

## DEVELOPMENT OF THREE-TIER DIAGNOSTICS INSTRUMENTS ON STUDENTS MISCONCEPTION TEST IN FLUID CONCEPT

Irwansyah<sup>\*1</sup>, Sukarmin<sup>2</sup>, Harjana<sup>3</sup>

<sup>1,2</sup> Postgraduate, Physics Education, Sebelas Maret University, Surakarta, Indonesia

<sup>3</sup> Department of Physics, Faculty of Mathematics and Science, Sebelas Maret University, Surakarta, Indonesia  
Jl. Ir. Sutami 36A Kentingan Jebres Surakarta 57126, Indonesia

\*Correspondence address: irwansyah200893.is@gmail.com

Received: July 18<sup>th</sup>, 2018. Accepted: October 18<sup>th</sup>, 2018. Published: October 28<sup>th</sup>, 2018

**Abstract:** *The purpose of this research was to know the validity and reliability of instrument three-tier diagnosis to identify the students' misconception in fluid concept. The stages of development research that used in this research were (1) The first research and collecting information, (2) planning and designing of development (3) validity and product evaluation. The instrument of the test was tested on 98 students of 3 schools. The instrument has been developed and categorized effectively because it has a valid category with the Aiken validity index value > from table V (0.75) and the reliability of the instrument has a value of 0.96 with a high category. The instrument of three-tier test that developed was able to identify the students' comprehension in students' concept and students' misconception. There are 27,58% students comprehend the concept, 45,29% students did not comprehend the concept, 24,74% students showed that misconception and 2,36% students experienced errors.*

© 2018 Physics Education, UIN Raden Intan, Lampung, Indonesia.

**Keywords:** *development, fluid, misconception, three-tier test*

### INTRODUCTION

One of the main problems that appear in physics education is students' misconceptions (Wijaya, Supriyono Koes, & Muhandjito, 2016). When the students learn about everything around them either through formal school education or non-formal education through daily experience, students tend to develop comprehension of something based on their own views. Because this concern, several researchers have done the research to describe the students' comprehension. The variety of comprehension types that are formed by students are called by several terms, such as "alternative conceptions", "misconception", "naive beliefs", "children's ideas", conceptual difficulties", "phenomenological primitives", "mental models" and many the others term (Gurel, Eryilmaz, & McDermott, 2015). Lack of comprehension through a concept can be improved by further instruction and

learning, while misconception is believed to able inhibiting the acceptance and development of knowledge and students' ability (Hasan, Bagayoko, & Kelley, 1999). Misconception must be overcome because it can give negative effect in the further learning process (Djanette & Fouad, 2014; Lucariello, Tine, & Ganley, 2014; Sholihat, Samsudin, & Nugraha, 2017). Based on this thing, it is necessary to identify the students' comprehension of whether the students have understood the concept well, students have not understood the concept well or the students faced misconception.

The misconception is the condition where the students' comprehensions are different from the experts' comprehension (Resbiantoro & Nugraha, 2017; Wijaya et al., 2016). Misconception can be also identified as the concept that contrary to theories that are scientifically accepted and generally accepted (Gurel et al., 2015). Misconception can be identified in

some methods such as using instrument diagnostic test that used to identified the students' misconception is Three-tier instrument (Zukhruf, Khaldun, & Ilyas, 2016). Three-tier tests are considered to have advantages in distinguishing the lower students' comprehension and students that faced fault incomprehension or misconception (Gurel et al., 2015).

