Design and validation of six-tier astronomy diagnostic test instruments with Rasch Model analysis

G P Utari1*, W Liliawati2*, and J A Utama3

1Departemen Pendidikan Fisika, sekolah Pascasarjana Universitas Pendidikan Indonesia, Jl. Dr. Setiabudi No. 229, Bandung 40154, Indonesia
2Departemen Pendidikan Fisika, sekolah Pascasarjana Universitas Pendidikan Indonesia, Jl. Dr. Setiabudi No. 229, Bandung 40154, Indonesia
3Departemen Pendidikan Fisika, sekolah Pascasarjana Universitas Pendidikan Indonesia, Jl. Dr. Setiabudi No. 229, Bandung 40154, Indonesia

*gina.puri@upi.edu, winny@upi.edu

Abstract. Misconceptions occur because students answer problems with explanations that are not in accordance with scientific concepts and are confident in their choice of answers. Identification of misconceptions is an important first step to uncover understanding of learning. The six-tier test was developed from a four-tier test by adding two questionnaire sources of choice of answers to see how the causes of misconception. This study aims to design and validate misconception diagnostic test instruments on Astronomy material. The research method uses mixed methods with the instruments used namely the six tier Astronomy diagnostic test sheet for aspects of reveal, answer, synchronize, relationship, clear, EYD and corresponding. Appraisers numbered four experts to assess the construct and contents of the test instruments. The validation test uses multi-rater test analysis on MINIFAC and MINISTEP software to be analyzed with Rasch. The results obtained from the six-tier Astronomy diagnostic test validation analysis are the suitability of the items fulfilling the appropriate category. Item validity assessment items obtained that the questions can be accepted in all aspects. The conclusion is that the six tier Astronomy diagnostic test instrument is appropriate for identify misconceptions on Astronomy.

1. Introduction

Students often misunderstand, modify or reject the scientific assumptions used as the rationale for how and why things happen [1]. Misconceptions occur because students answer a problem with an explanation that is not in accordance with scientific concepts and are confident in the choice of the answer [2]. Misconception is a high level of student confidence in a concept that is considered not in accordance with scientific knowledge. Identification of misconceptions is an important first step for misunderstanding of student learning [3]. Methods that have been used to measure misconceptions include: concept maps, interviews, description tests, and multiple choice diagnostic test instruments [4]. Interview, open test, and MCT (multiple choice test) are tests commonly used in physics education research [3]. Multilevel multiple choice is used to overcome several limitations in using interviews, open tests and multiple choice to identify misconceptions in students [3]. For this reason, the researchers added multiple choice into several test levels, both two levels [5], three levels [6] or the most widely used today is the four-level / four-tier test [7].
Based on previous research, identification of misconceptions is carried out on concepts that have the potential to experience misconceptions, one of which is in the study of physics in astronomy concept. As revealed by Uygar Kanli (2014) by using a three-tier diagnostic test, there are misconceptions commonly found in astronomy concept, including: (1) reason for seasons is the elliptical orbit of the Earth around the Sun, (2) the Sun is always overhead at noon, (3) Moon moves into earth’s shadow, (4) Moon is at full phase in solar eclipse, (5) Earth closer to Sun in summer [8].

Name of six-tier tests is because this diagnostic test instrument has six levels. Six-tier test with the first level is a multiple choice content, the second level is the level of student confidence in choosing answers and the third level is a questionnaire source of student answers in answering questions level one and two accompanied by the level of student confidence in choosing at each distractor. At the fourth level is the reason students answer questions, the fifth level is the level of student confidence in choosing a reason, and the sixth level is a questionnaire source of student answers in answering questions level four and five accompanied by the level of student confidence in choosing at each distractor. Three outlines of the steps used to overcome misconceptions, namely uncovering misconceptions experienced by students, finding the cause of misconceptions, and looking for appropriate treatment or solution [9]. This research was conducted to produce a misconception diagnostic test in a six-tier Astronomy diagnostic test (STADT) that is able to correctly identify misconceptions and causes of student misconceptions.

2. Methods
The method used in this research is by combining two forms of research, namely quantitative research and qualitative research. The research design used is sequential exploratory design. Sequential exploratory design is collecting and analyzing qualitative data, then collecting and analyzing quantitative data [10]. The following is a flowchart in sequential exploratory design research.

![Figure 1. Exploratory sequential mixed methods design](image)

The mixed sequential exploratory mixed research method in this study is characterized by collecting data and analyzing qualitative data in the first stage. In this design, researchers use qualitative methods to describe a phenomenon that occurs in the field at the first stage. The second stage is the collection and analysis of quantitative data, in order to reinforce the results of qualitative research conducted in the first stage. Quantitative data used in this study was obtained from the data processing of the six tier Astronomy diagnostic test given to prospective physics teacher students from various LPTK.
Data processing includes determining validity, reliability, difficulty level, distinguishing features and suitability of the STADT diagnostic test.

