: Summary Statistics

3. Difficulty Index

Difficulty index is a number that indicates the level of difficulty or ease of an item (Daryanto, 2010) . The results of data analysis using *the output table:* Item Measure can be seen in table 6. This table shows the level of difficulty of the questions, it can be seen from the order of items from top to bottom the difficulty level of the questions gets easier. This sequence can be seen from the column marked in blue. To see the categories of difficulty level of the items can be seen in the table marked in red and the categories adjusted according to those specified in Figure 5. It can be seen that the questions with item number 2B are the most difficult questions based on this test. Questions in the medium category are one of the questions with item number 2C. The questions that were easiest to answer by students were questions with item number 1A.

Figure 6. Output Table: Item Measures

4. Difference Power

Distinguishing power is a differentiator between each group of questions according to the differences that exist in that group (Bagiyono, 2017) . In this analysis *the output table* used is Summary Statistics, the same as reliability but for differential power it can be seen in Figure 4 which has been marked with a blue box. This different power table reading is different from the previous one because it requires a formula to obtain an H value of 4.74 and rounds to 5 which indicates that this instrument has five different questions. There are questions that are very easy, easy, medium, difficult and very difficult. However, the results of this power difference are not in sync with the results from *the output table:* Item Measure and the Wright map which shows that the resulting power difference is only 2 levels of questions.

View the wright map

At this stage, it is useful to see an overview of the distribution of students' abilities with the distribution of the difficulty levels of the questions. The distribution of problem difficulties with students' abilities can be seen when they have tested the product on students and the results can be seen in Figure 3:
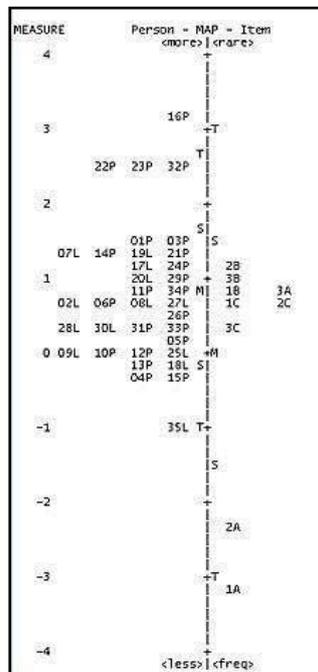

Figure 7. Wright Map Learners

It can be seen that students with number 16 P are students who have higher abilities than their peers with a logit value of more than +3 and are outside the T limit (*outlier*), this indicates that this student's ability is very high in workmanship. these instruments . These students can also answer all questions from each item, one of which is a difficult question, namely item 2B. Furthermore, students with number 35L are students who have lower abilities than their friends. However, these students were able to answer items number 1A and 2A where these two questions were the easiest questions with each logit value of more than -2 and -3. Item 1A lies outside the T limit (*outlier*), which means that this question is very easy because it lies outside the T limit. Questions that tend to be in the *extreme category* or too easy and difficult may be deleted or revised. Items with numbers 1B, 1C, 3A, 2C are questions that are in the medium category with a logit value of more than 0 and are located at the limit of M.

Product quality claims

At this stage it can be stated that based on testing on testing with the Facets and Ministep programs with the *output table:* Item fit order to test the validity and reliability have fulfilled the criteria but an inconsistency was found between the results of *the output table* : Item Measure and Separation. So, it can be claimed that this instrument does not meet the criteria of having the right index of difficulty and discriminating power.

Repeat steps 4-7

At this stage it is necessary to repeat at stages 4-7. Because the data that has been obtained does not match the criteria of the Rasch model.

Develop documentation

Developing documentation is meant to provide information to assist users in applying the instrument appropriately. Important information included in this documentation is the purpose of using the test instrument, the definition of the construct, and guidelines for managing the test instrument along with the assessment rubric and the level of understanding of students.

4.　Conclusion

Based on the results of the research and data analysis that the researchers have done, it can be concluded that the test instrument that can test students' macroscopic, sub-microscopic and symbolic understanding in acid-base titration material has been tested for validity and reliability but not with the difficulty index and discriminatory power . The problem is that research repetition is needed at stages 4-7.

Thank-you note

References

Ashraf, Z. A., & Jaseem, K. (2020). Classical and Modern Methods in Item Analysis of Test Tools. *International Journal of Research and Review*, *7*(7), 397–403.

Azwar, S. (2012). *Reliabilitas dan Validitas*. Pustaka Belajar.

Bagiyono. (2017). Analisis Tingkat Kesukaran dan Daya Pembeda Butir Soal Ujian Pelatihan Radiografi Tingkat 1 The Analysis of Difficulty Level and Discrimination Power of Test Items of Radiography Level 1 Examination. *Widyanuklida*, *16*(1), 1–12.

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, *4*(1), 1–13.

Daryanto. (2010). *Evaluasi Pendidikan*. Rineka Cipta.

Desnita, Yusmaita, E., Iswendi, & Iryani. (2021). Studi Tingkat Preferensi Panelis Terhadap Karakteristik Sensori Selai Kolang Kaling (Arenga Pinnata Fruits). *Jurnal Ilmiah Pengabdian kepada Masyarakat*, *5*(2), 75–81.

Eliza, W., & Yusmaita, E. (2021). Pengembangan Butir Soal Literasi Kimia pada Materi Sistem Koloid Kelas XI MIPA SMA/MA. *Jurnal Eksakta Pendidikan (JEP)*, *5*(2), 197–204.

Friatma, A., Syamsurizal, S., & Helendra, H. (2017). Analyzed Quality Question

of Final Exam Whole Semester on Biology Course Class XI IPA SMA Negeri District South of Solok Selatan Academic Year 2015/2016. *Bioeducation Journal*, *1*(2), 50–67.

Gabel, D. . (1993). *Handbook of Research on Science Teaching and Learning*. Maccmillan Company.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch: Pada Assessment Pendidikan*. Trim Komunikata.

Wei, S., Liu, X., Wang, Z., & Wang, X. (2012). Using rasch measurement to develop a computer modeling-based instrument to assess students' conceptual understanding of matter. *Journal of Chemical Education*, *89*(3), 335–345. https://doi.org/10.1021/ed100852t