Three-tier test instrument can identify students' comprehension of the concept and students do not need many times (Wijaya et al., 2016; Wisudawati, 2015). Besides the Three-tier test instrument can distinguish that students comprehend the concept, misconception, guesswork (do not have self-confidence), and do not the concept (Abbas, 2016). Besides it can identify the students' misconception, the three-tier test instrument also able to identify the students that comprehend the concept or less comprehend the concept (Aini, Ibnu, & Budiasih, 2016). The three-tier instrument is that used in the second level. The test participants are still given an alternative answer and sometimes the test participant just guesses and it is difficult to know the difficulties cause that experienced by the students (Suwanto, 2013). In this research, the researcher wants to develop the three-tier test instrument that able to identify the students' difficulties in the learning that is by using the three-tier instrument with the open-minded reason. In preparing the instrument or test instrument need to be done accurately based on the rule that determined by education measurement and curriculum (Prihatni, Kumaidi, & Mundilarto, 2016; Siregar, 2014). The instrument that used an instrument of collecting data must be done validity besides to acquire suggestion and improvement that purpose to give assessment through items in the instrument (Suryani, Kartowagiran, & Jailani, 2017). Validity is evidence and theory through the interpretation of the test score based on the purpose of test

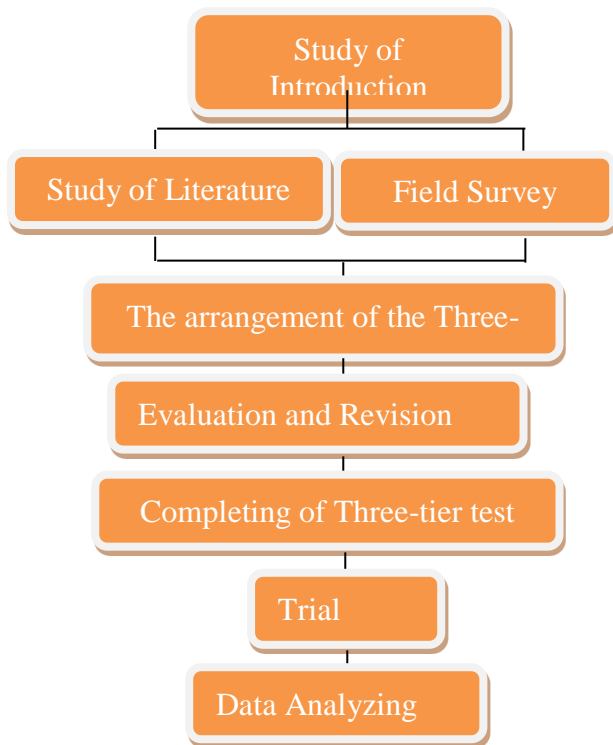
usage (Pruyn, Watsford, & Murphy, 2016).

### **Diagnostic Formative Assessment**

Assessment can be said as an act of collecting data in order to make decisions based on information obtained in the testing phase. Some of the assessment characteristics in learning such as the following: (1) Assessment is started by collecting some information about students in the learning; (2) In the assessment activity is done by analyzing and interpretation through data and information are that collected; (3) Interpretation results decisions about learning; (4) There is a follow-up to the decisions produced; (5) Assessment is carried out on an ongoing basis (Kusairi, 2012). Some knowledge about formative assessments include: (1) Formative assessment is a process carried out in learning; (2) Formative assessment results are not only used by teachers but also by students; (3) Formative assessment provides feedback on student learning and the learning process carried out by the teacher; (4) Feedback provided by formative assessment will be useful for students and teachers to make arrangements so that learning and learning can achieve curriculum goals (Hermawanto, Kusairi, & Wartono, 2013). With formative assessment in the form of objective tests, it can help students who have low learning outcomes to learn the subject matter as a whole. Formative diagnostic assessment is an instrument prepared to find out the learning outcomes that aim to identify or diagnose learning outcomes so that feedback occurs on the results of the use of formative diagnostic assessment.