2.1 validation test
Validity test is done by using multi-faceted Rasch measurement analysis on MINIFAC software to be analyzed with Rasch. After getting the trial data, the validation test is then performed using the output menu 10. Item (column): fit order in the MINISTEP software for analysis by Rasch. The measurement of the validity test of this instrument can be seen from the results of the scores on the Outfit mean square (MNSQ), Z-standard Outfit (ZSTD), and Point Measure Correlation (Pt Mean Corr). These criteria are used to determine the level of compliance items (item fit) . Knowing the validity of each question can be found by interpreting it in Table 1 as follows [11].

| Item output | Score    | Information |
|-------------|----------|-------------|
| MNSQ outfit | 0.5 <MNSQ <1.5 | Received    |
| ZSTD outfit | -2.0 <ZSTD <+2.0 | Received |

2.2 instrument reliability
Reliability test in this study uses MINISTEP software 4.3.1. Summary Statistics output menu that displays several reliability values, including person reliability, item reliability, and Cronbach alpha. The reliability value taken in this study is item reliability indicates the quality of the test items. Interpretations of item reliability values can be seen in Table 2 [12].

| Value of item reliability | Interpretation |
|---------------------------|----------------|
| 0.94 ≤ Value              | Special        |
| 0.91 ≤ Value <0.94        | Very good      |
| 0.81 ≤ Value <0.90        | Very nice      |
| 0.67 ≤ Value <0.80        | Enough         |
| Value <0.67               | Weak           |
2.3 difficulty level
The difficulty level of the STADT diagnostic test can be determined by using the MINISTEP 4.3.1 software on the output menu of Table 1 Variable (Wright) maps and Table 13 Item Measure output analyzed by Rasch. Output Variable (Wright) maps in this software aim to get an idea of the strength of each item. The difficulty class classification in this STADT instrument can be interpreted as in Table 3 as follows.

| Difficulty Level (TK) | Interpretation   |
|----------------------|------------------|
| 0.50 <TK             | Very difficult   |
| 0.00 <TK ≤ 0.50      | Hard             |
| -0.50 ≤ TK ≤ 0.00    | Easy             |
| TK < -0.50           | Very easy        |

2.4 distinguishing power
The distinguishing power of STADT can be determined using the MINISTEP 4.3.1 software with the Table 10 Item Fit Order output menu analyzed by Rasch. To be able to find out the distinguishing power categories for each of the STADT test instrument items in the PT-MEASURE CORR. Column, Smiley (2015) provides an interpretation of each value as presented in Table 4 as follows [13].

| Pt Mean Corr | Interpretation |
|--------------|----------------|
| 0.40 <ID     | Very good      |
| 0.30 ≤ ID ≤ 0.40 | Well        |
| 0.20 ≤ ID <0.30 | Not good    |
| ID <0.20     | Ugly           |

3. Result and Discussion
Six tier Astronomy diagnostic test developed to identify misconceptions on Astronomy material. The questions are 18 questions with 2 parts (parts A and B), where the questions section A is a question with three levels namely questions at the first level, confidence at the second level and the source chooses answers. While part B questions are questions with three levels namely reasons for choosing answers at the first level, beliefs at the second level and sources for choosing reasons. The results of the analysis of the validity of the questions obtained that the question number 16 part A (question) is the best problem of the other questions with the highest logit value, while question number 1 part B (reason) is a bad question because it has the lowest logit value. Questions that have good logit score are 14B, 15A, 15B, 17A, 3A and 6A.

Item validity assessment items consist of 7 aspects: (1) reveal, (2) answer (3) synchronize (4) relationship, (5) clear, (6) EYD, (7) corresponding. The validity of the aspects obtained results that the most difficult item to achieve is the "relationship" (highest logit) to the aspect indicators in the form of images used to help students understand the questions. While the item validity of an item that is easily achieved is " corresponding " (the lowest logit) with an aspect indicator in the form of a sentence the instructions filling in an item are written in a clear, correct and easy to understand order. Items with clear aspects, EYD and reveal are included in the category of "medium" where it is not too difficult and not too easy to achieve.

Figure 3. is a problem measurement report that displays the results of expert validation for the problem categories with RASCH analysis through MINIFAC software.
Based on Figure 3, problem measurement report, seen the reliability value of the questions obtained by 0.51, it can be concluded that the quality of the items included in the category of weak. The quality of the instrument in terms of the whole item is seen from the value of separation with a value of 1.02.

The results of the analysis of the difficulty level of the questions are categorized into several categories. Categories are very difficult on item problem numbers: 16A, 11A, 12A. Difficult categories for item number: 5A, 13A, 13B, 18A, 18B, 11B, 16B, 3A, 6A, 14B, 15A, 15B, 17A. Easy category on item problem numbers: 9A, 17B, 4B, 6B, 8A, 9B, 10A, 12B, 2B, 3B, 4A, 7B, 8B, 10B, 7A, 2A, 1A. Categories are very easy on item about numbers: 1B.