**METHOD**

The Research Procedure



**Figure 1.** The Research Procedure

This research is a type of research that aims to analyze the validity, reliability, level of difficulty, differentiation, deception index of students' misconception diagnosis instruments and analyze students' learning difficulties that cause the students do not understand the concept or students that faced misconception in the fluid concept. The research is started with preliminary study through literature review and empirical research (Field Survey). The result of the literature review showed that an instrument was said to be good if it has validity, reliability, level of difficulty, differentiation and a good deception index. The research begins in December 2017 until March 2018. The initial drafting of the instrument, validation, and test is conducted in December 2017 until February 2018. The instrument trials are conducted from February to March 2018. The research was conducted in SMA and

Madrasah aliyah in the Karanganyar region of Java Middle, with high, medium and low categories based on the results of 2014, 2015 and 2016 national examinations. The subjects of this research were 98 students of class XI IPA consisting of 1 class of SMA NI Karanganyar, 1 class of SMA Gondangrejo, and 1 class of MA N Gondangrejo. Before the instrument was used to identify student misconceptions, a Focus Group Discussion (FGD) was carried out by involving experts in the fields of Physics, Physics Education, measurement and Physics Education practitioners. Based on the FGD, there are several instruments that need to be revised. After the revision is complete, the instrument can be used to identify student misconceptions, but first, it is necessary to analyze the validity, reliability, differentiation, level of difficulty and index of the deception.

Tests of validity, reliability, level of difficulty, distinguishing power and the test index of the test instruments to identify student misconceptions involved 1 class from each school with a total of 98 students. In the testing phase of the instrument, it is intended to obtain information about the feasibility, adaptability, and functioning of the test instrument on field conditions. In addition, the trials carried out can also provide information about the validity, reliability, differentiation, level of difficulty and deception index of an instrument. The validity of an instrument does not apply generally to all measuring objectives. A test generally only produces a valid measure for a particular measurement goal.

Validity of the contents of a test instrument can be done by looking at the suitability of the indicators in the grid with the operational definition of the test instrument and with the operational definition of constructing instruments, determination of content validity can be

done through focus group discussions (PGD) conducted by at least 5 experts and practitioners in measured field. An instrument will be said to be valid if it has an Aiken validity index greater than V table. According to the V table compiled by Aiken, a question is said to be valid with the number of validators 8 people at least the question has a validity index of 0.75 (Aiken, 1985).

Based on the validity of the test items, the value of 0.74 indicates good and acceptable validity (Mohamad, Sulaiman, Sern, & Salleh, 2015; Yasin, Yunus, Rus, Ahmad, & Rahim, 2015). Aiken's validity is formulated by the equation:

$$V = S / [n * (c - 1)] \text{ Where } S = \sum ni(r - \ell_0)$$

Where:

V: Aiken validation index

C: Number of categories or criteria

$\ell_0$ : The lowest category

$ni$ : Number of assessors who choose criteria i

r: Criteria to i

n: Number of all assessors

In addition to the validity of an instrument, the reliability of an instrument is also much needed in the preparation of an instrument, this is because the reliability of an instrument shows the consistent validity of an instrument (Mohamad et al., 2015; Yasin et al., 2015). The reliability of an instrument is a determination of the results of a test, or if a change occurs, then the change can be said to be meaningless or significantly, Reliability itself can be said as the instrument's freedom of error and can produce a consistent result. The reliability of an instrument can be obtained through the results of testing instruments to test participants, in addition by testing the test instruments, information on the difficulty of the items, the item distinguishing power, and the index of the items about an instrument (Yamtinah, Saputro, & Utami, 2015). A test instrument is said to

be reliable if the instrument items are consistent in measuring a person so that the results of measurements are performed on the same subject even at different times. There is no agreement in general what is the minimum coefficient of an instrument so that the instrument can be said to be reliable, but for a test to be used to decide about students, it must have a reliability coefficient of 0.70 (Fitriatun & Sukanti, 2016). Determination of the reliability of the assessment instrument developed in this study is to use the equation or formula Cronbach Alpha as follows (Fitriatun & Sukanti, 2016):

$$\alpha = \frac{k}{k-1} \left( 1 - \frac{\sum Si^2}{S^2} \right)$$

Where:

K = Number of test items

$Si^2$  = Variant total test score

$\sum Si^2$  = Number of all item variants forming test

**Table 1.** Interpretation of reliability values refers to Guilford's opinion

Score $\alpha$	Reliability
< 0,70	Low
$\geq 0,70$	High

Cronbach Alpha is generally used when we measure tests that have standard multiple-choice items or in the form of essays. Cronbach Alpha in principle includes measuring homogeneity in which it focuses on two important aspects, namely the content aspect and the heterogeneity aspect of the test. If the test item is heterogeneous, it means measuring more than one characteristic, traits or attributes and will cause the alpha coefficient to be lower. Conversely, if the test is more homogeneous then the alpha coefficient price will be higher which means the test is more consistent (Shirali, Shekari, & Angali, 2017).

Reliability is the stability of the score obtained by the same person when tested again with the same test in different situations or from a measurement of other

measurements, so that reliability can be said as the level of consistency of the two measurements of the same thing. By using a reliable test tool, a person will have a constant score if given the same test even at different intervals, for which a good test instrument must also have a good test item difficulty level.

The level of difficulty of a question can show the level of quality of a test item. Test items can be stated as items that are good if the items are not too difficult and not too easy, in other words, the degree of difficulty of an item is moderate or sufficient. Good questions are questions that are not too easy and not too difficult, questions that are too easy do not stimulate students to improve their efforts to solve them and vice versa if the problem is too difficult will cause students to despair and not have the enthusiasm to try again. Equations are used to determine the level of difficulty with the correct proportion of answers (Fitriatun & Sukanti, 2016):

$$P = \frac{\sum x}{S_m N}$$

Where:

- P = The proportion answers correctly or the level of difficulty
- $\sum x$  = The number of test participants who answered correctly
- $S_m$  = Maximum Score
- N = Number of test takers

**Table 2.** Difficulty level categories

Score of P	Question category
$P \leq 30$	Difficult
$0,30 < P \leq 0,70$	Medium
$0,70 < P \leq 1,00$	Easy

(Arikunto, 2013)

Distinguishing power is one of the points that must be considered in the preparation of instruments or questions in the analysis of learning outcomes. Analysis of the differentiation of items was conducted to determine the difference power which can be seen from the discrimination index value for each item.

The differentiating power of a problem is the ability of a question to distinguish between students who are clever (highly capable) with students who are stupid (low ability). The distinguishing factor of a test functions to determine whether or not a question can distinguish groups in aspects measured according to differences in the group, in principle the distinguishing index is calculated on the basis of group division into two parts, namely the upper group which is a group of capable test takers high with the lower group, namely the low-ability group of test takers. Different power analysis can use equations:

$$rbis = \frac{Y_b - Y_s}{SD} \cdot \frac{nb.bs}{un \sqrt{n^2 - n}}$$

$$rpbis = \frac{X_b - X_s}{SD} \sqrt{pq}$$

Where:

- $X_b, Y_b$  = The average score of students who answer correctly
- $X_s, Y_s$  = The average score of students who answer incorrectly
- SD = Standard Deviation
- Nb, ns = Number of students who answered correctly, Number of students who answered incorrectly
- p = Proportion of the correct answer for all answers
- q = 1-p
- u = Ordination of a normal curve

The distinguishing criteria are interpreted using the following references:

**Table 3.** Classification of differentiation (D)

Distinction index (D)	Category
$D \leq 0,20$	Poor
$0,20 < D \leq 0,40$	Satisfactory
$0,40 < D \leq 0,70$	Good
$0,70 < D \leq 1,00$	Excellent

(Arikunto, 2013)

The answer pattern in a question is known by calculating the number of tests who chose each option provided (Fitriatun & Sukanti, 2016). A deception index can be said to be good or has performed its

function well if it has such attractiveness that the test participants feel hesitant and hesitant so that in the end, they become fooled to choose a distractor as the correct answer. In general, what applies to the evaluation of learning outcomes is that distractors have been able to carry out their duties well if the distractors have at least been chosen by 5% of all test participants (Fitriatun & Sukanti, 2016). Distractor efficiency (DE) is calculated based on the number of NFD in an item and ranges from 0 - 100%. NFD is choices other than keys when chosen not often by respondents (<5%), and does not carry out their functions (D'Sa & Visbal-Dionaldo, 2017; Mahjabeen et al., 2018).