The criteria used to check the suitability of items that are not appropriate are: (1) questions for which the mean square outfit (MNSQ) is not accepted are numbers: 5A, 13A, 13B, 14A, 18A, 18B, 5B, 11B, 16B, 3A, 6A, 14B, 15A, 15B, 17A, 9A, 4B, 6B, 8A, 9B, 10A, 12B, 2B, 3B, 4A, 7B, 8B, 10B, 1A, 1B. (2) questions for which the Z-standard (ZSTD) outfit is not accepted are numbers: 16A, 1B. (3) questions that do not match point measure correlation (Pt Mean Corr) are absent. Based on the results of the criteria analysis, there are no items that do not meet the three criteria for item suitability. Items that only do not meet one of the criteria can then be maintained without being discarded. So it can be concluded that the suitability of the items meets the appropriate category.

Figure 4, measurement aspect report shows the results of expert validation for the problem categories with RASCH analysis through MINIFAC software.
Based on the aspect of measurement report in Figure 4, it can be seen that the reliability value is 0.98, it can be concluded that the quality of the aspects is included in the special category. The quality of the instrument in terms of all aspects seen from the value of separation with a value: 6.87. The results of the analysis of the difficulty level of aspects to be achieved are categorized in several categories. Categories are very difficult in aspects: relationship and synchronize.

Difficult categories in aspects: clear, reveal. Easy categories in aspects: EYD while categories are very easy in aspects: answer, corresponding.

The criteria used to check the suitability of items that are not appropriate are: (1) aspects for which the mean square outfit (MNSQ) value is not accepted are: clear, reveal, EYD, corresponding. (2) Aspects that are not Z-standard outfit (ZSTD) values are not accepted are absent. (3) Aspects that are worth points measuring correlation (Pt Mean Corr) are not appropriate are: answer, corresponding. Based on the results of the analysis of the criteria, it does not include aspects that do not meet the criteria for conformity of aspects. So it can be concluded that the questions can be accepted in all aspects.

4. Conclusion
The six-tier Astronomy diagnostic test for students on the suitability of the item meets the appropriate category and for the suitability aspects of reveal, answer, synchronize, relationship, clear, EYD and corresponding to the results obtained in the appropriate category. This research needs to be continued by implementing the use of diagnostic test instruments on students to determine the effectiveness of the six-tier Astronomy diagnostic test instrument in identifying misconceptions.

5. References
[1] Osborne R J, Bell B F and Gilbert J K 1983 Science teaching and children’s views of the world Eur. J. Sci. Educ. 5 1 p. 1–14
[2] Peşman H and Eryılmaz A 2010 Development of a three-tier test to assess misconceptions about simple electric circuits J. Educ. Res. 103 3 p. 208–222
[3] Kaltakci-Gurel D, Eryılmaz A and McDermott L C 2017 Development and application of a four-tier test to assess pre-service physics teachers’ misconceptions about geometrical optics Res. Sci. Technol. Educ. 35 2 p. 238–260
[4] Chu H E, Treagust D F and Chandrasegaran A L 2009 A stratified study of students’ understanding of basic optics concepts in different contexts using two-tier multiple-choice items Res. Sci. Technol. Educ. 27 3 p. 253–265
[5] Treagust D F and Chandrasegaran A L 2007 The Taiwan national science concept learning study in an international perspective Int. J. Sci. Educ. 29 4 p. 391–403
[6] Eryılmaz A 2010 Development and application of three-tier heat and temperature test: Sample of bachelor and students graduate Egit. Araştırmaları - Eurasian J. Educ. Res. 40 p. 53–76
[7] Caleon I S and Subramaniam R 2010 Do students know what they know and what they don’t
know? Using a four-tier diagnostic test to assess the nature of students’ alternative conceptions Res. Sci. Educ. 40 3 p. 313–337

[8] Kanli U 2014 A study on identifying the misconceptions of pre-service and in-service teachers about basic astronomy concepts Eurasia J. Math. Sci. Technol. Educ. 10 5 p. 471–479

[9] Suparno P 2005 Miskonsepsi dan Perubahan dalam Pendidikan Fisika (Jakarta: Gramedia)

[10] Creswell J W 2017 Research Design Qualitative, Quantitative, and Mixed Methods Approaches (Lincoln: SAGE Publication)

[11] Boone W J, Staver J R and Yale M S 2014 Rasch Analysis in the Human Sciences (New York Springer: Science Business Media)

[12] Sumintono, Bambang and Widhiarso W, 2015 Aplikasi Pemodelan RASCH pada AssessmentPendidikan (Cimahi: Trim Komunikata)

[13] Smiley J 2015 Classical test theory or Rasch- A personal account from a novice user Shiken 19 1 p. 16–29