**Table 4.** Questions Category based on Distinction index Factor (Fitriatun & Sukanti, 2016)

Non-functional distractors Item	Items
0	Very Good
1	Good
2	Enough
3	Less Good
4	Not Good

The instruments used in this research used conceptual diagnostic tests consisting of three levels or three-tier tests. The three-level diagnostic test instrument used in this research consisted of 28 items from 11 fluid concepts. Research to analyze students' misconceptions about fluid concepts is carried out in several stages, the initial stage of data collection, data processing, description and discussion of findings in the study. Data analysis uses the concept of analysis compiled by Kaltakci (Kamilah & Suwarna, 2016).

**Table 5.** Misconception Analyzing

The first stage	The second stage	The third stage	Categories
Correct	Correct	Believed	Understand the concept
Correct	Correct	Not believed	Not understand the concept
Correct	Incorrect	Believed	Positive False or

The first stage	The second stage	The third stage	Categories
Correct	Incorrect	Not believed	Misconception Not understand the concept
Incorrect	Correct	Believed	Negative false or error
Incorrect	Correct	No believed	Not understand the concept
Incorrect	Incorrect	Believed	Misconception
Incorrect	Incorrect	Not believed	Not understand the concept

## RESULTS AND DISCUSSION

### Validity

An instrument that has been prepared in draft form before used it must be validated. The purpose of validation is to get feedback, criticism, and suggestions for improvement in accordance with the expertise of each validator. Expert validation aims to provide an assessment of the instruments that have been prepared. The assessment carried out by the validator can be in the form of conformity with the indicators that have been prepared, the suitability of the material, conformity of the choice of answers, the suitability of the language or conformity of the instrument as a measuring instrument.

An instrument has that prepared to be validated by 4 experts and 4 teachers, the data were analyzed by using Aiken equation. Based on the validity analyzing was getting the data that 28 questions that were prepared and had the valid category where for the Aiken validity index more than V table for 8 validators is 0.75. Based on the validity analyzing was getting the data that:

**Table 6.** Validity Analyzing Items

Items Number	Category
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28	Valid
-	Invalid

Based on the results of the analysis of the validity of the items it was found that 28 questions were stated in the valid

category where for the Aiken validity index results for 8 validators higher than V tables or higher than 0.75. If the item is valid then the question is feasible to use.

**Reliability**

The reliability analysis of the items was carried out on 28 questions that had good validity, based on the results of the reliability analysis using the Cronbach Alpha formula with the help of the Quest program, it was obtained data that:

**Table 7.** Reliability analysis of test items

Types of Reliability	Score of Reliability	Category
Reliability item	0,96	High

Reliability analysis of the questions was carried out on 28 valid questions. Based on the results of the reliability analysis using the Cronbach Alpha formula, the reliability of the questions was 0.96 with a high category.

**Difficulty level**

Difficulty level analysis of test items was carried out on 28 questions that had been tested for validity. Based on the analysis of the difficulty level of the test items, it was found that:

**Table 8.** Analysis of the difficulty level of test items

Test items	Category
2,3,5	Easy
1,4,6,7,8,10,11,12,13,15,16,17,18,19,21,22,23,27,28	Medium
9,14,24,25,26	Difficult

**Table 9.** Percentage of the difficulty level of test items

Category	Percentage
Easy	10,71 %
Medium	71,42 %
Difficult	17,85 %

Based on the results of the data analysis of the difficulty level of the test items shown in the table, it can be seen

that from 28 test questions that have been tested, information is obtained that the difficulty level of the test questions is divided into three categories, namely easy, moderate, and difficult. Questions that have an easy level of difficulty are shown no 2,3,5 or with a percentage of 10.71%, questions that have a level of difficulty with the category being shown no.1, 4, 6, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18, 19, 21, 22, 23, 27, 28 or with a percentage of 71.42% and questions that have a level of difficulty with difficult categories are shown no 9,14,24,25,26 or with a percentage amounting to 17.85%.

**Differentiating Power**

Differential power analysis was conducted to determine the quality of the items in distinguishing between the upper groups who answered correctly and the lower groups who answered correctly. Distinguishing power analysis was carried out on 28 items that had good validity. Based on the distinguishing analysis, data were obtained that:

**Table 10.** Distinguishing power Analyzing

Items number	Category
1,3,5,7,14,24,25	Bad
2,4,6,10,11,13,15,18,20,21,27,28	Sufficient
8,9,12,16,17,19,22,23,26	Good

**Table 11.** Percentage of Distinguishing power Analyzing

Category	Percentage
Bad	25 %
Sufficient	42,85 %
Good	32,14 %
Very good	0 %

Based on the results of the analysis of distinguishing data shown in the table, it can be seen that out of 28 questions compiled and tested on students, information was obtained that the differentiating problem was divided into 3 categories, namely bad, sufficient and good. Test items that have a differentiation with a bad category are numbered 1,3,5,7,14,24,25 or with a

percentage of 25%, test items that have distinguishing features from the category are simply indicated number 2,4,6, 10,11,13,15,18,20,21,27,28 or with a percentage of 42.85% and test items that have distinguishing features with good categories shown in numbers 8,9,12,16,17,19, 22,23,26 or with a percentage of 32.14%.

### Deception Index

The deception index analysis was carried out to find out how the functioning of the alternative answers to the errors given to the multiple-choice test items. Based on the analysis of the alternative index, the answers to the test items were found that:

**Table 12.** Analyzing of deception alternative answer

Alternative Answer	Category
2 / (b,c,d), 5/(b,c,d), 8/(a,c,d)	$\leq 0,05$ ((less good)

**Table 13.** Percentage Analyzing of deception alternative answer

$\leq 0,05$ (less good)	$\geq 0,05$ (good)
6,42%	93,57%

Based on the results of the analysis of the alternative answers to the deception data shown in the table, it is known that from 28 test items with 5 alternative answers on each test item, it was found that alternative answers b, c, and d on test items number 2, b, c and d on Test items number 5 and a, c and d on item number 8 have not had good or not selected deception at least 5% of the total test takers.

### Conception of Students

Based on the results of student misconception analysis of the fluid concept using the three-tier test instrument, the data obtained that:

**Table 14.** Percentage of students' concept comprehension level about the fluid concept

Level of comprehension concept	Percentage
Understand the concept	27,58
Not understand the concept	45,29
Misconception	24,74
Error	2,36

Table 14 showed that from 98 students in SMA XI IPA grade was gotten data that 27,28% students understood the fluid concept. 45,29% students have not understood the fluid concept, 24,74% students have misconceptions and 2,36% students have an error. Based on the analyzing of data identified students' misconception, there are many students that faced misconception. The difference between students that did not understand and the students and students faced misconception lies in the students' belief about their answer was given. If the students believed through their answer is given but their answer was wrong so that the students were categorized as a misconception, even though if the students did not believe with their answer given so that the students were categorized that they did not understand the concept.

Misconceptions that occur in students are not limited simply because their answers are wrong, however, based on an analysis of understanding the concepts of students with three-tier test instruments with open reasons or students not given alternative answers, it is found that students find the difficult to give reasons for their answers given it to the first level question. It shows the difficulty of students in communicating in conveying what they understand.



## CONCLUSION AND SUGGESTION

The results of the validation by eight validator experts found that the test items developed were included in the valid category. Based on the results of the analysis of the diagnostic test instrument, it was found that the 28 test items used in this study had the reliability of 0.96 with a high category so that the instruments prepared were reliable and feasible to be used as test kits. The level of difficulty of items consists of 3 items with easy categories, 19 items with medium categories and 5 items with difficult categories. The differentiation of test items consisted of 7 items with a bad category, 12 items with enough categories and 9 items in a good category, for the fraudulent index the alternative answers that had not been selected were at least 5% of the total test participants, namely alternatives B, C and D on item number 2, B, C and D in points 5 and A, C and D in number 8.

The next research for the researchers suggested that the test sample diagnostic test instruments use more varied schools so that the diagnostic test instruments compiled have better validity and reliability. Item Questions developed should be made more in-depth in identifying misconceptions experienced by students so that the information obtained by the teacher will be more complete. In addition, diagnostic instruments to identify student misconceptions should be arranged in the form of four-tier tests for open reasons.

## REFERENCES

- Abbas, M. L. H. (2016). Pengembangan Instrumen Three Tier Diagnostic Test Miskonsepsi Suhu dan Kalor. *Ed-Humanistics*, 01(02), 83–92.
- Aiken, L. R. (1985). *Psychological Testing and Assessment* (5th Editio). Massachusetts: Allyn and Bacon, Inc.
- Aini, R. G., Ibnu, S., & Budiasih, E. (2016). Identifikasi Miskonsepsi dalam Materi Stoikiometri pada Siswa Kelas X di SMAN 1 MALANG Melalui Tes Diagnostik Three-Tier. *Jurnal Pembelajaran Kimia*, 01(2), 50–56.
- Arikunto, S. (2013). *Prosedur Penelitian Suatu Pendekatan Praktik*.
- D'Sa, J. L., & Visbal-Dionaldo, M. L. (2017). Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency. *International Journal of Nursing Education*, 9(3), 109. <https://doi.org/10.5958/0974-9357.2017.00079.4>
- Djanette, B., & Fouad, C. (2014). Determination of University Students' Misconceptions about Light Using Concept Maps. *Procedia - Social and Behavioral Sciences*, 152, 582–589. <https://doi.org/10.1016/j.sbspro.2014.09.247>
- Fitriatun, A., & Sukanti. (2016). Analisis Validitas, Reliabilitas dan Butir Soal Latihan Ujian Nasional Ekonomi Akuntansi di MAN Maguwaharjo. *Jurnal Kajian Pendidikan Akuntansi Indonesia*, 8, 1–11.
- Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science and Technology Education*, 11(5), 989–1008. <https://doi.org/10.12973/eurasia.2015.1369a>
- Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the Certainty of Response Index (CRI). *Physics Education*, 34(5), 294–299. <https://doi.org/10.1088/0031-9120/34/5/304>
- Hermawanto, Kusairi, S., & Wartono. (2013). Pengaruh Blended Learning Terhadap Penguasaan Konsep dan

- Penalaran Fisika Peserta Didik Kelas X. *Jurnal Pendidikan Fisika Indonesia*, 9(57), 67–76.
- Kamilah, D. S., & Suwarna, I. P. (2016). Pengembangan Three-Tier Test Digital untuk Mengidentifikasi Miskonsepsi pada Konsep Fluida Statis. *EDUSAINS*, 8(2), 212–220.
- Kusairi, S. (2012). Analisis Asesmen Formatif Fisika SMA Berbantuan Komputer. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 16(3), 68–87.
- Lucariello, J., Tine, M. T., & Ganley, C. M. (2014). A formative assessment of students' algebraic variable misconceptions. *Journal of Mathematical Behavior*, 33(1), 30–41.  
<https://doi.org/10.1016/j.jmathb.2013.09.001>
- m. duskri, kumaidi, S. (2012). Jurnal Penelitian dan Evaluasi Pendidikan. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 16(2), 492–510.
- Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. (2018). Difficulty Index, Discrimination Index and Distractor Efficiency in Multiple Choice Questions. *Annals of PIMS*, 310–315.
- Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the Validity and Reliability of Research Instruments. *Procedia - Social and Behavioral Sciences*, 204(August), 164–171.
- Prihatni, Y., Kumaidi, K., & Mundilarto, M. (2016). Pengembangan Instrumen Diagnostik Kognitif Pada Mata Pelajaran Ipa Di Smp. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 20(1), 111.  
<https://doi.org/10.21831/pep.v20i1.7524>
- Pruyn, E. C., Watsford, M. L., & Murphy, A. J. (2016). Validity and reliability of three methods of stiffness assessment. *Journal of Sport and Health Science*, 5(4), 476–483.  
<https://doi.org/10.1016/j.jshs.2015.12.001>
- Resbiantoro, G., & Nugraha, A. W. (2017). Miskonsepsi Mahasiswa Pada Konsep Dasar Gaya Dan Gerak Untuk Sekolah Dasar. *Jurnal Pendidikan Sains (JPS)*, 5(2), 80–87.
- Shirali, G., Shekari, M., & Angali, K. S. (2017). Assessing Reliability and Validity of an Instrument for Measuring Resilience Safety Culture in Sociotechnical Systems. *Safety and Health at Work*, 9(3), 296–307.
- Sholihat, F. N., Samsudin, A., & Nugraha, M. G. (2017). Identifikasi Miskonsepsi dan Penyebab Miskonsepsi Siswa Menggunakan Four-Tier Diagnostic Test Pada Sub-Materi Fluida Dinamik: Azas Kontinuitas. *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, 3(2), 175–180.  
<https://doi.org/10.21009/1.03208>
- Siregar, S. (2014). *Statistik Parametrik untuk Penelitian Kuantitatif*. Jakarta: Bumi Aksara.
- Suryani, H., Kartowagiran, B., & Jailani. (2017). Development and Validity of Mathematical Learning Assessment Instruments Based on Multiple Intelligence. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 21(1), 93–103.
- Suwarto. (2013). Model-model Instrumen Diagnostik. *WIDYATAMA*, 22(1), 64–70.
- Wijaya, C. P., Supriyono Koes, H., & Muhardjito. (2016). The diagnosis of senior high school class X MIA B students misconceptions about hydrostatic pressure concept using three-tier. *Jurnal Pendidikan IPA Indonesia*, 5(1), 14–21.  
<https://doi.org/10.15294/jpii.v5i1.5784>
- Wisudawati, A. W. (2015). Pengembangan Instrumen Three – Tier Test Untuk Mengidentifikasi Representasi Tingkat Mikroskopis

- Perubahan Wujud Air Sebagai Alternatif Assesment Integrasi Islam-Sains ( A Preliminary Study ). *Seminar Nasional Kimia Dan Pendidikan Klmia VII*.
- Yamtinah, S., Saputro, S., & Utami, B. (2015). Content Validity and Scoring of Two Tier as Measuring Instrument of Science Process Skills for Knowledge Aspects in Chemistry Learning. *Prosiding ICTTE FKIP UNS, 1*(1), 911–916.
- Yasin, R. M., Yunus, F. A. N., Rus, R. C., Ahmad, A., & Rahim, M. B. (2015). Validity and Reliability Learning Transfer Item Using Rasch Measurement Model. *Procedia - Social and Behavioral Sciences, 204*(November 2014), 212–217.
- Zukhruf, K. D., Khaldun, I., & Ilyas, D. S. (2016). Remediasi Miskonsepsi Dengan Menggunakan Media Pembelajaran Interaktif Pada Materi Fluida Statis. *Jurnal Pendidikan Sains Indonesia, 04*(02), 56–68.